Search as News Curator:

The Role of Google in Shaping Attention to News Information

Daniel Trielli
School of Communication
Northwestern University
dtrielli@u.northwestern.edu

Nicholas Diakopoulos
School of Communication
Northwestern University
nad@northwestern.edu

ABSTRACT

This paper presents an algorithm audit of the Google Top Stories box, a prominent component of search engine results and powerful driver of traffic to news publishers. As such, it is important in shaping user attention towards news outlets and topics. By analyzing the number of appearances of news article links we contribute a series of novel analyses that provide an in-depth characterization of news source diversity and its implications for attention via Google search. We present results indicating a considerable degree of source concentration (with variation among search terms), a slight exaggeration in the ideological skew of news in comparison to a baseline, and a quantification of how the presentation of items translates into traffic and attention for publishers. We contribute insights that underscore the power that Google wields in exposing users to diverse news information, and raise important questions and opportunities for future work on algorithmic news curation.

CSS CONCEPTS

• Information systems~Web search engines • Information systems~Page and site ranking

KEYWORDS

search engines; news curation; news diversity; algorithm audit

ACM Reference format:
Daniel Trielli and Nicholas Diakopoulos. 2019. Search as News Curator: The Role of Google in Shaping Attention to News Information. In *2019 CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glagsow, Scotland, UK*. ACM, New York, NY, USA. 13 pages. https://doi.org/10.1145/3290607.3300683

CHI 2019, May 4-9, 2019, Glasgow, Scotland, UK.

ACM ISBN 978-1-4503-5970-2/19/05...$15.00.
DOI: https://doi.org/10.1145/3290605.3300683

1 INTRODUCTION

When it comes to the interaction of the public with news, search engines are an increasingly powerful intermediary, both in exposing audiences to news information and assisting them in making sense of it. A Pew survey from 2017 [1] showed that 43% of Americans get their news online, up from 38% in the previous year and closing in on the percentage of people who got their news from television (50%). In a survey by the Reuters Institute in 2017, 24% of respondents from around the world said that search engines are their main gateway to news, compared to 23% for social media [2]. And Google is the dominant search engine, handling 63% of all search queries in the United States in April 2018 according to Comscore [3].

Google is also a powerful force within the news economy. According to the Parse.ly media referrer dashboard in August 2018, 50% of external traffic (and 22.4% of overall traffic) to online publishers was referred by Google search, and another 25% from Facebook [4]. That impact is consistently growing. According to Chartbeat, traffic from Google to publishers has increased more than 25% from January 2017 to February 2018, mostly as a result of mobile search [5]. Van Aelst et al. [6] pointed out that media concentration is increasing, in part because of market pressures. As a significant referrer of internet traffic, Google is one of these market pressures.

News curation algorithms can have important implications because of how ranking impacts attention. Users both click more often and believe a result is more relevant if it is in a higher position [7]. As a result, search engines can affect users' attitudes, shape opinions, alter perceptions or reinforce stereotypes, and impact how voters come to be informed during elections [8, 9, 10, 11]. An experiment of 2,150 people using mock search results during the 2014 Indian elections indicated that 24.5% of undecided voters could be swayed by biased rankings in search results [12]. Search algorithms thus play a key role in how people are exposed to information and may develop robust and diverse viewpoints on societally relevant issues, having deep political ramifications [13].

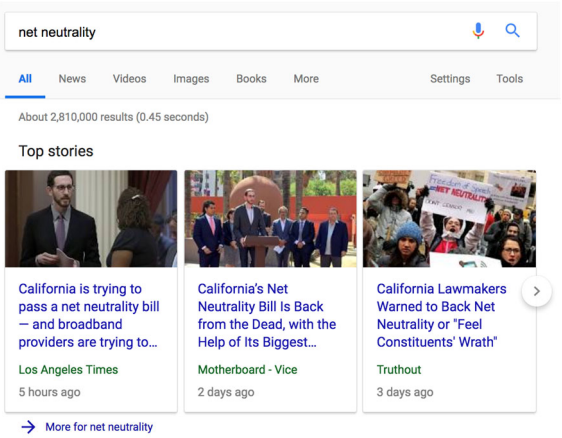

Figure 1: The Google Top Stories Component shown for a query of "net neutrality".

But for all the societal importance and impact that Google and its algorithms have for news media and news audiences, little is known about what drives its algorithm to select and curate the sources of news information that it does. How does the dominant search engine shape attention to news information? Does it provide a diverse sampling of sources and ideological perspectives in the news it curates? What editorial criteria drive its selections? This work grapples with these questions and sheds light on the algorithmic curation of news in Google Search. In particular we focus on the issue of news source diversity, but also delve into how this translates into perspective diversity, and finally what diversity in curation on Google means for the attention that people ultimately pay to the news.

To address these questions, we undertake an algorithm audit [14, 15, 16] focusing on the Google Top Stories box (See Figure 1). This box is a component of search engine results pages (SERPs) that highlights news articles with headlines, images, and links. In previous audit work, this component was found in the first position in about 30% of search queries [17], underscoring its attention-attracting prominence for many queries. We audit the results of the Top Stories box for almost 200 queries relating to news events over the course of a month in late 2017.

We contribute a series of novel analyses that provide an in-depth description and characterization of news source diversity and its implications for attention via Google search. Our results show that just twenty news sources account for more than half of the article impressions in the Top Stories component. Source diversity is query specific, however, with some queries reflecting a fair degree of diversity and others suffering from a severe concentration. We observed a left-leaning ideological skew in Google's selection of sources, only slightly exacerbating the background baseline of media we measured. Analysis of the time stamps of articles indicates Google's algorithmic news curation has a strong predilection towards news articles less than 24 hours old, reinforcing the traditional daily news cycle. Finally, by combining our observations of Top Stories with data from Chartbeat about referrals from Google, we quantify the relationship between the presence (and position) of a news article on Google and how that translates into a referral and attention for that news article. Our results and contributions underscore the power that Google wields in exposing users to diverse news information, raising important questions and opportunities for future work on algorithmic news curation systems.

2 BACKGROUND

In examining the role search engines play in directing attention to news information this work builds on ideas in three key areas of related work: editorial values in algorithms, media diversity as a key editorial value, and algorithm auditing as a technique to illuminate editorial values in search algorithms.

2.1 Editorial Values in Algorithms

In journalism studies editorial news values have traditionally referred to the criteria used by journalists to assess the newsworthiness of content and select what is published or gains prominence in a publication medium [18]. Contemporary journalism practices may consider criteria like recency, conflict, unexpectedness, relevance, proximity, and social impact, among others [19]. Yet journalists are not the only ones exercising editorial judgement in today's media environment. Algorithmic curators play an increasingly important role in the flows of information that reach news consumers [20]. The algorithmic application of editorial news values determines what is included, excluded, highlighted, or de-emphasized in an information display, such as a search results page, social media feed, or news recommendation app [21, 22].

Recent research has explored the encoding of journalistic news values in algorithms that support news content production [23, 24, 25, 26, 27]. One example related to curation is Park et al. [24], which developed a set of algorithmically applied editorial criteria, such as article and conversational relevance, that were useful to news comment moderators in identifying high-quality comments. However, literature on the study of editorial values in more highly automated algorithmic curation systems is more limited. An exception in this vein studied the editorial values that are apparent in the Facebook

Newsfeed [28]. Analysis of intellectual property filings and press releases indicated a set of nine criteria that influence inclusion in the feed, which at times diverge from traditional journalistic news values. These include the importance of friend relationships, explicit and implicit user interests, prior engagement, post age, content quality and so on.

In this research we aim to help fill a gap in research by studying the expression of editorial values in algorithmic curation of news, in particular on search engine result pages. By analyzing news articles to which Google orients audience attention via its Top Stories box, we investigate what type of news information is privileged by the algorithm, and what editorial values are apparent. Specifically, we leverage an auditing methodology to provide observational data on two key editorial dimensions: diversity (detailed more next) and timeliness.

2.2 Media Diversity

Media diversity can be defined and measured in numerous ways, such as the demographics of those working in a newsroom, the political viewpoints presented, the plurality of sources available to news consumers, or the ownership of those sources [29]. One typology for media diversity distinguishes *source diversity* (i.e. of news organizations, their demographic constitution, and economic structure), *content diversity* (i.e. of perspectives, viewpoints, or ideas presented), and *exposure diversity* (i.e. whether audiences actually consume a diverse array of content) [30]. While scholars have warned that source diversity does not necessarily imply content diversity or exposure diversity [30, 31] it is reasonable to assume that in general more source diversity should encourage more content diversity, if not necessarily more exposure diversity. In this research we focus on source diversity, including both the identity and ideological position of news sources, while acknowledging that future research should more directly examine the content and exposure diversity of news as mediated through algorithmic curation.

The importance of media diversity for news audiences hinges on different philosophical conceptions of democratic society [32, 33]. Lack of media diversity could make it difficult to discover new perspectives or ideas thus limiting the quality of arguments, it could curtail users' autonomy of information selection, or it could stifle the awareness needed to contest an idea or issue [34, 35]. McQuail & Van Cuilenburg [34] express the benefit of diversity to society as a whole, defining diversity as "the free expression of alternative goals and solutions to problems. The more the alternatives, the better the prospects for individual and collective welfare." Empirical work has found a correlation between diversity of media exposure and reception to diverse ideas [35].

A common concern surrounding algorithmic mediation of news relates to how exposure diversity could be diminished by filter bubbles that reinforce the tendency for individuals to be selectively exposed to attitude-confirming information and reinforce partisan polarization [36, 37]. However, recent empirical results have downplayed the effect that algorithmic curation might have in an individual's exposure to diverse news [38, 39, 40, 41]. When users are isolated from content they do not agree with, that may be more a function of their individual choices than algorithmic effects [42].

Yet, despite limited empirical support for filter bubbles that create *individual* cocoons, algorithmic news curation still represents a concern for source diversity since it can concentrate societal attention on a narrow range of privileged outlets. For instance, studies of Google News have found that it over-represents some news publishers while under-representing other highly-frequented outlets [41, 43, 44]. Recent survey results [45] show that users of search engines for news reported exposure to a greater number of news sources in comparison to those who did not use search engines, but this is only in relation to the top 30 most popular outlets in each locale surveyed. In other words, search engines may diversify exposure to a point, but are ultimately limited by the small set of largely mainstream sources they present [44]. Society-wide over-reliance on content from a small number of sources can still undermine content and exposure diversity, something we specifically examine in this research in terms of the concentration of news sources in search results.

2.3 Search Engine Audits

One methodological approach towards better understanding the social and political influence of search engines is to systematically observe their results under a range of conditions; to audit them [14, 15, 16]. Previous audits of search results have examined a range of issues including the degree of personalization to location, demographic profile, or preferences [17, 46, 47], the geographic origin of results [48], the degree of state imposed censorship [49], the information quality related to the inclusion of "fake news" sites [50], the presence of representational issues like gender bias in images [11, 51] and commercial anti-competitive tendencies stemming from preferential treatment of some sources over others [52]. These studies expose a range of methodological challenges in search engine audits that we consider,

including the choice of search terms, language settings, geolocation, search history, and logged-in status [53]. While some recent studies have emphasized ecological validity by leveraging plugins which scrape search results from real users' results [17, 54, 55, 56], here we opt for the control, consistency, and scale afforded by automated means of scraping results.

While some previous audit research on search engines has examined news information as one component of search results [8. 17, 46, 56], here we focus exclusively on the news information conveyed via search results pages in order to uncover editorial values that may be at play. The question we pose is: How do search engines shape the availability and consumption of news media, particularly in regards to the share of attention they provide to specific news sources? The next section describes our methods for approaching this question in more detail.

3 DATA COLLECTION METHODS

In order to collect data on news article impressions in Google search's Top Stories box, a method was developed to identify the most relevant news stories each day, determine relevant search queries for those news stories, and then use those queries in automated Google searches. Results for those automated searches were then scraped and analyzed.

3.1 Selecting Stories to Track

News stories for each day in the sampling period were selected using Google Trends. The website contains a ranking of "Stories trending now", which can be filtered by geographical area and are based on a random unbiased sample of Google search data[1]. This ranking lists collections of "stories" that are generated by Google based on news articles and search trends indicating jumps in search traffic. As the user clicks on one of those stories, they are directed to a page about that particular story, top news articles about it, a search trend timeline, a map with the search interest by region, and a list of "Trending queries", that is, terms that were used by web users and were related to that story. The ranking of "Stories trending now" typically has more than 200 stories, from a wide range of topics, such as politics, economy, international news, society, crime, celebrities, and sports.

As previously stated, we are interested in the wide societal impact of the Google Top Stories box curation. These are classically defined as "hard news" by journalism practice and scholarship. We focus on hard news because the public interest value for access to diverse information about hard news is greater than for other types of soft news relating to topics like entertainment, sports, or celebrity.

Reinemann, et al [57] base the classification of hard and soft news on three dimensions: topics / events, focus, and style. The most important, in their view, is the topic dimension: "the extent to which the content of a news item deals with norms, goals, interests, and activities related to the preparation, assertion, and implementation of authoritative, generally binding decisions about societal conflicts."

Therefore, we operationalized this definition of "hard news" in our data collection procedure by selecting stories from Google Trends that had broad societal impact or were covered as such. On one end of the spectrum, stories about government, elections, global affairs, and macroeconomics have clear societal impact. On the other end, stories about the personal lives of celebrities, individual sports achievements, or stock exchange performance of a company had no widespread societal impact. In the middle, and dependent on more subjective evaluation, were stories that had personal and individual agents as part of stories that had societal impact. For instance, when an athlete or a team was involved in a political debate, when a celebrity was involved in a sexual abuse scandal, or when an individual company generates a controversy that has industry-wide effects. These stories have societal impact and, therefore, are considered hard news. Sports results, celebrities getting married or divorced, and updates on individual companies that have no wider impact are not considered hard news.

To select hard news stories, Google Trends was visited every day at 11am CST. One of the authors went through the list of trending topics each day and applied the definition of hard and soft news (societal versus individual issues) to select only hard news topics. Due to constraints in scraping and handling data we chose to limit the number of stories tracked to 30 per day. The next step was selecting the terms to search about each story.

3.2 Selecting Query Terms

As previously mentioned, Google Trends itself provides a list of "Trending queries" for each story; terms that were used by web users and were related to that story, ranked by popularity of use. The appropriateness and accuracy of queries varies from story to story however. Trending queries do not necessarily reflect the news story trending at the moment. Our approach was to select the highest-

[1] https://support.google.com/trends/answer/4365533

ranking relevant search query for each story. The terms on the list were examined one by one to see if they had relevance to the story that was being tracked. We checked the appropriateness of all search terms at the time of their selection, searching them in incognito mode to see if the Top Stories articles they yielded were related to the stories we selected. If they were relevant, they were manually selected; if they were not, they were skipped in favor of the next available and relevant term.

A total of 224 search terms (188 unique terms, with some of them repeated for more than one collection cycle) were selected. Only one search term was selected per trending news story. Some terms are not unique because a news story might have had new developments in subsequent days. A list of all search terms and their context is available at https://goo.gl/M18K77.

3.3 Collecting Search Results

Query terms about the trending stories were selected every day between 11am and 12pm. Selected query terms were stored in a database, with dates and times for starting and ending collection cycles for each term. Data for each search term was collected once per minute and each term was used to scrape results for 24 hours, starting at 12pm and ending at 12pm the following day. The collection included the 3 articles featured in the Top Stories box (Google has since increased the number of links to 10, however there are still only 3 visible initially without interaction), including their URL, title, time information, and position within the box (i.e. whether it was the link on the left, the middle, or the right). The scraper was implemented to read the day's search terms and collect data on them for that 24-hour cycle. Data was collected from October 30 to November 30, 2017. An interruption caused by lack of server capacity led to a gap in collection between the evening of November 19 and 12 pm on November 20.

To minimize personalization, automated searches were made using a desktop browser configured with no user history, without being logged-in, and with language set to English. One remaining source of personalization could have come from server location, in this case Ohio. However, previous work shows that location personalization impacts mostly localized services (such as "airport" and "pizza"), and has a significantly smaller impact on more general terms, such as controversial topics and names of politicians [47]. Searches were run on the main google.com domain, reflecting search results tailored to the U.S. [48]. Across 224 search terms (188 unique), 6,302 links to news articles were collected from the Top Stories box. Each day we scraped 7.1 terms on average (M=7, min=2, max=13).

4 RESULTS

This work seeks to shed light on how the Google search engine shapes the availability of news media, particularly in regards to the share of attention they provide to particular news sources. To do that, we analyze the data through different perspectives. First, we examine the diversity of news sources in the Google Top Stories box; then, we investigate the diversity of ideological leaning of the articles surfaced; next, we investigate if there is a preference for articles in relation to their age; and finally, we assess the relationship between the appearance of articles in the Google Top Stories box and the volume of referrals to news sources' websites.

4.1 Diversity of News Sources

In this section we examine the diversity of news sources observed in the Top Stories box by looking at their distribution and concentration. In particular we consider overall source diversity as well as source diversity by query term. To measure source diversity, we define an *impression* as the appearance of a link in the Top Stories box and then aggregate news article impressions by their root domain. Subdomains that belong to the same news organization (i.e. money.cnn.com) were aggregated to their root domains (cnn.com) so that the measurement is representative of entire news organizations.

4.1.1 Overall Diversity

The top 20.0% of news sources (136 of 678) account for 86.0% of all impressions; and 52.1% of impressions go to the top 20 news sources (See Figure 2). The top three, CNN, The New York Times, and The Washington Post, account for 23.0% of impressions observed. One reason some sources

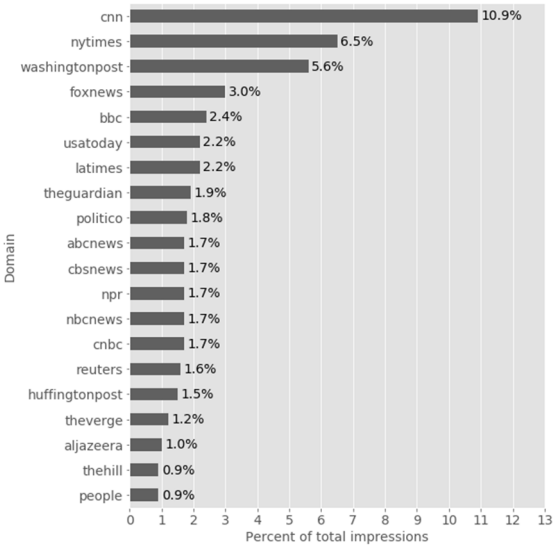

Figure 2: Top 20 in share of impressions

may have more overall impressions than others is because they have articles selected by Google across a greater number of topics or search queries. For instance, links from CNN are present in 106 of the 188 search terms used in the data collection, whereas 313 of the sources were observed in only one query (many of which are local news sources). To control for this variance, we compute the number of impressions for each source normalized by the number of query terms in which the source was observed.

When ranked by the average number of impressions per search term where the source was observed (and filter news sources found in fewer than ten search queries so as to maintain focus on major publishers), CNN and New York Times are still at the top, but other sources shift positions and drop out of the top 20 ranking (See Figure 3). The five news sources that are in the top 20 in total impressions but not in the top 20 for average impressions per search term are: ABC News, CBS News, HuffPost, NBC News, and The Hill. These sources get selected by the platform for a wide variety of topics (i.e. queries), but individually per topic they are not as successful in garnering impressions as some other sources. On the other hand, five news sources that were not in the top 20 for total impressions but are in the top 20 for average impressions per search term include: Deadline, Forbes, Independent, QZ, and Wired. This means that while they are not among the overall largest impression-getting news organizations in the Google Top Stories box, they are very successful for the topics they do get picked up on. These five reflect news niches where the source may be particularly authoritative or apt in their coverage; Deadline for entertainment news, Forbes and QZ for business, Wired for technology, and The Independent as a reflection of UK news as a niche within the American news environment.

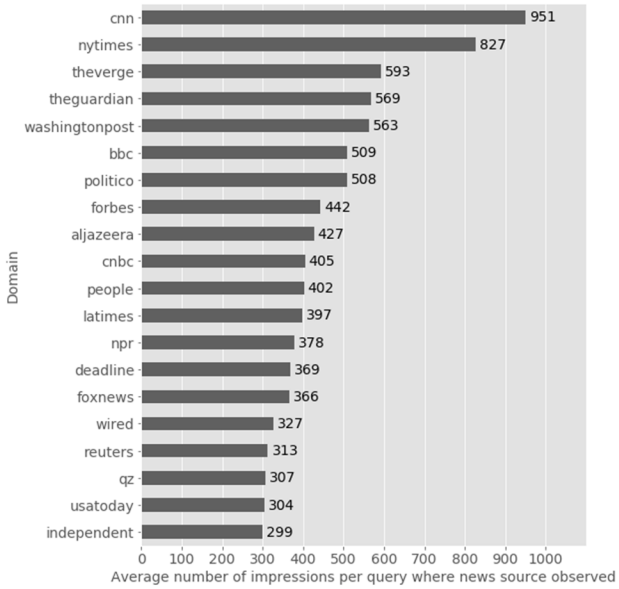

Figure 3: Top 20 sources in average of impressions per query, among news sources with ten or more queries.

Without the filter of news sources that were present in at least ten search queries, all the top 20 domains change, and the ones with the higher average impression per search term in which they appear are local news organizations, with the exception of the United Nations. That happens because these news sources are covering search terms that are about their niche news – either local news, or specific news, such as "United Nations". They have a high average impression per search term because they have less competition in those terms (being present in more impressions) while having a small denominator, since they are present in only one or two search terms.

In sum, although a source may be picked up for many topics, it does not imply they are a consistently dominant source: they may simply have a lower number of impressions aggregated across topics; conversely, a news source that is picked up for a small niche of topics can have a high number of impressions in the terms in which they are present. These results suggest that source diversity should be evaluated within individual search terms, which we turn to next.

4.1.2 Diversity Across Search Terms

Each search term tracked had an average of 19.0 news sources selected by Google (SD = 17.4; M = 14.5). Out of the 188 news queries tracked, 57 (30.3%) were covered by ten or fewer sources. The query with the largest number of news sources selected was "Thanksgiving" (here encompassing all the hard news elements of the holiday, e.g. traffic reports), with 159 sources receiving Google impressions; the query with the fewest number of sources

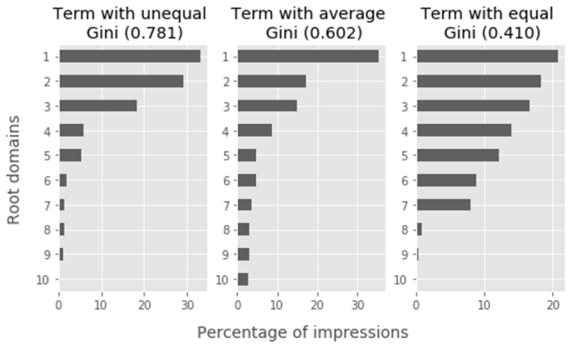

Figure 4: Examples of terms and their distribution of impressions across domains, according to their Gini

covering it was "united nations", with only one source, the United Nations website itself. While most queries drew on more than a dozen sources, some had far more and others still had virtually no diversity at all.

To compare queries by how concentrated the distribution of impressions is across sources, we utilize the Gini coefficient. In this context, the Gini coefficient measures, for each query, how unequal the distribution of impressions is across different sources. The Gini coefficient ranges from 0 (most equal) to 1 (most unequal). A term that has a high Gini coefficient (more unequal) has a high number of impressions concentrated in a few domains, and few impressions distributed across many domains. Terms with lower Gini coefficient (more equal) still can have some domains with more impressions and some with fewer, but that distribution is more even (Figure 4).

Across the 188 unique search terms tracked, 162 were trending in only one cycle, while the other 26 appeared in more than one cycle. We calculated the Gini index of just those terms that appeared in one cycle in order to avoid distortions caused by different collection periods. For those 162 terms, the average Gini coefficient is 0.580 (SD = 0.177; M = 0.610). The distribution of impressions is predominantly less equal with respect to the Gini coefficient: 41.4% of search terms have a Gini coefficient below the mean, while 58.6% are above the mean (See Figure 5).

The search term with the highest inequality in source impressions (See Table 1) is "Rex Tillerson" (then U.S. Secretary of State who was rumored to be replaced by President Donald Trump). Although 38 news sources appear for that search term, two sources are responsible for 75.2% of the 4,296 impression it has: New York Times (41.3%) and CNN (33.9%).

On the opposite end of the spectrum, five search terms had a Gini coefficient of zero, meaning they have an even

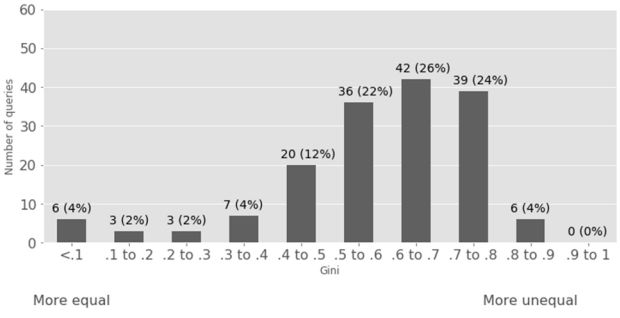

Figure 5: Distribution of Gini coefficients for individual search terms, calculated based on distribution of impressions across sources for each term.

Search term	Gini Coefficient	# News Sources
rex tillerson	0.849	38
time person of the year	0.824	21
russell simmons	0.817	25
zimbabwe news	0.815	44
halloween	0.814	69

Table 1. The five terms with the highest Gini coefficient

distribution of impressions across root domains that cover them: "corelogic" (a company that released a house price index), "big bear billings mt" (a city that had an active shooter situation on that day), "humboldt tn" (regarding a new food factory announced in the city), "sammamish" (city in Washington state that recorded a crime that day), and "united nations". Aside from the previously mentioned "united nations", the other search terms had three sources covering each, and each of these sources had a third of impressions across each of those key terms, that is, they had the same news sources at every cycle of collection. Notably, though all these terms were trending nationally according to Google Trends, they are all related to local or highly specialized news. Of these four search terms, two only had local news as sources ("big bear billings mt" and "sammamish") and two had a mix of local and finance news websites ("corelogic" and "humboldt tn"). When considering terms that had the median number of news sources covering them (i.e., 12 or 13), the most equal search query is “libya slave trade", with a Gini coefficient of 0.431 (See Table 2).

4.2 Ideological Diversity

A key motivator for content diversity is the desire to provide a range of perspectives and viewpoints, such as across different ideological or political positions. Although

Search term	Gini Coefficient	# News Sources
libya slave trade	0.431	12
athens	0.494	12
kristina cohen	0.53	13
milo yiannopoulos	0.534	13
texas trooper killed	0.547	13
mar a lago	0.576	13
sophia robot	0.586	12
texas state university	0.593	13
marissa mayer	0.612	13
john boehner	0.646	12
boston dynamics	0.694	13
pyramid	0.723	13

Table 2. Gini Coefficients for the 12 search terms with a median number of news sources.

we do not measure content diversity directly, in this section we examine the overall ideological position (i.e. left or right-lean) of sources as a proxy for content diversity across different political viewpoints.

4.2.1 Overall Ideological Diversity

To measure the ideology of news sources we considered several alternatives. For instance, Pew has surveyed media consumers, their ideological leanings, and their trust in news organizations to determine what news are preferred by which political group [58]. AllSides uses surveys and community feedback to generate a Media Bias Rating (www.allsides.com/media-bias/about-bias). While valid, these studies introduce a confound related to the *perception* of bias; they are self-report data rather than observational data. Another alternative is to classify news organizations by the types of words or phrases used in article text. Gentzkow & Shapiro [59] tagged news organizations based on phrases that are most commonly used by left or right-leaning politicians. However, their study only covers local and regional newspapers, leaving out websites that, in our data, represent a large share of impressions.

We decided to use ratings data published in Bakshy, Messing & Adamic [42] indicating the ideological alignment of the top 500 most-shared news organizations on Facebook. These ratings were calculated from 10.1 million Facebook users in the U.S. by comparing the sources of news they share with their stated political affiliation. The ratings do not measure the slant of the media outlet itself, but the alignment of preference for sharing content among left/liberal and right/conservative users. Each news organization has a score that ranges from -1 (more left/liberal) to 1 (more right/conservative). An advantage of using these political alignment ratings is that they cover a large proportion of top shared websites while not relying on evaluation of each of them. Since the method consists of observing actual content shared by Facebook users and the users' stated political affiliation, the dataset is representative of users' actual exposure and sharing of news sources, as opposed to self-reported perceptions of media bias. One caveat comes from recent findings that demonstrate that link sharing is more complex than agreement or disagreement [60]. However, sharing still demonstrates preferential attention for a source, even if it does not always imply agreement.

To calculate the ideological slant of the Google Top Stories box impressions, we aggregated news sources by subdomain, as opposed to root domain. That is because the data provided Bakshy, Messing & Adamic [42] had detected different ideological alignments within root domains (e.g. money.cnn.com has a more conservative alignment than cnn.com). In total, there are 727 subdomains in our dataset of which 187 are covered by the dataset in Bakshy, Messing & Adamic [42]. While we only have data for 187 of 727 domains (25.7%) this covers an outsized proportion (74.1%) of impressions observed. From the 187 subdomains for which the ideology is known, 139 have an ideological score of less than zero, meaning they are left/liberal leaning sources; and the other 48 have a score of more than zero, representing right/conservative leaning sources.

For just the 187 subdomains with ideology ratings, the average ideology weighted by the proportion of impressions observed in our data is -0.24, indicating an overall trend towards impressions of left/liberal sources (we compare this to a baseline of media later in this section). If we also include in that calculation the presence of subdomains for which the ideology is not known, and consider them as missing values, the proportionally weighted average ideological lean is -0.16. From all the impressions of the 727 subdomains, 62.4% have a left/liberal slant and 11.3% have a right/conservative slant (See Figure 6). Among the 10 domains with most impressions, only one (Fox News) leans conservative.

The remaining 26.3% of impressions have no ideological data. These impressions come from 540 news organizations (M=120 impressions each). The largest subdomain with an unknown ideology score is ESPN, a sports news site, with 8,416 impressions (0.9% of the total). However, having ideological data for these sources would not substantively change the result. If the 26.3% unknown impressions were split equally by ideology, the percentages of impressions would be 75.5% left/liberal and 24.5% right/conservative. Even if we imagine that all unknown impressions are right/conservative, the divide would still break in favor of left/liberal, at 62.4% versus 37.6%. In short, even in the most drastic scenario, the split would change the scale, but not the direction of the results.

The observed bias of impressions towards left-leaning sources may be the result of different mechanisms, such as

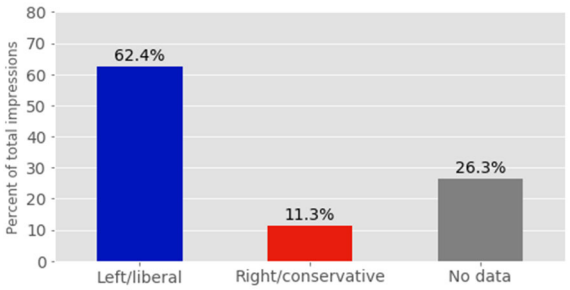

Figure 6: Ideological leaning of impressions. The majority of impressions are from left leaning news sources.

(1) the Google algorithm itself is biased towards selecting left-leaning sources; and (2) there is more left/liberal news content being produced and published online and the observed results simply reflect a greater availability of news content on the left.

While the proprietary nature of the Google Top Stories box precludes an in-depth look at the first possibility, we can examine the second possibility by using a statistical baseline comprised of a wide sampling of news articles published online. To do that we analyzed data from GDELT, a system that monitors and aggregates news output from hundreds of thousands of sources around the world [61]. We used GDELT's DOC 2.0 API to collect data on all the articles that were published for the queries we tracked in our study during the same timeframe. Of the 188 unique search terms for which we have collected data from the Google Top Stories box, we were able to gather GDELT results for 178, (other terms returned no articles in GDELT, either because they were short and therefore not valid for GDELT queries, such as "nfl" or "snl" or they were longer phrases that could be found by Google search but did return results from GDELT such as "san pablo ca" or "texas trooper killed").

The results show that the Google Top Stories box is more left-leaning in comparison to articles collected on GDELT. While on the Google Top Stories box 19.1% of sources are left-leaning, 6.6% are right-leaning and 74.3% are unknown, on GDELT, those proportions are 2.3%, 1.4%, and 96.3%, respectively. The high proportion of unknowns in GDELT is a reflection of its broad coverage of sources in comparison to the more modest number of sites for which we have ideological ratings [42]. If we consider that some sources publish a greater volume of articles than others, then we find a higher proportion of articles of known ideology. We therefore also compare the number of articles in each category. In the Top Stories box, 51.9% of the articles come from left-leaning sources versus 16.2% from right-leaning sources (3.2 times as many left-leaning), and on GDELT, those numbers are 11.0% and 4.9% (2.2 times as many left-leaning). (See Table 3). These results indicate a greater availability of left-leaning sources and articles in the Top Stories box, although the baseline itself already skews left.

	Left	Right	Unknown
Google Sources	139 (19.1%)	48 (6.6%)	540 (74.3%)
GDELT Sources	168 (2.3%)	100 (1.4%)	6,907 (96.3%)
Google Articles	3,305 (51.9%)	1,028 (16.2%)	2,029 (31.9%)
GDELT Articles	178,979 (11.0%)	79,200 (4.9%)	1,371,930 (84.2%)

Table 3: Ideology in GDELT versus in Top Stories box.

4.2.2 Ideological Diversity Across Search Terms

Each search term also has its own characteristic ideological bias in terms of impressions in the Top Stories box. Out of 188 search terms, 161 had a left-leaning average ideological score, and 22 had a right-leaning average ideological score (the other five only had impressions from sources with unknown ideology).

The search query that had the most left-leaning average score (-0.591) was "juan manuel santos", in reference to news about the president of Colombia who announced a record seizure of cocaine (See Table 4). The query with the most right-leaning average (0.399) was "joe scarborough", a journalist and television host who was trending that day because of he had called for the cabinet to remove president Donald Trump from office. Queries that are more partisan also display a greater percentage of impressions that are slanted one way or another: 66.7% of the impressions for "juan manuel santos" came from left-leaning sources, with the other third unknown. On the other hand, "joe scarborough" had 39.8% of impressions coming from right-leaning sources, and 36% from left-leaning sources, with the remaining 24.2% unknown.

search term	Average alignment	Percentage of impressions		
		Left	Right	Unknown
juan manuel santos	-0.591	66.7%	0%	33.3%
san pablo ca	-0.581	33.3%	0%	66.7%
gothamist	-0.563	96.3%	0%	3.7%
safari west	-0.542	79.4%	0%	20.6%
word of faith fellowship	-0.527	66.6%	0%	33.4%

search term	Average alignment	Percentage of impressions		
		Left	Right	Unknown
joe scarborough	0.399	36%	39.8%	24.2%
denise young smith	0.365	0%	16.4%	83.6%
nancy pelosi	0.337	74.3%	13.9%	11.8%
isis new york	0.262	12.3%	33.2%	54.6%
kaepernick	0.145	26.4%	13.6%	60.1%

Table 4: Terms with most biased ideology of impressions; Most liberal (top), most conservative (bottom).

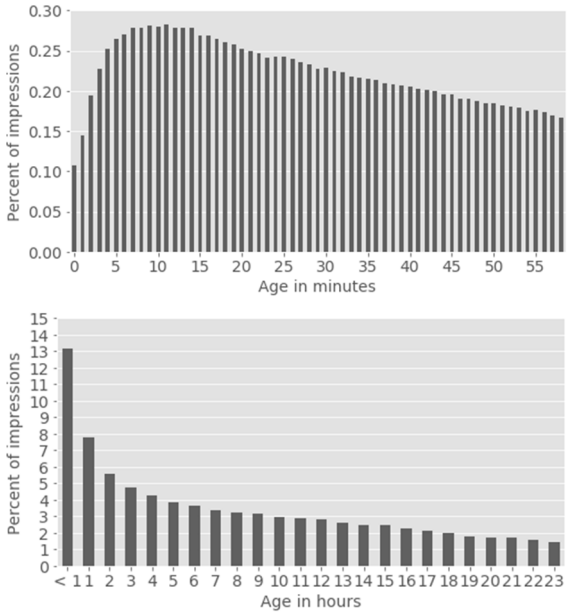

Figure 7: Impressions based on age of the article across different time scales.

4.3 Article Timeliness

So far, results have been focusing on describing matters of source diversity. But the data collected can also provide insight into how Google's search algorithm considers timeliness in its curation of news articles regardless of source. Timeliness is an important editorial criterion, particularly in breaking news scenarios, but also more broadly as it dictates how quickly an algorithmic curation system will churn through content (even if it's still relevant to an ongoing issue).

The Top Stories box provides the approximate age of articles to users in the interface, indicating how many minutes, hours, days, or weeks ago an article was published or updated. Using this information, we calculate the distribution of impressions across the age of articles.

The sample collected shows that Google impressions tend to concentrate on articles that are more recent in age (See Figure 7). Out of 927,494 impressions collected, 83.5% were for articles that were less than 24 hours old. This included 70.4% of articles between one and 24 hours old and 13.1% less than one hour old. Another 16.5% were more than one day old. The distribution of impressions ramps up until the age of the articles gets to 10 minutes where it plateaus until an age of 15 minutes, and then steadily decreases again for older articles.

A closer analysis shows some distinctions among articles that have different ages. Among articles that are recent, such as the ones that are up to one hour old, the articles with more impressions are the classic breaking news types, such as natural disasters, crimes, or political developments (CNN's "Indonesia volcano: Mount Agung eruption closes Bali's main airport", or "Eight arrested in protests as Milo Yiannopoulos speaks at Cal State Fullerton" by the LA Times).

The most successful articles that are a few days old, on the other hand, are stories that have a longer time frame of interest, such as the ones about expected events in the near future, such as a rocket launch from Wallops Island, in Virginia ("ISS resupply mission launching from Wallops Island on Saturday", by local NBC affiliate WAVY), or news about incoming holidays ("The best and worst times to drive and fly this Thanksgiving", by The Atlanta Journal-Constitution).

For articles that are weeks old but still get impressions, many are news articles that describe background or contextual information that remains relevant for an extended period of time, such as news articles from specific topics that did not get coverage from a wide variety of sources. An example of that is in the search term "Ohio cultivation license", which was covered by only seven sources. Cleveland.com's November 5, 2017 article on the topic, titled "First round of Ohio medical marijuana grow license winners announced", appeared in the Top Stories box four weeks after it was initially published.

4.4 News Curation and Attention

Whether we consider source diversity, content diversity as a reflection of source ideology, or other editorial criteria like timeliness, a question that remains open is how the curatorial decisions of search algorithms like Google are presented to end-users and result in users paying attention to various news items. To take one example, a query could have a Gini index of zero, indicating that impressions were equally apportioned to each of three sources. But the position of results on the page could interact with perception and cognition such that one of those sources still receives more attention [62]. The aim of this section of the analysis is to examine how article impressions in the Top Stories box, including their position in the box, convert to actual user clicks and exposure to those news sources.

In order to study this issue, we make use of data provided by Chartbeat, a news analytics provider which tracks the traffic for their clients' articles, including where that traffic comes from (i.e. it's referrer) and how engaged users are on the page once they get there. They do this by embedding a bit of JavaScript code on the news publisher's page that

	Model 1		Model 2	
Variable	β	P>|z|	β	P>|z|
Organic Imp	0.0015	$< 10^{-3}$		
TS Imp	0.0017	$< 10^{-3}$		
Organic Imp			0.0015	$< 10^{-3}$
Rank 1 TS Imp			0.0014	$< 10^{-3}$
Rank 2 TS Imp			0.0024	$< 10^{-3}$
Rank 3 TS Imp			0.0011	$< 10^{-2}$

Table 5. Results of negative binomial regression for two different models with the dependent variable as the number of referrals. All independent variables are significant and the magnitude and sign of the β coefficient is representative of the variable's effect.

sends pings of data back for storage. For each article link observed in our data set, Chartbeat provided timestamped data indicating the referrer for all of the traffic to that article for the 24-hour period in which we were tracking the query in which that article was observed. Across the 188 unique search queries monitored, Chartbeat had referral data for 41.9% of the articles observed (2,639 of 6,302).

We further filtered the data provided to only include data coming from google.com as referrer. While this excludes data from other Google services such as Google News, this filter is not able to differentiate traffic from the Top Stories box and traffic from organic search results. This is an important caveat to the following results, since referral data will overestimate the amount of traffic due solely to an article's placement in Top Stories. In the modeling we describe later in this section we also make use of data we collected about the appearance of news articles in the main organic search results pages, which allows us to consider impressions in Top Stories as well as impressions in organic results in accounting for referral volume. Another filter we apply to the Chartbeat data is to only consider referrals that occurred after the first observation of the article in the Top Stories box. This allows us to further isolate the impact of impressions in the Top Stories box on referral volume.

The relationship between the number of impressions in the Top Stories box and the number of referrals from Google can be calculated by dividing the latter by the former for each article link. This provides an *impression conversion ratio* (ICR) indicating the degree to which impressions are associated with article referrals. Because different search terms will have different degrees of interest and numbers of people searching those terms we calculate the average of the ICR for articles within each search term. Of the 188 search queries, 117 had more than five articles with Chartbeat data, allowing for some degree of aggregation in ICR values. The average ICR calculated this way is 287.1 referrals per impression (SD = 627.7; M= 67.6). As indicated by the high standard deviation, the ICR varies substantially between search queries, reflecting the long-tailed range of different magnitudes of attention searching for different topics. While 58% of search terms had an average of less than 100 referrals per impression, 8.5% have more than 1,000 referrals per impression. One search term, "matt lauer" (in the wake of revelations that the television presenter had sexual harassment complaints against him) had 3,961 referrals per impression.

In order to better understand the relationship between impressions and referrals we use a regression model with top stories impressions and organic impressions as the independent variables and referrals as the dependent variable. The number of referrals is an overdispersed count variable, therefore we use a negative binomial regression model. Because the magnitude of attention flowing to different search terms varies so widely we group articles by search term in the model by using a dummy variable coded for each group. For this modeling we use the entire set of 2,676 articles across 179 queries that had any Chartbeat data (37 articles appeared in more than one query but were treated as distinct). We build two models, Model 1 takes the total count of impressions for each article observed in top stories and in organic results as independent variables, whereas Model 2 takes the count of top stories impressions for each article broken down by position (i.e. rank 1 is the left-most, rank 2 is the middle item, and rank 3 is the right-most item in the Top Stories carousel – See Figure 1), as well as the total count of organic impressions. Model 2 contains the same information as Model 1 and is not meant to be an improvement but rather a more fine-grained assessment of the impact of the position of impressions on the page. By including the counts of organic impressions in the models we are able to isolate the impact of those impressions distinctly from impressions in Top Stories.

Results for the regressions are shown in Table 5. The regression coefficients (β) indicate the effect of the independent variables on the dependent variable. In a negative binomial model the β is interpreted as the log of the ratio of how much the response variable is expected to change for a one unit change in the predictor variable. The result of a likelihood ratio test comparing each model to the null model ($p << 10^{-3}$ for both) indicates they are appropriate to characterize the effects of the independent variables. For Model 1, the β of 0.0017 for Top Stories (TS)

impressions indicates that 1 additional impression on the Top Stories box is associated with 0.17% more referrals. Note however that 1 impression in our data is measured each minute. Therefore, an article receiving 60 top stories impressions in an hour might be expected to receive 10.7% more referrals. Organic impressions account for a shade less in terms of referrals. An article receiving 60 organic impressions in an hour might be expected to receive 9.4% more impressions. An article may of course be receiving impressions both organically and via the top stories box, and from a number of search terms, leading to an even greater boost in referral volume.

Taking the position of the impressions into account in Model 2, we see that the positioning associated with the largest boost in referrals is the middle article in the carousel (rank 2), followed by the left-most (rank 1), and then the right-most (rank 3). An impression in rank 1 is worth about 27% more referrals than an impression in rank 3 and an impression in rank 2 is worth about 71% more than an impression in rank 1. An article receiving 60 impressions in rank 2 over the course of an hour would be expected to receive 15.5% more referrals than if it didn't receive those impressions. Organic impressions have the same weight as in Model 1 and, in fact, rank 1 and rank 3 impressions in top stories are less powerful than organic impressions for driving referrals. Rank 2 impressions in top stories are however about 60% more effective in driving referrals than organic impressions.

5 DISCUSSION

This study has shown that the Google search algorithm provides different degrees of attention to different news sources. Overall, some sources are selected more often than others for the Top Stories box. Results indicate that a majority of impressions went to only 20 news sources, all of which can be considered mainstream, national news outlets. Two sources, CNN and NYT, accounted for 17.4% of the impressions observed. These results are both consistent with and more extreme than previous audits [12] indicating the dominance of CNN and NYT. These two sources dominate when considering the breadth of their coverage as well as within individual query terms. Still, it remains unclear whether the dominance of particular sources is a result of successful strategic behavior by sources to achieve "algorithmic recognizability" [63], or from emergent biases based on the signals the Google algorithm attends to in producing a ranking of news sources.

Our results underscore the degree to which source diversity varies between individual queries. Some sources do fare better in niche areas. But almost a third of queries tracked had 10 or fewer sources, and almost 12% of queries had Gini coefficients more than 1 standard deviation above the mean Gini coefficient. In some cases, such as the query "rex tillerson", CNN and NYT accounted for three-quarters of the impressions observed. Other politically-oriented queries, such as those relating to "russell simmons" and "zimbabwe news" also had high Gini coefficients. These results indicate that for some queries of public importance and social consequence the lack of source diversity can be quite extreme. An implication is that future audits of news on search engines should very clearly motivate the importance and reason for tracking particular queries, ideally in terms of human information seeking behavior.

Our analysis also considered the ideological slant in the distribution of impressions observed. Results showed that Google Top Stories box impressions tend to have a more left-leaning than right-leaning inclination. This can also vary by search term, with 161 terms having an overall left-leaning score and 22 having a right-learning score. It is important to note that a baseline of news content on the internet provided by GDELT showed that a left-leaning slant is the general tendency. Some pieces of news, such as related to the "gothamist" query that received no impressions from right-leaning sources, may simply receive very little to no coverage from the right to begin with, making it difficult or impossible for Google to make diverse selections of sources. In comparison to the GDELT results, the Google Top Stories box may be slightly increasing the disparity between left and right sources (e.g. from 2.2 times as many left-leaning articles than right-leaning in the GDELT baseline to 3.2 times as many in the Google Top Stories box), though the underlying issue in source diversity appears to hinge on a greater availability of news material on the left.

We further show that the Google search algorithm embodies other editorial values that impact the availability and attention to news sources. One of the editorial values observed is a predilection towards recency. By privileging articles that are more recent, the algorithm reflects the journalistic value of timeliness in the way content and news sources are selected. News organizations that have the potential to generate fresh iterations of content may be better positioned to gain impressions from the platform, privileging larger and more well-resourced news organizations that also prioritize timeliness (e.g. CNN). Where a greater diversity of sources is desired, Google may consider relaxing the timeliness constraint to widen the

scope of sources available to its curation algorithm. High-quality journalism can often have a longer shelf-life with respect to user attention, suggesting that news curation algorithms may benefit from dynamically considering the timeliness of selections [64].

Finally, though an analysis of our impressions data in combination with Chartbeat's referrer data, we showed that articles receiving impressions in the Top Stories box (or organically) do tend to receive greater numbers of referrals from Google, and that the positioning of articles within the Top Stories box matters to the volume of referrals received. Impressions in Google's Top Stories box are consequential to source diversity because they do convert to real and substantial amounts of user attention. In order to get closer to addressing the end-goal of exposure diversity algorithmic curators such as Google may consider the relationship between source diversity and placement on the page as well as perhaps other presentation factors linked to reading patterns like size and image use and selection [30].

When viewed through the lens of the economic health and competitiveness of the larger news ecosystem, our results further speak to the implications that powerful algorithmic curators like Google have for mediating attention to news information. We found that publishers that are selected for inclusion in the Top Stories box receive a significant boost in traffic, up to about a sixth more if optimally positioned for just an hour. Because of the importance of Google in referring traffic to news sites, less source diversity implies the unequal capture of economic benefits, such as advertising revenue or the ability to convert users to subscribers, with potential to impact the vitality of the media landscape [29]. None of the top 20 sources in terms of impressions could be considered "local outlets", raising questions about the relationship of source diversity to larger issues in the media landscape like the decline of local news and the appearance of local news deserts [65]. Future work on media diversity should more closely consider the economic ramifications of the ability of algorithmic intermediaries like Google to pick winners and losers based on the large flows of attention they direct. Algorithmic curators will need to make careful tradeoffs between what's desirable for individuals (e.g. relevance), what's desirable for society (e.g. diversity), and what's desirable for news organizations (e.g. fairness in a competitive environment).

5.1 Limitations and Future Work

The results presented in this paper are necessarily constrained by the queries we observed as external auditors of Google results. The terms we selected to track are only a sample of the newsworthy topics (i.e. those related to hard news) that users may have searched for in the timeframe of study. A key limitation of our sample is that it relies on Google Trends, which likely has the effect of skewing the terms we observe towards the more popular. Future approaches to auditing could consider generating search terms through user-center methods, such as via surveys.

Another limitation of the data collected for this study is that it only captures desktop results, despite the growing use of mobile search. Future research may establish comparisons between source diversity in mobile versus desktop scenarios. An additional source of complexity not covered by our study is the potential variation across countries. This study was focused on the U.S. version of Google, with U.S.-centric search terms. It is unclear if the results would be the same throughout the world, and how different rates of media availability in different national contexts may impact results [48].

There are many avenues of continuing investigation. For instance, we did not tailor the data collection to focus on analysis of local versus national news, but we did observe underrepresentation of local news outlets in the overall results, suggesting there may be a rich area of future work in that direction. Additionally, although we focused on the Top Stories box, due to its prominence on many results pages, future work could examine news in the main "organic" Google results more closely. Given the limitations of our scraping method, another possible direction is to utilize data collection plugins [17, 54] to investigate whether personalization has an impact on the news sources found on Google. Future work may also consider computational methods for classifying source and article ideology in order to provide a more comprehensive baseline. And finally, a longitudinal analysis is warranted to determine how the diversity of sources may be changing over time.

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation Grant, award IIS-1717330. The authors would like to thank Chartbeat for their collaboration, particularly Christopher Breaux, Director of Data Science, and Jeiran Jahani, Senior Research Data Scientist. We would also like to thank the reviewers for their thoughtful input.

REFERENCES

[1] Elisa Shearer and Jeffrey Gottfried. 2017. *News use across social media platforms 2017.* Pew Research Center. Retrieved from http://www.journalism.org/2017/09/07/news-use-across-social-media-platforms-2017/

[2] Nic Newman, Richard Fletcher, Antonis Kalogeropoulos, David A. L. Levy, and Rasmus Kleis Nielsen. 2018. *Reuters Institute Digital News*

Report 2018. Reuters Institute. Retrieved from http://media.digitalnewsreport.org/wp-content/uploads/2018/06/digital-news-report-2018.pdf?x89475

[3] 2018. Share of search queries handled by leading U.S. search engine providers as of July 2018. *Search Engines & SEO*. Retrieved from https://www.statista.com/statistics/267161/market-share-of-search-engines-in-the-united-states/

[4] 2018. Explore traffic source trends for digital publishers. Retrieved from https://www.parse.ly/resources/data-studies/referrer-dashboard/

[5] John Saroff. 2018. Google referrals are up: Why that's good and how to make the most of it. Retrieved from https://digitalcontentnext.org/blog/2018/02/14/google-referrals-thats-good-make/

[6] Peter Van Aelst, Jesper Strömbäck, Toril Aalberg, Frank Esser, Claes de Vreese, Jörg Matthes, David Hopmann, Susana Salgado, Nicolas Hubé, Agnieszka Stępińska, Stylianos Papathanassopoulos, Rosa Berganza, Guido Legnante, Carsten Reinemann, Tamir Sheafer, and James Stanyer. 2017. Political communication in a high-choice media environment: a challenge for democracy? *Annals of the International Communication Association* 41, 1: 3–27. https://doi.org/10.1080/23808985.2017.1288551

[7] Bing Pan, Helene Hembrooke, Thorsten Joachims, Lori Lorigo, Geri Gay, and Laura Granka. 2007. In Google We Trust: Users' Decisions on Rank, Position, and Relevance. *Journal of Computer-Mediated Communication* 12, 3: 801–823. https://doi.org/10.1111/j.1083-6101.2007.00351.x

[8] Nicholas Diakopoulos, Daniel Trielli, Jennifer Stark, and Sean Mussenden. 2018. I Vote For—How Search Informs Our Choice of Candidate. In *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple*, Martin Moore and Damian Tambini (eds.). Oxford University Press.

[9] Silvia Knobloch-Westerwick, Benjamin K. Johnson, Nathaniel A. Silver, and Axel Westerwick. 2015. Science Exemplars in the Eye of the Beholder. *Science Communication* 37, 5: 575–601. https://doi.org/10.1177/1075547015596367

[10] Robert Epstein. 2018. Manipulating Minds: the Power of Search Engines to Influence Votes and Opinions. In *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple*, Martin Moore and Damian Tambini (eds.). Oxford University Press.

[11] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 3819–3828. https://doi.org/10.1145/2702123.2702520

[12] Robert Epstein and Ronald E. Robertson. 2015. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *National Academy of Sciences* 112, 33: E4512–E4521. https://doi.org/10.1073/pnas.1419828112

[13] Tarleton Gillespie. 2014. The relevance of algorithms. *Media technologies: Essays on communication, materiality, and society*, 167.

[14] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on Internet platforms.

[15] Nicholas Diakopoulos. 2015. Algorithmic Accountability: Journalistic Investigation of Computational Power Structures. *Digital Journalism*. 3, 3. https://doi.org/10.1080/21670811.2014.976411

[16] Rob Kitchin. 2016. Thinking critically about and researching algorithms. *Information, Communication & Society*, 20(1), 14–29. https://doi.org/10.1080/1369118X.2016.1154087

[17] Ronald E. Robertson, David Lazer, and Christo Wilson. 2018. Auditing the Personalization and Composition of Politically-Related Search Engine Results Pages. In *Proceedings of the 2018 World Wide Web Conference*, 955–965. https://doi.org/10.1145/3178876.3186143

[18] Bob Franklin, Martin Hamer, Mark Hanna, Marie Kinsey, and John E. Richardson. 2005. *Key concepts in journalism studies*. SAGE Publications Ltd.

[19] Tony Harcup and Deirdre O'Neill. 2017. What is news? *Journalism Studies* 18, 12: 1470–1488. https://doi.org/10.1080/1461670X.2016.1150193

[20] Kjerstin Thorson and Chris Wells. 2016. Curated Flows: A Framework for Mapping Media Exposure in the Digital Age. *Communication Theory* 26, 3: 309–328. https://doi.org/10.1111/comt.12087

[21] Matthew S. Weber and Allie Kosterich. 2017. Coding the News. *Digital Journalism* 6, 3: 310–329. https://doi.org/10.1080/21670811.2017.1366865

[22] Taina Bucher. 2012. Want to Be on the Top? Algorithmic Power and the Threat of Invisibility on Facebook. *New Media & Society* 14 (7): 1164–80. https://doi.org/10.1177%2F1461444812440159

[23] Måns Magnusson, Jens Finnäs, and Leonard Wallentin. 2016. Finding the news lead in the data haystack: Automated local data journalism using crime data. In *Computation + Journalism Symposium*.

[24] Deokgun Park, Simranjit Sachar, Nicholas Diakopoulos, and Niklas Elmqvist. 2016. Supporting Comment Moderators in Identifying High Quality Online News Comments. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 1114–1125. https://doi.org/10.1145/2858036.2858389

[25] Alejandro Montes-García, Jose María Álvarez-Rodríguez, Jose Emilio Labra-Gayo, and Marcos Martínez-Merino. 2013. Towards a journalist-based news recommendation system: The Wesomender approach. *Expert Systems with Applications* 40, 17: 6735–6741. https://doi.org/10.1016/j.eswa.2013.06.032

[26] Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Sameena Shah, Robert Martin, and John Duprey. 2017. Reuters Tracer: Toward Automated News Production Using Large Scale Social Media Data.

[27] William Perrin. 2017. Local News Engine: Can the machine help spot diamonds in the dust? In *Data Journalism Past, Present, Future*, John Mair, Richard Lance Keeble and Megan Lucero (eds.). Abramis academic publishing.

[28] Michael A. DeVito. 2016. From Editors to Algorithms. *Digital Journalism* 5, 6: 753–773. https://doi.org/10.1080/21670811.2016.1178592

[29] Natali Helberger. 2018. Challenging Diversity— Social Media Platforms and a New Conception of Media Diversity. In *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple*, Martin Moore and Damian Tambini (eds.). Oxford University Press.

[30] Philip M. Napoli. 2011. Exposure Diversity Reconsidered. *Journal of Information Policy* 1: 246–259. https://doi.org/10.5325/jinfopoli.1.2011.0246

[31] Paul S. Voakes, Jack Kapfer, David Kurpius, and David Shano-Yeon Chern. 1996. Diversity in the News: A Conceptual and Methodological Framework. *Journalism & Mass Communication Quarterly* 73, 3: 582–593. https://doi.org/10.1177/107769909607300306

[32] Engin Bozdag. 2013. Bias in algorithmic filtering and personalization. *Ethics and Information Technology* 15, 3: 209–227. https://doi.org/10.1007/s10676-013-9321-6

[33] Natali Helberger, Kari Karppinen, and Lucia D'Acunto. 2018. Exposure diversity as a design principle for recommender systems. *Information, Communication & Society* 21, 2: 191–207. https://doi.org/10.1080/1369118X.2016.1271900

[34] Denis McQuail and Jan J. Van Cuilenburg. 1983. Diversity as a Media Policy Goal: a Strategy for Evaluative Research and a Netherlands Case Study. *International Communication Gazette* 31, 3: 145–162. https://doi.org/10.1177/001654928303100301

[35] Richard van der Wurff. 2011. Do audiences receive diverse ideas from news media? Exposure to a variety of news media and personal characteristics as determinants of diversity as received. *European Journal of Communication* 26, 4: 328–342. https://doi.org/10.1177/0267323111423377

[36] Eli Pariser. 2011. The filter bubble: How the New Personalized Web Is Changing What We Read and How We Think. Penguin Press.

[37] Natalie Jomini Stroud. 2010. Polarization and partisan selective exposure. *Journal of Communication* 60, 3: 556–576. https://doi.org/10.1111/j.1460-2466.2010.01497.x

[38] Michael A. Beam. 2013. Automating the News. *Communication Research* 41, 8: 1019–1041. https://doi.org/10.1177/0093650213497979

[39] Richard Fletcher and Rasmus Kleis Nielsen. 2017. Are people incidentally exposed to news on social media? A comparative analysis. *New Media & Society* 20, 7: 2450–2468. https://doi.org/10.1177/1461444817724170

[40] Elizabeth Dubois and Grant Blank. 2018. The echo chamber is overstated: the moderating effect of political interest and diverse media. *Information, Communication & Society* 21, 5: 729–745. https://doi.org/10.1080/1369118x.2018.1428656

[41] Mario Haim, Andreas Graefe, and Hans-Bernd Brosius. 2018. Burst of the Filter Bubble? *Digital Journalism* 6, 3: 330–343. https://doi.org/10.1080/21670811.2017.1338145

[42] Eytan Bakshy, Solomon Messing, and Lada A. Adamic. 2015. Political science. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239: 1130–2. https://doi.org/10.1126/science.aaa1160

[43] Roland Schroeder and Moritz Kralemann. 2005. Journalism Ex Machina — Google News Germany and Its News Selection Processes. *Journalism Studies* 6, 2: 245–247. https://doi.org/10.1080/14616700500057486

[44] Efrat Nechushtai and Seth C. Lewis. What kind of news gatekeepers do we want machines to be? Filter bubbles, fragmentation, and the normative dimensions of algorithmic recommendations. *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2018.07.043

[45] Nic Newman and Richard Fletcher. 2018. Platform Reliance, Information Intermediaries and News Diversity. In *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple*, Martin Moore and Damian Tambini (eds.). Oxford University Press.

[46] Chloe Kliman-Silver, Anikó Hánnak, David Lazer, Christo Wilson, and Alan Mislove. 2015. Location, location, location: The impact of geolocation on web search personalization. In *Proceedings of the 2015 Internet Measurement Conference*, 121–127. https://dl.acm.org/citation.cfm?doid=2815675.2815714

[47] Anikó Hannák, Piotr Sapiezynski, Arash Molavi Khaki, David Lazer, Alan Mislove, and Christo Wilson. 2017. Measuring Personalization of Web Search. *arXiv:1706.05011 [cs.CY]*.

[48] Andrea Ballatore, Mark Graham, and Shilad Sen. 2017. Digital Hegemonies: The Localness of Search Engine Results. *Annals of the American Association of Geographers* 107(5): 1194–1215. https://doi.org/10.1080/24694452.2017.1308240

[49] Min Jiang. 2012. The Business and Politics of Search Engines: A Comparative Study of Baidu and Google's Search Results of Internet Events in China. *New Media & Society* 16, 2: 212–233. https://doi.org/10.2139/ssrn.2027436

[50] P. Takis Metaxas and Yada Pruksachatkun. 2017. Manipulation of search engine results during the 2016 US congressional elections. In *Proceedings of the ICIW 2017*.

[51] Gabriel Magno, Camila Souza Araújo, Wagner Meira Jr, and Virgilio Almeida. 2016. Stereotypes in Search Engine Results: Understanding The Role of Local and Global Factors. *arXiv:1609.05413 [cs.CY]*.

[52] Mohammed A. Alam and Doug Downey. 2014. Analyzing the content emphasis of web search engines. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 1083–1086. https://doi.org/10.1145/2600428.2609515

[53] Jacob Ørmen. 2016. Googling the news: Opportunities and challenges in studying news events through Google Search. *Digital Journalism* 4, 1: 107–124. https://doi.org/10.1080/21670811.2015.1093272

[54] Connor McMahon, Isaac L. Johnson, and Brent Hecht. 2017. The Substantial Interdependence of Wikipedia and Google: A Case Study on the Relationship Between Peer Production Communities and Information Technologies. In *Proceedings of the Eleventh International Conference on Web and Social Media*, 142–151.

[55] Cornelius Puschmann. 2017. How significant is algorithmic personalization in searches for political parties and candidates? *Alexander von Humboldt Institute for Internet and Society (HIIG)*. Retrieved from https://www.hiig.de/en/personalized-search-results-elections/

[56] Ronald E. Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. Auditing Partisan Audience Bias within Google Search. *Proceedings of the ACM on Human-Computer Interaction 2* (CSCW): 148. https://doi.org/10.1145/3274417

[57] Carsten Reinemann, James Stanyer, Sebastian Scherr, and Guido Legnante. 2012. Hard and soft news: A review of concepts, operationalizations and key findings. *Journalism* 12, 2: 221–239. https://doi.org/10.1177/1464884911427803

[58] Amy Mitchell, Jeffrey Gottfried, Jocelyn Kiley, and Katerina Eva Matsa. 2014. *Political polarization & media habits*. Pew Research Center. Retrieved from http://www.journalism.org/2014/10/21/political-polarization-media-habits/

[59] Matthew Gentzkow and Jesse M. Shapiro. 2010. What drives media slant? Evidence from US daily newspapers. *Econometrica* 78: 35–71. https://doi.org/10.3982/ECTA7195

[60] Matti Nelimarkka, Salla-Maaria Laaksonen, and Bryan Semaan. 2018. Social Media Is Polarized, Social Media Is Polarized: Towards a New Design Agenda for Mitigating Polarization. In *Proceedings of the 2018 on Designing Interactive Systems Conference*, 957-970. https://doi.org/10.1145/3196709.3196764

[61] Kalev Leetaru and Philip A. Schrodt. 2013. GDELT: Global data on events, location, and tone. Retrieved from http://data.gdeltproject.org/documentation/ISA.2013.GDELT.pdf

[62] Alamir Novin and Eric Meyers. 2017. Making Sense of Conflicting Science Information: Exploring Bias in the Search Engine Result Page. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, 175–184. https://doi.org/10.1145/3020165.3020185

[63] Tarleton Gillespie. 2017. Algorithmically recognizable: Santorum's Google problem, and Google's Santorum problem. *Information, Communication & Society* 20, 1: 63–80. https://doi.org/10.1080/1369118X.2016.1199721

[64] Carlos Castillo, Mohammed El-Haddad, Jürgen Pfeffer, and Matt Stempeck. 2014. Characterizing the life cycle of online news stories using social media reactions. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, 211–223. https://doi.org/10.1145/2531602.2531623

[65] Philip M. Napoli, Matthew Weber, Katie McCollough, and Qun Wang. 2018. *Assessing Local Journalism: News Deserts, Journalism Divides, and the Determinants of the Robustness of Local News*. DeWitt Wallace Center Media & Democracy.