The Thirty-Second International Florida
Artificial Intelligence Research Society Conference (FLAIRS-32)

Opening Up the Black Box: Auditing Google's Top Stories Algorithm

Emma Lurie, Eni Mustafaraj
Department of Computer Science
Wellesley College, Wellesley, MA
elurie,emustafaraj@wellesley.edu

Abstract

Algorithmic auditing has emerged as an important methodology that gleans insights from opaque platform algorithms. These audits often rely on the repeated observations of an algorithm's outputs given a fixed set of inputs. For example, to audit Google search, one repeatedly inputs queries and captures the resulting search pages. Then, the goal is to uncover patterns in the data that reveal the "secrets" of algorithmic decision making. In this paper, we introduce one particular algorithm audit, that of Google's Top stories. We describe the process of data collection, exploration, and analysis for this application and share some of the insights. Concretely, our analysis suggests that Google may be trying to burst the "filter bubble" by choosing less known publishers for the 3rd position in the Top stories. In addition to revealing the behavior of the platform, the audit revealed illustrated that a subset publishers cover certain stories more than others.

Introduction

In the aftermath of the 2016 U.S. presidential election, reports that stories from fake news sources on Facebook received more engagement than news articles from reputable sources (Silverman 2016) led to changes to the Facebook News Feed algorithm. These changes decreased the number of news articles that appeared on News Feed (Isaac 2018), reducing users exposure to news on Facebook. Media reports have suggested that this change may have caused users to rely more heavily on Google to discover news (Moses 2018). But what do users typically find when they search Google for current events? Very often they find "Top stories", a user interface component that displays several news articles (typically ten) in a scrollable carousel as shown in Figure 1. This particular addition to Google's search page was initially introduced in February 2016 on mobile phones (Sterling 2016), and in December 2016 on desktop computers too (Schwartz 2016).

When does the Top stories component appear in the Google search results? There is no academic research that answers this question, and commercial organizations that provide search engine optimization (SEO) services give varying reports. There is evidence that Google displays the Top stories feature when the volume of search results for a particular query suddenly increases or is large, indicating the public's interest in the topic. On at least one occasion, the Top stories panel contained news from unreliable sources (the infamous 4chan board) which was spreading false information about the suspect of the Las Vegas mass shooting in 2017 (Shieber 2017). Given the broad reach of Google search, this component, which appears at the top of the search page, can impact how users perceive a developing story. There are thousands of online news sources on the Web, yet Top stories initially displays only three headlines, so what "editorial" decisions is Google's algorithm making to choose which stories to display? Do some sources appear more frequently than others? The latter question is especially timely because Google is currently facing allegations of anti-conservative bias. Google's response[1] is that its algorithms don't favor any political ideology and that the rankings take into consideration a large number of signals, including the popularity of certain publishers, freshness of content, and relevance to the query. There is legitimate concern from the public about how the new gatekeepers of information (e.g. Google) rank news stories. Despite these concerns, Google refuses to explain how its algorithms work to avoid manipulation by bad faith third-party actors. To resolve this information asymmetry, researchers have introduced a series of auditing methods (Sandvig et

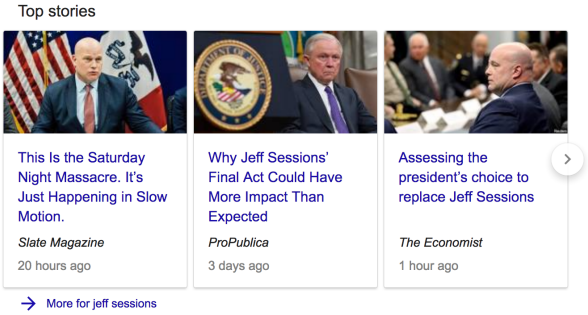

Figure 1: Top stories panel that appeared on Google search for the query "jeff sessions" on Nov 15, 2018 at 12pm.

[1] Google says Trump's bias claims are 'not true', Aug 30, 2018, https://www.bbc.com/news/technology-45354306.

al. 2014), including the scraping audit, which scrapes outputs from platforms like Google and investigates the results. Previous research has revealed the potential power of the "search engine manipulation effect" (SEME) to alter users' perceptions of political candidates based on the ranking of the search results (Epstein and Robertson 2015; Epstein et al. 2017). However, (Robertson, Lazer, and Wilson 2018) found no evidence of Google partaking in partisan manipulation of search results.

We do not aim to measure the SEME in Top stories, but rather are interested in understanding what ranking choices the algorithm makes when certain variables are kept constant. Additionally, we aim to discover what this auditing process tells us about news publishers and their perceptions of which topics and events are newsworthy. In this context, this paper presents the following research contributions:

- A novel audit of the Google's Top stories panel that provides insights into its algorithmic choices for selecting and ranking news publishers (RQ1 & RQ2).
- Evidence about the potential of using audit results from news aggregation platforms (e.g., Google) to answer questions relevant to media communication theory such as media selection bias (e.g., which publishers cover which stories) (RQ3).

Data Collection Methodology

When searching Google, we need to consider that there are many variables that influence the search results. A non-exhaustive list of such variables is discussed in the following:

- **Query phrase:** Early research in search engines logs established that users prefer short queries, between 2-3 words (Silverstein et al. 1999). Often these phrases are stripped of "unneccessary" words and try to convey the intention of the query with as few words as possible. For example, instead of searching for "news about trump", users will simply search for "trump" or "trump news". While the semantic intention behind these phrases is the same, the results can be different. Therefore, when trying to capture search results pages relevant to an event, we should attempt to create an entire set of queries that mimics user query formulation. For example, to create a query set for one of the biggest political stories of September 2018, the confirmation process of U.S Supreme Court judge Brett Kavanaugh, we employed multiple phrases: "kavanaugh", "brett kavanaugh", "kavanaugh hearing", "kavanaugh confirmation", etc.
- **Location:** To generate search results, Google accounts for the location of a user, which is inferred by the IP address of the device. This is easily verifiable with the query phrase "weather", which shows the weather prediction for the current location. However, it is not clear to what extent queries unrelated to location lead to different results (Robertson, Lazer, and Wilson 2018). It is plausible to expect that Google might show local publishers as part of Top stories. We don't explore this question in this current research, but leave it to future work.
- **Personal search history:** Google's algorithms learn over time users' informational preferences and adjust themselves to better adapt to them. This is known as personalization, but another name for it is the "filter bubble", a term popularized by Eli Pariser (Pariser 2011). His explanation of the phenomenon involves an example showing side-by-side screenshots of two searches for Egypt (performed by two of his friends), with one page containing news about the protests in Egypt and the other with no mention of the protests. Although personalization of algorithmic platforms is often regarded as a possible cause for the perceived increase in public's political polarization (Flaxman, Goel, and Rao 2016), previous research on auditing Google search for political results has not been able to verify it (Robertson, Lazer, and Wilson 2018). To avoid any possible effects of personalization, we use a browser without search history for our data collection.
- **Device type:** Mobile phones and laptop computers differ in many ways (internet speed, page loading time, screen size, modes of interaction, etc.), thus, when creating a list of results for different devices, Google takes these factors into account. Additionally, in order for content to appear to participate in Top stories, publishers need to use the Google-introduced publication system AMP[2] (accelerated mobile page), that loads content faster on mobile platforms. According to Google, mobile-friendly content performs better in the ranking for those who search on mobile[3]. Automatically capturing content from mobile phones is challenging, thus, we currently focus our experiment on traditional devices (e.g., a laptop).

Compounding to all this, there are different products for searching news, including Google News[4], the tab "News" on Google search, or Google search itself, that shows the Top stories. Each of these products seems to operate in different ways. Moreover, Google is constantly performing A/B testing of its products, and this affects the results too. For our study, we only focus on the Top stories shown on Google search.

Researchers who audit Google search results (Robertson et al. 2018), typically recruit geographically diverse participants to study how location and personalization influence the search results. In this study, we are interested in the baseline behavior of the algorithm with respect to choosing news sources without the adjustments it makes for location, personalization, or device type. In order to do this, we decide to keep many of the above-mentioned variables constant, that is, we use the same laptop computer, with the same IP address, and the same blank-state browser in incognito mode (no search history).

We use the web browser automation tool Selenium controlled through a Python program, to open a new browser instance every time we perform a search, in order to avoid query session interference. Once the search engine result page (SERP) is loaded, we save it as an HTML file. We keep

[2]https://www.ampproject.org/

[3]https://webmasters.googleblog.com/2018/03/rolling-out-mobile-first-indexing.html

[4]https://news.google.com

Attribute	Value
query	jeff sessions
timestamp	2018-11-15-12pm
title	Why Jeff Sessions' Final Act Could Have More Impact Than Expected
source	ProPublica
url	https://www.propublica.org/article/why-jeff-sessions-final-act-could-have-more-impact-than-expected
moment	3 days ago
position	2

Table 1: An example of the data structure for storing one of the articles shown in Figure 1. The attributes of title, source, url, and moment are extracted from the HTML code, and we automatically add fields for the query, the timestamp of the search, as well as the position of the story in the Top stories panel (assuming 1 for most-left story and up to 10).

Data Type	Amount
Query phrases	48
Observation moments	472
SERPs	18,844
SERPs with Top stories	15,729 (83.5%)
SERPs without Top stories	3,155 (16.5%)
SERPs with 10 stories/panel	11,934 (76%)
SERPs with 3 stories/panel	3,050 (19.4%)
Story observations	132,553
Unique stories	20,256
Unique sources	1,392

Table 2: Statistics about the data collection, before cleaning the field "source". See Data Cleaning section for details.

all the original files (so that we can always verify the results of our automatic scraping). Details about the collected data are described in the following.

Data Collection Timeframe and Statistics

Our data collection is on-going, but for this paper we are using data collected between Sept 6 - Nov 15, 2018. We started with a list of 23 query phrases about politics and over time added new queries to follow breaking news events. For example, we started with some queries about Brett Kavanaugh, but when Christine Blasey Ford came forward with her story,[5] we added new phrases to the list.Similarly, we added queries for the shooting in a Pittsburgh synagogue[6] and the mail bombs send to politicians and media in October 2018[7]. The current analysis contains data for 48 queries, though not all of them have the same number of observations, due to their later addition in our set of queries.

In total, we collected data on 472 occasions, initially 4-10 times per day, and since November 5, 12 times per day every two hours. Occasionally, our data collection failed and we are missing a few observations. In total, we have 18,844 SERPs as HTML files to analyze.

[5]https://en.wikipedia.org/wiki/Christine_Blasey_Ford

[6]https://en.wikipedia.org/wiki/Pittsburgh_synagogue_shooting

[7]https://en.wikipedia.org/wiki/October_2018_United_States_mail_bombing_attempts

Extracting Top Stories

We parsed the HTML files to extract the information in the Top stories panel. Most of the stories are in the format displayed in Figure 1, three visible stories and a total of ten stories accessed by scrolling left. However, there is an alternate format that shows stories in groups of three, not in a carousel. While most Top stories show in either groups of ten and three, a very small percentage (4.5%) have a configuration between 1 to 9 stories. There are also SERPs where the Top stories panel did not appear. We stored each story as a JSON dictionary of keys and values, as shown in Table 1. Following this processing, we calculated the statistics shown in Table 2. While there are 132,553 story observations in our dataset, there are much fewer stories, (20,256), because many stories appear several times in the Top stories panel, especially for less newsworthy events (there are not many sources who cover such stories).

Data Cleaning

The data points in our dataset are structured as shown on Table 1. At first glance, it looks like these data points are well-structured and clean. However, when exploring the "sources" attribute, we noticed that the list of sources contained duplicates. For example, we found both *HuffPost* and *Huffington Post* or *The Hill* and *TheHill*. Given our interest in studying sources, it was important to fix this issue by replacing all duplicates. We did this by extracting the domain name from the URL of each unique story. For example, from the URL in the Table 1, we extract the domain `probublica.org`. This way, the various names of a source are merged together to a single domain name. As a result of this cleaning, the number of sources was reduced by 19%, from 1,392 to 1,125 sources. A noteworthy duplication case was the domain name `cbn.com` that corresponds to *The Christian Broadcasting Network*. This domain was matched to five different names (variations of CBN, CBN.com, CBN News, etc.). In future work, we will examine why publishers use multiple names to sign their stories.

Data Exploration and Analysis

There are many interesting questions we can explore in this dataset, but due to lack of space, we will focus only on a few.

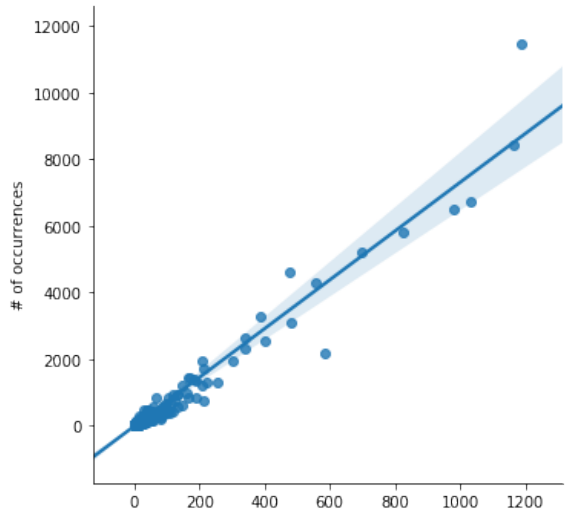

Figure 2: Correlation between number of unique stories by each source and their total number of occurrences in our observations. The clump close to the origin indicates that the majority of sources have 1 or 2 stories.

RQ1: Are all publishers represented equally?

Our dataset contains 1,125 publishers that have at least one story occurring in the Top stories panel during the audit. Some of these publishers are big media corporations like *CNN* and *BBC News*, while others are local newspapers or blogs. Thus, we cannot expect an equal representation of all publishers in the dataset. However, how do big organizations compare to one another? And which are the big organizations when it comes to political events? In Figure 2, we can notice that there is high correlation between the number of unique stories published and the total occurrences in the Top stories observations [r=0.97, n=1125, p=0.0].[8] However, we also notice that the majority of publishers have up to 10 stories (n=950, 84%), with a negligent number of stories (n=2,318, 11%). Meanwhile, 175 publishers (16%) are responsible for the majority of stories (n=17,938, 89%) and the majority of occurrences in our observations, 90%.

Furthermore, an even smaller number of publishers, a total of 11, have each published between 400-1200 stories. These can be easily observed in the upper right part of the graph in Figure 2. In Table 3, we provide statistics for each of these top publishers. In the table, we find some big news organizations, such as *CNN*, *Fox News*, and *The New York Times*, but also a few publishers that are mostly focused on politics, such as *The Hill* and *Politico*. Also worth noticing is that the clearest outlier in the data is *CNN*, which is appearing in the Top stories more often than its produced number of stories would predict. Finally, this table encapsulates the inequality problem in news production: 1% of publishers (11 out of 1,125) produce 41% of total articles and are present in 46% of observations of the Top stories panel.

[8] r: correlation coefficient, n: sample size, p: probability

Publisher	# st.	% st.	# occ.	% occ.
cnn.com	1,188	5.9%	11,473	8.7%
thehill.com	1,165	5.8%	8,438	6.4%
foxnews.com	1,032	5.1%	6,736	5.1%
washingtonpost.com	980	4.8%	6,490	4.9%
nytimes.com	824	4.1%	5,809	4.4%
usatoday.com	696	3.4%	5,205	3.9%
yahoo.com	584	2.9%	2,197	1.7%
nbcnews.com	559	2.8%	4,306	3.2%
huffingtonpost.com	480	2.4%	3,096	2.3%
politico.com	477	2.4%	4,600	3.5%
cnbc.com	403	2.0%	2,533	1.9%
Totals		**41.4%**		**45.9%**

Table 3: The 11 most prolific publishers in our dataset. They have each more than 400 unique stories. Stories (#st) and occurrences (#occ) are correlated, but there are a few outliers too, such as CNN and Yahoo. This table shows how 1% of publishers receives 46% of presence in Top stories.

Position in Top stories	# of sources	% of sources
1st position	376	33.4%
2nd position	534	47.5%
3rd position	774	68.8%

Table 4: Distribution of publishers by position in Top stories (at least one occurrence). Twice as many publishers show in the 3rd position compared to the 1st position.

RQ2: Which publishers show on the top 3 positions?

Research on user behavior on search engines has established that users primarily consider and click on the first two links (Granka, Joachims, and Gay 2004). While such research for Top stories doesn't exist yet, we can hypothesize a user preference for the first 3 positions (which are the only ones visible at first). Therefore, the publishers and stories that appear in these three positions likely impact how users perceive the story. We calculated the number of times each publisher occurred at least once in the 1st, 2nd, and 3rd position in the Top stories panel and the results are summarized in Table 4. Across the 48 queries, we find that only 33.4% of publishers in our dataset occurred at least once in the 1st position of Top stories. The percentages increase for the 2nd and 3rd positions.

It is important to notice that the number of sources in the 3rd position is more than double that of sources in the 1st position. Comparing the lists of sources that show in these three positions, we find that 307 sources (27% of all sources) that appear in 3rd position, never occurred in the 1st or 2nd position. This indicates that the algorithm doesn't treat articles from all sources equally. Some are more likely than others to show in the 1st or 2nd place and others have a lower chance. However, the wide range of sources in the 3rd position leads us to hypothesize that Google's algorithm might be trying to balance the exploration and exploitation dilemma, a strategy used to improve "learning to rank" tasks in information retrieval (Hofmann, Whiteson, and de Rijke 2013), and other reinforcement learning tasks in general.

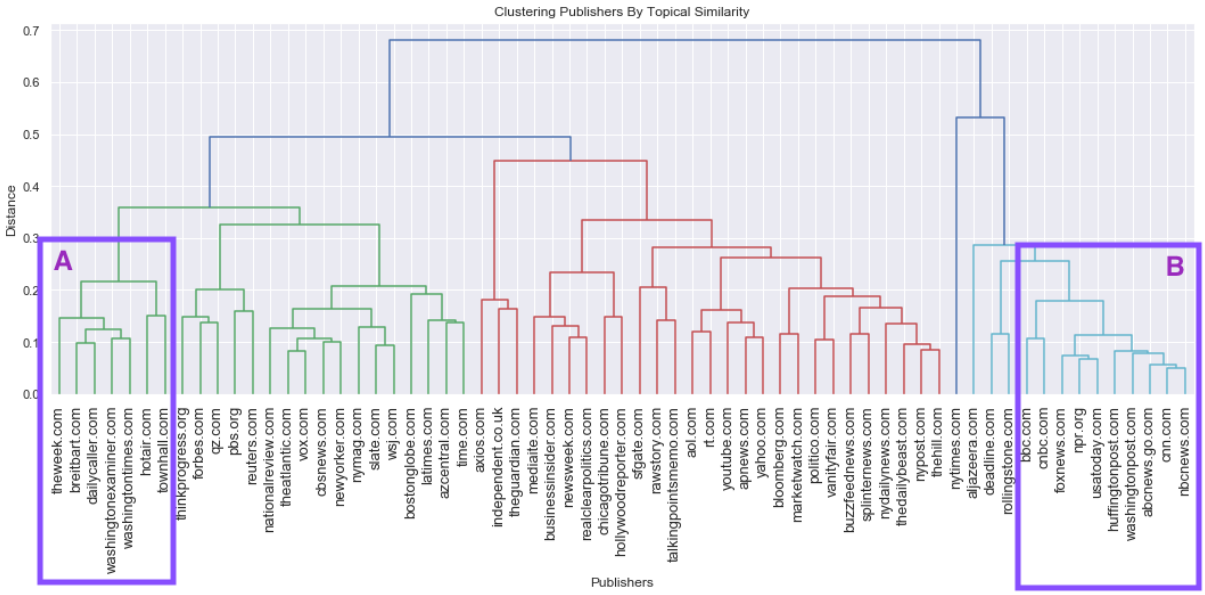

Figure 3: Hierarchical clustering for 65 publishers who have written stories for more than 25 queries in our dataset. Each publisher is represented as a vector of the 48 queries. Cluster A is mostly composed of right-leaning sources. Cluster B groups together the largest mainstream publishers.

Most of the times, the algorithm shows in the first 2 positions popular sources (exploitation), and then uses the 3rd position for exploration, providing users with unfamiliar sources.

The race for "making it" into the Top stories seems to be more competitive when an event is extremely newsworthy. For example, we extracted from the data all stories within the timeframe Sep 5 - Oct 10, 2018, which corresponded to the confirmation process of Brett Kavanaugh. This extraction identified 204 different sources. However, doing the same extraction for the period Oct 11 - Nov 15, 2018, we discovered 241 sources, but 118 of them had never appeared during the confirmation timeframe. It is unlikely that they were not writing about Kavanaugh, but because the major publishers were covering the event non-stop, with continuous updates these 118 sources were not selected by the algorithm.

RQ3: Do publishers prefer certain stories more than others?

In the social sciences, there is an established theory about the selection bias of media (McCarthy, McPhail, and Smith 1996). Given the many events that happen everyday, a news publisher only selects a subset of them to cover. Our data collection, albeit small, might provide some insights into this phenomenon. What events or people are publishers choosing to report on any given day? It is natural to expect that most publishers will cover the major events of the day, however, their selection bias will lead them to cover their preferred topics as well. For example, for the query "hillary clinton", the top sources are *Washington Examiner* and *Fox News* (both, right-leaning sources) that continue to keep Secretary Clinton in the news, despite her current retirement from politics, while for "michelle obama" (a less polarizing political figure), the top sources are *CNN* and *The Hill*, known more as center of left-center leaning sources.

To understand how different publishers behave with respects to different events and political figures, we decided to cluster publishers by representing each a of them as a vector of the queries in our dataset. This representation consists of a 48-dimensional vector, where the values are the proportions of unique stories observed for the source. For example, if a source has a total of 100 stories and 14 from them were about "midterm elections", the corresponding value for this dimension is 0.14. This way, we normalize the data to handle the fact that some organizations are more prolific. Through this representation, we discover that 50% of our sources contain a single direction (meaning, they have occurred for a single query), emphasizing one more time how narrow the pool of news sources is (despite the first impression of a big list of sources). To avoid sparsity and to make the visualization of hierarchical clustering legible, we focus on publishers that have dimension values greater than 0 for at least 25 of their 48 dimensions. This leaves us with 65 publishers, whose hierarchical clustering we depict in Figure 3. There are many small clusters of size two that are recognizable for media experts, such as *Bloomberg* and *Market Watch*.While an explanation of all clusters requires expertise knowledge, two clusters annotated as A and B in the graph are easy to interpret. All but one source in Cluster A are right-leaning sources, in fact they are some of the most highly partisan sources in the entire dendrogram. Meanwhile, cluster B is populated by the mainstream TV and newspapers sources. The fact that *Fox News* is here clustered together with *NPR* (the national public radio) and *USA Today* (the most circu-

lated newspaper in the U.S, because of its centrist positions), indicates how Fox News covers all major stories as much as these two non-partisan sources.

Discussion and Future Work

Auditing algorithms is an emerging research direction that is still trying to establish itself. While there have been other audits of search engine results (Epstein et al. 2017; Robertson, Lazer, and Wilson 2018), and one previous work that also considers Top stories (Robertson et al. 2018), none of them had a longitudinal focus. As the research direction is so new, we are still in experimenting with research design. For example, it is not clear how often we should perform daily audits. Our dataset is currently mixed. For the first eight weeks we only collected data between 8am-10pm, at somewhat random times, as a user will do. However, later we switched to an equidistant regime that captures data every 2 hours. We believe that this will allow us to avoid issues of time zones that might lead to oversampling of U.S sources.

Although our dataset is not perfect, our analysis provided several new insights: a) news coverage as presented in Top stories relies too much on a small number of mainstream publishers, with 1% of sources making up 46% of observations (see RQ1); c) Google's algorithm appears to be attempting to provide a larger group of publishers to select news from in the 3rd position (see RQ2) and c) the media "selection bias" phenomenon is observable through a clustering process of longitudinal data, allowing us to group publishers by their choice of topics (see RQ3).

The (RQ2) insight is especially interesting, because it indicates that Google might be actively trying to pop the so-called "filter bubble". By showing a wide range of unfamiliar sources in the 3rd position, the algorithm is providing users with an opportunity to explore different news sources and break the bubble.

This analysis revealed opportunities for new research questions. For example, Figure 2 indicates that CNN's behavior is an outlier compared to all other news publishers. That is, CNN stories have some particular features that provide them some advantage in Google's Top stories ranking. Such behavior is interesting to investigate in future work. Additionally, we will examine the differences in representation of sources known for partisanship (does the algorithm has a political bias), and explanatory models that predict ranking based on past behavior.

Acknowledgments

We are grateful to contributions from Wellesley Cred Lab members and partial funding from the National Science Foundation, through grant IIS 1751087.

References

Epstein, R., and Robertson, R. E. 2015. The search engine manipulation effect (seme) and its possible impact on the outcomes of elections. *PNAS* 112(33):E4512–E4521.

Epstein, R.; Robertson, R. E.; Lazer, D.; and Wilson, C. 2017. Suppressing the search engine manipulation effect (seme). *Proc. CSCW* 1:42:1–42:22.

Flaxman, S.; Goel, S.; and Rao, J. M. 2016. Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly* 80(S1):298–320.

Granka, L. A.; Joachims, T.; and Gay, G. 2004. Eye-tracking analysis of user behavior in www search. In *Proc. of the 27th ACM SIGIR*, 478–479.

Hofmann, K.; Whiteson, S.; and de Rijke, M. 2013. Balancing exploration and exploitation in listwise and pairwise online learning to rank for information retrieval. *Information Retrieval* 16(1):63–90.

Isaac, M. 2018. Facebook overhauls news feed to focus on what friends and family share. *The New York Times*.

McCarthy, J. D.; McPhail, C.; and Smith, J. 1996. Images of protest: Dimensions of selection bias in media coverage of washington demonstrations, 1982 and 1991. *American sociological review* 478–499.

Moses, L. 2018. As promised, facebook traffic to news publishers declines again, post news-feed change. *Digiday.com*.

Pariser, E. 2011. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin.

Robertson, R. E.; Jiang, S.; Joseph, K.; Friedland, L.; Lazer, D.; and Wilson, C. 2018. Auditing partisan audience bias within google search. *Proceedings of CSCW* 2:148.

Robertson, R. E.; Lazer, D.; and Wilson, C. 2018. Auditing the personalization and composition of politically-related search engine results pages. In *The Web Conference*, 955–965.

Sandvig, C.; Hamilton, K.; Karahalios, K.; and Langbort, C. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* 1–23.

Schwartz, B. 2016. Google replaces 'in the news' box with 'top stories' on desktop. *SearchEngineLand.com*.

Shieber, J. 2017. How reports from 4chan on the las vegas shooting showed up on google top stories. *TechCrunch.com*.

Silverman, C. 2016. This analysis shows how viral fake election news stories outperformed real news on facebook. *BuzzFeed News*.

Silverstein, C.; Marais, H.; Henzinger, M.; and Moricz, M. 1999. Analysis of a very large web search engine query log. In *ACM SIGIR Forum*, volume 33, 6–12. ACM.

Sterling, G. 2016. Google's amp top stories now live in mobile search results. *SearchEngineLand.com*.