Auditing Partisan Audience Bias within Google Search

RONALD E. ROBERTSON, Northeastern University, USA
SHAN JIANG, Northeastern University, USA
KENNETH JOSEPH, Northeastern University, USA
LISA FRIEDLAND, Northeastern University, USA
DAVID LAZER, Northeastern University, USA
CHRISTO WILSON, Northeastern University, USA

There is a growing consensus that online platforms have a systematic influence on the democratic process. However, research beyond social media is limited. In this paper, we report the results of a mixed-methods algorithm audit of partisan audience bias and personalization within Google Search. Following Donald Trump's inauguration, we recruited 187 participants to complete a survey and install a browser extension that enabled us to collect Search Engine Results Pages (SERPs) from their computers. To quantify partisan audience bias, we developed a domain-level score by leveraging the sharing propensities of registered voters on a large Twitter panel. We found little evidence for the "filter bubble" hypothesis. Instead, we found that results positioned toward the bottom of Google SERPs were more left-leaning than results positioned toward the top, and that the direction and magnitude of overall lean varied by search query, component type (e.g. "answer boxes"), and other factors. Utilizing rank-weighted metrics that we adapted from prior work, we also found that Google's rankings shifted the average lean of SERPs to the right of their unweighted average.

CCS Concepts: • **Information systems** → **Page and site ranking**; **Content ranking**; **Personalization**; • **Social and professional topics** → *Political speech*; • **Human-centered computing** → *User interface design*;

Additional Key Words and Phrases: Search engine rankings; quantifying partisan bias; algorithm auditing; political personalization; filter bubble

ACM Reference Format:
Ronald E. Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. Auditing Partisan Audience Bias within Google Search. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 148 (November 2018), 22 pages. https://doi.org/10.1145/3274417

1 INTRODUCTION

In 1928, Edward Bernays, the nephew of famous psychoanalyst Sigmund Freud and the "father of public relations" [93], published *Propaganda*, introducing some of the first ideas on how people's unconscious heuristics could be leveraged to shape public opinion [9]. These ideas were refined in the decades that followed, with researchers positing a one-step, "hypodermic needle," model in which the mass media directly influenced individuals [58], and Katz and Lazarsfeld producing

Authors' addresses: Ronald E. Robertson, rer@ccs.neu.edu, Northeastern University, 1010-177 Network Science Institute, 360 Huntington Ave. Boston, 02115, MA, USA; Shan Jiang, sjiang@ccs.neu.edu, Northeastern University, USA; Kenneth Joseph, k.joseph@northeastern.edu, Northeastern University, USA; Lisa Friedland, l.friedland@northeastern.edu, Northeastern University, USA; David Lazer, d.lazer@neu.edu, Northeastern University, USA; Christo Wilson, cbw@ccs.neu.edu, Northeastern University, USA.

2573-0142/2018/11-ART148 $15.00
https://doi.org/10.1145/3274417

their seminal work on a "two-step" model that noted the importance of information rebroadcasts by influential individuals [54].

Today, one-step models have seen a revival of research interest due to the growing capacities of online platforms to target and personalize content [6, 62, 76], and two-step models have been expanded to study digital traces of behavior spreading on social networks [32, 37, 38]. However, in light of the current debate on the impact that online platforms might have had on the 2016 US election [53], it is now evident that these advances in measuring and shaping human behavior have come with unexpected consequences and challenges for democratic societies [59, 88].

Spurred by these concerns, research on the spread of fake news on social media has flourished [1, 51, 60, 96]. While the spread of misinformation on social media can have serious impacts, such as the massive dip in the stock market that resulted from a fake 2013 tweet claiming Barack Obama had been injured in an explosion [26], these concerns encompass only part of a larger digital information ecosystem. For example, some researchers have pointed out that exposure to fake news on social media accounts for a relatively small amount of news exposure for the average American, and that the real threat to the democratic process is the tendency of the media to favor scandal and sensationalism over substantive discussions of policy [97].

Web search, on the other hand, has been under-explored relative to the dominant role that search engines play in the information ecosystem. Indeed, a total of eight experiments involving over 8,000 subjects has shown that partisan bias in search engine rankings can substantially influence the voting intentions of undecided voters, even when design interventions for suppressing the effect are used [28, 30]. Research on whether such partisan bias actually occurs on search engines is thus crucial [29], but also a difficult task fraught with challenges, such as defining bias in a valid way, and controlling for confounding platform features like personalization [21].

The importance of web search and its symbiotic relationship with the media and politics motivated us to conduct a controlled *algorithm audit* [85] to assess the *partisan audience bias* within Google Search. In early 2017, following Donald Trump's inauguration, we recruited 187 participants to complete a survey and install a custom browser extension that conducted a set of static and dynamic Google Search queries in participants' web browsers. For each query, our extension preserved a Search Engine Result Page (SERP) from both a personalized (standard) and an unpersonalized (incognito) browser window. In total, we obtained over 30,000 SERPs.

To quantify the bias of websites, we developed an audience-based metric to score the partisanship of web domain audiences using a unique dataset containing the sharing propensities of registered Democrats and Republicans on Twitter. To quantify bias in web search, we merged our bias scores with the domains from our search data and adapted metrics recently developed for measuring ranking bias in social media search [57].

Our work makes the following contributions:

- Extending prior work [61], we developed a practical and sustainable partisan audience bias metric for web domains by leveraging the sharing patterns of a digital panel of Twitter users linked to US voter registration records. We found that our measure is strongly correlated with similar bias scores developed in prior and ongoing work [2, 3], as well as with bias scores we collected from human raters.
- Operationalizing our bias scores, we conducted a controlled audit of partisan audience bias within Google Search and found that the results positioned toward the bottom of the SERPs we collected were more left-leaning than the results positioned toward the top, a finding consistent with prior work [21, 29].
- Exploring the interaction between personalization and partisan audience bias, we found little evidence to support the "filter bubble" hypothesis in web search [73].

- We build upon prior work by identifying and exploring partisan bias in a diverse set of 14 Google search ranking *components* and *subcomponents* (*e.g.,* `news-card` and `twitter-card`), and found that their average bias varied by search query and rank.
- Compared to the other components, we found a consistent right-lean in the `twitter-card` components we identified. Given that searches for the query "Donald Trump" almost always returned a `twitter-card` component linking to his Twitter account in our dataset, and that "Donald Trump" was one of Google's top ten searches of 2016 [40], it is possible that the inclusion of these components in Google's SERPs may have provided Donald Trump with a virtuous cycle of amplification during the election.

Outline. We organized the rest of our study as follows. First, we review prior work on quantifying partisan bias and auditing Google Search (§ 2). Next, we describe our auditing methodology, the survey and search data we collected, and the partisan audience bias scores that we developed (§ 3). Then, we explore how our bias metric varies by query, component type, rank, and participants' political preferences (§ 4). Finally, we discuss the implications of our findings and limitations (§ 5).

2 BACKGROUND

The findings from several studies conducted over the past five years have shown that search engines play a pivotal role in the public's access to the digital information ecosystem. For example, a 2017 international survey found that 86% of people use search engines on a daily basis [24], and surveys from Pew (2014) and Reuters (2017) found that more people get their news from search engines than social media [69, 77]. Other findings include that search engines are one of the first places people go to seek out information [23], and the second most common use of the internet (email being first) [25].

When people do obtain information through social media, 74% of participants in a recent international survey reported using search to check the accuracy of that information, and 68% reported that the information they found by searching was "important to influencing their decisions about voting" [24]. The vast majority of internet-using US adults also depend upon search engines to find and fact-check information [24, 25] and place more trust in information found through search engines than through social media [8, 42, 78]. This trust is exemplified by a recent ethnographic study of conservative search engine users, where 100% of participants reported that "doing your own research" started with a Google Search, with one participant believing that Google "works as a fact checker" [94].

To understand how this dependence and trust in search engines might have undesirable effects on democracy, and what we can do to measure those effects, below we review the literature on the impact of partisan bias, methods for quantifying it, and techniques for conducting algorithm audits on search engines.

The Impacts and Origins of Partisan Bias. Partisan bias has been shown to influence voting behaviors through newspapers [17, 22, 35], television (*e.g.,* the "Fox News Effect" [20]), social media [10] (see also "digital gerrymandering" [100]), and search engines (*e.g.,* the "Search Engine Manipulation Effect (SEME)" [28, 30]). Studies of bias in traditional media, such as television, suggest the potential to shift around 10,000 votes [20]. In contrast, an experiment on social media motivated 340,000 people to vote [10], and the SEME experiments found that partisan bias in election-related search rankings can sway the preferences of undecided voters by 20% or more [28].

Of particular concern here, is that, even without intentionally designing bias into a platform like Google Search, machine learning algorithms can still naturally surface and amplify the societal biases reflected in the data they use or the people that contruct them [14, 36, 48]. In the context

of partisan bias, concerns arise regarding personalization technologies placing users in internet "filter bubbles" that entrench their existing political beliefs by limiting exposure to cross-cutting information [5, 7, 73]. These concerns are rooted in a combination of the theory of selective exposure, which posits that people seek information that agrees with their existing beliefs while avoiding information that does not [87], and the personalization technologies that might have a synergistic effect with that human tendency. While scholars have recently suggested that such concerns may be overstated with respect to social media [3, 44], the same may not be true for search engines, where research on the interaction between personalization and partisan bias is limited.

Quantifying Partisan Bias. Generally speaking, partisan bias is a complex, nuanced, and subjective phenomenon that is difficult to quantify in a valid way [12, 71]. The existing methods can be sorted into three primary approaches: audience-based, content-based, and rater-based.

Audience-based measures depend on the social network principle of homophily ("birds of a feather flock together" [46, 64]), to recover the political leaning of a news outlet or web domain from the political affiliations of its audience (*e.g.,* likes and shares on Facebook [3]). Content-based measures leverage linguistic features to measure the differential usage of phrases (*e.g.,* in congressional speeches [34] or debates [50]). Finally, rater-based methods have people rate the sentiment or partisan lean of webpages or text [12, 21, 29].

Compared to rater-based methods, audience-based measures are not hindered by biases among raters or the high cost of gathering manual webpage ratings [12, 29]. Audience-based measures are also conceptually simpler and less computationally expensive than content-based methods, which can produce scores that are limited by the training data used [14, 50, 74, 80, 99].

Recent audience-based metrics of partisan bias have leveraged associations among users' behavior on Facebook [3, 81] and Twitter [57], or the sharing propensities of Twitter users who followed manually identified landmark accounts [61]. However, these approaches are also limited due to their reliance on users' self-reported political affiliation, Facebook's opaque "interested" classification of users, or inference of users' characteristics based on social media behavior.

Auditing Search Engines. Although the algorithms used by search engines are opaque and unavailable to external researchers, *algorithm auditing* methodologies provide a useful tool for systematically assessing their output based on a controlled input [67, 85]. Algorithm audits have previously been used to investigate various aspects of Google Search, including the personalization, stability, and locality of search rankings [45, 56, 63, 66]. Although prior work has audited partisan bias on social media [3, 57], similar research on web search is limited to a small number of recent studies [21, 29, 91].

Most relevant here are a recent book chapter detailing a series of case studies on partisan bias in non-personalized search [21], and a recent white paper on auditing partisan bias using a primarily passive data collection methodology [29]. In the book chapter, Diakopoulos *et al.* examined (1) the degree of support for candidates on the first page of search results, (2) the "Issue Guide" feature that appeared at the top of the search results for queries containing candidate's names during the 2016 election season, (3) the presentation of news about candidates in Google's "In the News" results, and (4) the visual framing of candidates in the image collages that sometimes appear in the side bar of search results [21]. To quantify partisan bias, these researchers collected rater-based sentiment scores of search results, utilized audience-based scores from prior work [3], and operationalized content-based measures to assess coverage bias in political candidate quotations and visual framing in images. Overall, Diakopoulos *et al.* found (1) a higher proportion of negative articles in searches for Republican candidates than Democratic ones, (2) a mean left-leaning partisan bias of -0.16 (using

the -1 (left-lean) to 1 (right-lean) partisan bias scores from [3]) in the sources used in Google's "Issue Guide",[1] (3) that CNN and the New York Times dominated Google's "In the News" results, and (4) that the sources of images in Google's image collages came from left-leaning sources more often than right-leaning sources [21].

In the white paper, Epstein and Robertson utilized a browser extension and a Nielsen-ratings-type network of 95 confidants spread across 24 states to collect Google and Yahoo search data [29]. Their extension passively collected SERPs whenever the confidants conducted a search with a query that contained a word from a manually curated list of election-related keywords [27].[2] Similar to prior work examining the stability of search engine results [66], this study did not differentiate among the various result types that populate Google SERPs [56, 84]. To quantify bias, these researchers used a rater-based approach [12] to score individual web pages, though few specifics about their method or controls for rater bias were reported. Overall, for a 25-day period from October 15 to November 8, 2016, these researchers found (1) a mean bias of 0.17 on a scale of -1 (pro-Trump) to 1 (pro-Clinton), (2) a greater pro-Clinton bias in Google (0.19) than Yahoo (0.09) search results, (3) that the pro-Clinton bias increased towards the bottom of the search rankings, and (4) that the bias decreased as the election approached and passed.

The findings from these two studies agree with prior research on quantifying media bias that found an overall left-lean [43], though the subjectivity of partisanship and the complexity of journalistic motivations limits the causal claims that can be drawn from practically any measure of bias [16, 71]. This is especially true when auditing search engines, where it is difficult to tease apart confounds inherent to the scale and complexity of the web [95], the constantly evolving metrics (*e.g.,* "relevance") that search engines optimize for [21], and the efforts of search engines to prevent gaming by third parties [66]. Similarly, how users formulate and negotiate search queries within web search interfaces is an under-researched topic that involves not only the information retrieval algorithms at play, but also the cultural history of the human interacting with them [21, 94]. The lack of up-to-date research on this topic is due, in part, to the difficulty associated with obtaining real user queries [11], the ever-evolving nature of users' information needs [4], and the opaque interactions between users and autocomplete algorithms that influence the process of query selection [70].

3 METHODOLOGY AND DATA OVERVIEW

In the following section we describe the participants we recruited, the browser extension we designed, the search data we collected, and the bias scores we derived.[3] We provide details on participants' demographics, the composition of the Google SERPs they received during our audit, and provide a rationale, formula, and validation for our partisan audience bias metric.

3.1 Survey and Participants

We utilized two established crowdsourcing services [72, 75] to recruit a demographically diverse sample of 187 participants to take our survey. Using built-in features on each service, we restricted the visibility of our ads to the US and recruited participants once a week during January and February 2017, coinciding with and following Donald Trump's inauguration. At the end of the recruitment phase, we had recruited 74% of our sample through Prolific (http://prolific.ac) and 26% from Crowdflower (http://crowdflower.com).

[1] This mean was obtained with low domain coverage - their scores only matched 50% of the domains that appeared in the Issue Guide. Considering only the top 10 sources that appeared in the Issue Guide, the bias score was -0.31.

[2] Although not mentioned in the whitepaper, we learned from the authors that participants were shown the list of search terms and instructed to conduct at least one search from that list per day in order to remain a participant.

[3] This study was IRB approved (Northeastern IRB #16-11-23) and summary data and code are available at http://personalization.ccs.neu.edu/

As with prior research on crowdsourcing websites [13], the majority of our participants were White (66%) and male (52%), and 44% reported having at least a bachelor's degree. The mean age was 32 (σ = 12.3), participants reported a median household income of $50,000 to $74,999, and our sample leaned Democratic (47%) and liberal (50%). The mean rating of Donald Trump on an 11-point bipolar Likert scale (used in prior research [28, 30] and ranging from *negative* -5 to +5 *positive*) was -2.4 (σ = 3.4), and 22% of participants gave him a "positive" rating on a binary scale (*positive* or *negative*). Compared to Trump's approval rating at the time of our audit (42.3%), our participants' ratings of the US President were somewhat low [33].

In terms of online habits, we asked participants which Alphabet services[4] they were regular users of, defined as using the service at least once a week. The median number of Alphabet products that participants reported regularly using was 4, and Gmail or YouTube were the most popular, with 90% using these services regularly. Consistent with prior research on search usage [18, 30], 88% of participants reported a preference for Google Search, 82% reported a preference for Chrome, and the average number of searches that participants reported conducting per day was 14.2 (σ = 16.9).

3.2 Browser Extension and SERP Collection

To collect search data from within participants' browsers–maintaining the personalization due to their cookies, search history, and account logins–we designed a custom Chrome browser extension that could automatically conduct searches and save the raw HTML. After completing our survey, participants were instructed to install the extension and input a unique token (a pseudonymous ID) that started the searches. Upon providing the token, the extension opened both a standard and an incognito Chrome browser window and began conducting simultaneous searches in each window, saving a standard-incognito SERP pair for each query.[5]

Root Queries and Suggestions. To obtain a diverse set of static and dynamic queries, we seeded our extension with a predefined list of 21 *root queries* covering six topic areas that we intentionally focused around a major political event that coincided with our audit: Donald Trump and his inauguration (Table 1 in Supplementary Material [SM]).[6] For each of these root queries, we collected its Google autocomplete search suggestions and added them to the search queue for SERP collection.[7] As one might expect, we found that suggestions for the same root query changed over time and participants. Across all participants and all dates, we collected 549 unique queries (including our 21 roots), whereas if Google's suggestions were static and uniform, we would only expect 105 unique queries (Table 1 in SM).

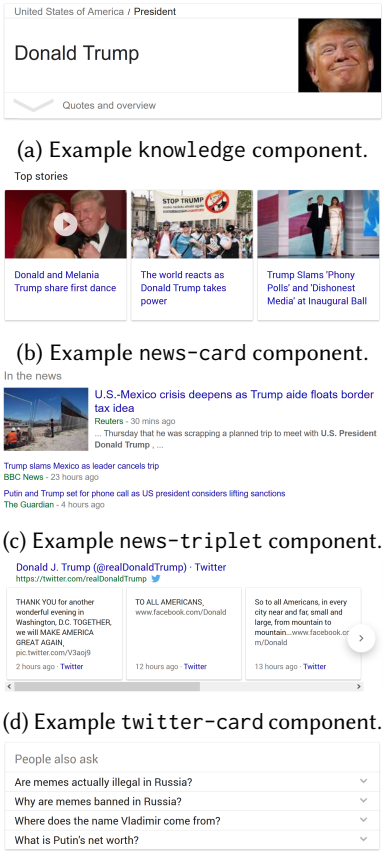
To reduce the ambiguity of the queries we collected and quantify their partisan bias, we relied on the ability of humans to classify ambiguous queries [89]. We obtained crowdsourced ratings for each query, asking workers to (1) classify it as "Political," "Not Political," or "Ambiguous" and (2) if the query was political, rate how favorably it depicted the Democratic and Republican Party on two 5-point Likert scales ranging from "Very Negative" to "Very Positive". We collected two ratings for each query, breaking classification ties manually and calculating the mean bias by combining the

[4]Including Android, Gmail, Calendar, Docs, Drive, Google+, Groups, Maps, and YouTube.

[5]Although Google's documentation is somewhat ambiguous on the details of incognito mode, it generally indicates that the searches conducted from an incognito window are not personalized [41], and that any carryover or location effects [45, 56] would affect both windows equally.

[6]The topics and queries were *US President* ("2017 US President", "US President"), *Inauguration* ("Trump inauguration", "inauguration", "President inauguration"), *Political Party* ("Democrat", "Republican", "Independent"), *Political Ideology* ("liberal", "moderate", "conservative"), *Political Actors* ("Donald", "Trump", "Donald Trump", "Mike", "Pence", "Mike Pence"), and *Foreign Entities* ("China", "Russia", "Putin", "UN"). Our method here was exploratory and we limited our investigation to 21 root queries to avoid triggering Google's rate limits, which would have negatively impacted our participants.

[7]At the time of our audit, Google provided a maximum of four suggestions per query.



(a) Example `knowledge` component.

(b) Example `news-card` component.

(c) Example `news-triplet` component.

(d) Example `twitter-card` component.

(e) Example `people-ask` component.

Fig. 1. Examples of prominent component types.

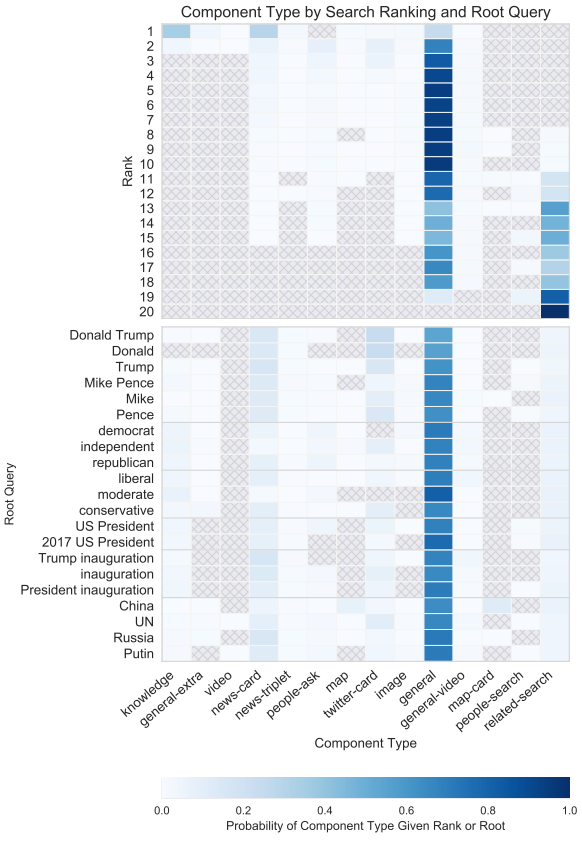


Fig. 2. Probability matrices for observing each component (columns) given the search ranking it appeared at (top) or given its root query (bottom). Gray cells indicate the absence of a component given that rank or root. Rows sum to one.

two bias ratings onto a single scale [12]. Overall, we classified 46.6% of queries as political, 40.8% as not political, and 12.6% as ambiguous (close to the 16% found in prior work [89], see Table 1 in SM).

SERP Composition. At the end of data collection, we had collected 15,337 standard-incognito SERP pairs. We broke each of these SERPs, the first page of results, down into a set of *components*–corresponding to the vertically ranked items in the center column of a Google SERP–and their *subcomponents*–the clusters of similar items that Google groups horizontally (*e.g.,* `news-card` (Figure 1b) and `twitter-card` (Figure 1d) components) or vertically (*e.g.,* `news-triplet` (Figure 1c)). In total, we parsed 456,923 components and subcomponents.

We identified 14 components in our sample of Google SERPs, including the:

- `general` components that consist of the ubiquitous blue title and webpage snippet, the related `general-extra` components that include extra links to domain present in the title link, and `general-video` components that include a video thumbnail.
- `knowledge` components that draw from Google's "Knowledge Graph" and have been previously studied in-depth but in isolation [63] (Figure 1a).

- `news-card` and `news-triplet` components that contain three news stories as subcomponents and that Google labels "Top Stories" and "In the News," respectively (Figures 1b & 1b). Both of these have been studied to some extent in prior work [21, 56].
- `image` components that present a collage of image results pulled from Google's image search feature, some aspects of which have been studied in prior work [21, 55, 91].
- `video` components that embed a YouTube video in the SERP.
- `map` and `map-card` components that present results from a Google Maps search directly in the search results. The `map` component was previously identified in prior work on mobile Google Search [56].
- `people-also-ask` components that present a vertically stacked set of four or five questions that drop down to reveal an answer parsed from the content of various websites.[8]
- `twitter-card` components, which consist of a distinct header that displays a Twitter user's account, or a Twitter search term, in a similar fashion to `general` components. This header is followed by a set of horizontally arranged subcomponents, each containing a tweet (Figure 1d).[9]
- `people-also-search` and `suggested-search` components that are generally found at the footer of the SERP and provide internal links to the search queries they suggest.

The average SERP consisted of 13.5 ($\sigma = 3.4$) components,[10] and there were no significant differences between the SERPs produced by the standard and incognito windows in this respect. The presence or absence of components was related to the root query that produced the SERP, and `knowledge` and `news-card` components were especially prominent in the rankings (Figure 2).

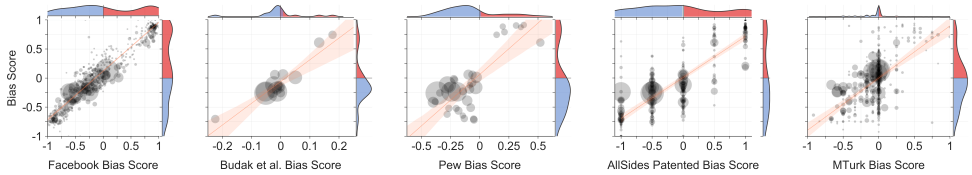
3.3 Partisan Audience Bias Scores

To score partisanship of websites, we developed a *partisan audience bias score* for web domains by leveraging the sharing propensities of a virtual panel of Twitter users. To apply these scores to the SERPs we collected, we adapted several rank-weighted metrics from recent work on measuring partisan bias within social media search [57].

Scoring Method. To construct our scores, we first linked 519,000 Twitter accounts to US voter registration records.[11] Our linking procedure was as follows. For each voter record: (1) we searched the name on Twitter, (2) if accounts using that name were found, we checked the location field of each matching account, and (3) if both the name and location were matched, then we linked the voter record and the Twitter account. If multiple matches were found for the name and location of a given voter's record, we discarded it to avoid false positives. Prior work has used this panel matching method to classify political stance and identify individuals whose online discussions accurately characterized election events [47, 52].

Beginning in October of 2017, we used the Twitter API to download up to the last 3.2K public tweets of each linked account every 15 days, collecting a total of 113 million tweets. Tweets that did not contain a URL were filtered out and are not included in the final count. Before calculating our partisan audience bias score, we reduced each URL to its second-level domain (*e.g.,* http://www.bbc.com/news/business-38686568 was reduced to bbc.com), as done in similar work

[8] We note that URLs were frequently missing in these components, perhaps due to their dynamically loading nature.
[9] We utilize only the URLs in the subcomponents, as the header always links to Twitter.
[10] This number was calculated by counting each component and ignoring it's subcomponents. For example, a `twitter-card` consisting of five subcomponents only counted as one component. Including subcomponents, the mean is 14.9 ($\sigma = 3.3$).
[11] These data contained information on US voters' names, current political party registrations, and location. We limit our focus to the two major US parties: Democrats (left-leaning) and Republicans (right-leaning).



(a) Bakshy *et al.*'s audience-based bias scores [3] ($r = 0.96^{***}$).

(b) Budak *et al.*'s rater-based bias scores [12] ($r = 0.93^{**}$).

(c) Pew's rater-based bias scores [68] ($r = 0.78^{***}$).

(d) AllSides bias scores [2] ($\rho = 0.64^{***}$).

(e) Domain ratings we crowdsourced on MTurk ($\rho = 0.50^{***}$).

Fig. 3. We found high agreement ($r > 0.90$) with peer-reviewed partisan bias metrics (3a, 3b) and moderate agreement ($0.5 < \rho < 0.82$) with the three other methods (3d, 3c, 3e). Point size represents the number of shares a domain had in our data. $^{*}P < 0.05$;** $P < 0.01$;*** $P < 0.001$.

measuring partisan bias on social media [3]. After this coarse-graining, we maintained a set I of 1,019,730 unique domains, and then computed the bias score for each domain $i \in I$ using the following formula:

$$\text{bias-score}(i) = \frac{\frac{r_i}{\sum_{j \in I} r_j} - \frac{d_i}{\sum_{j \in I} d_j}}{\frac{r_i}{\sum_{j \in I} r_j} + \frac{d_i}{\sum_{j \in I} d_j}}, \tag{1}$$

where r_i (respectively d_i) is the count of unique registered Republicans (Democrats) that shared a domain i, and $\sum_{j \in I} r_j$ ($\sum_{j \in I} d_j$) represents the sum of counts over all domains shared by unique registered Republicans (Democrats). The resulting bias score is continuous and scales from -1 to 1, where a score of -1 means that a domain i was shared exclusively by Democrats ($r_i = 0$, $d_i > 0$), and a score of 1 means that a domain i was shared exclusively by Republicans ($r_i > 0$, $d_i = 0$). A domain i has a score of 0 if and only if it was shared by the same proportion of Democrats and Republicans. Finally, to reduce noise and spam–63% of URLs were shared by only one unique Twitter account–we only keep domains shared by 50 or more unique accounts ($r_i + d_i \geq 50$), which reduced the number of domains to 19,022.

Construct Validity. We compared our domain scores to five other sets of bias scores[12] and found the strongest correlation between our scores and the audience-based scores crafted from Facebook data by Bakshy *et al.* in 2015 (Pearson's $r = 0.96^{***}$) [3] (Figure 3a). We found a similarly high correlation between our scores and the rater-based scores carefully collected by Budak, Goel, and Rao ($r = 0.93^{***}$) in 2016 [12] (Figure 3b).[13]

Next, we compared our scores to the crowdsourced domain bias ratings aggregated by the website AllSides [2], a source used in prior examinations of partisan bias [91]. AllSides rates websites on a five-point scale (left, left-lean, center, right-lean, and right) and provides two scores, one based purely on crowdsourced ratings (community score) and one based on their patented technology that incorporates the bias of raters (controlled score). After matching our scores to 200 of the 254 domains rated by AllSides we found a moderate correlation with their controlled score ($\rho = 0.73^{***}$)

[12]We are grateful to Ceren Budak for sharing the data collected in [12].

[13]We used Pearson's r to compare our continuous score with Bakshy *et al.*'s continuous metric and switched to Spearman's ρ for the AllSides and MTurk ratings, which were collected on ordinal scales. Using Spearman's ρ for our comparison with Bakshy *et al.*'s findings had a minimal effect on the correlation ($\rho = 0.95^{***}$). On the other hand, our correlation with Budak *et al.*'s metric, which was measured on an ordinal scale but aggregated by domain onto a continuous scale, was substantially different using a Spearman's test ($\rho = 0.73^{***}$). More details on these correlations are available in SM.

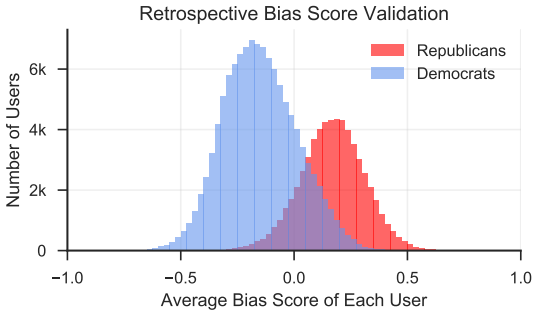


Fig. 4. A comparison of the average bias scores generated for Republican and Democrat panel members. The majority of panel members have an average score that leans in the expected direction, providing retrospective validation for our scoring methodology and demonstrating homophily in the sharing patterns of panel members by their political party registration.

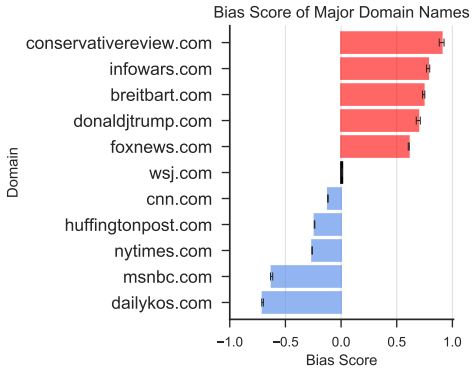


Fig. 5. The mean bias of mainstream media and other popular domains matched expectations, providing face validity for our partisan audience bias score.

and a smaller correlation with their community score ($\rho = 0.64^{***}$) (Figure 3d). We also found a moderate correlation between the number of shares each domain received in our dataset and the number of individuals who contributed to the community scoring of each domain in AllSides data ($r = 0.70^{***}$), suggesting some degree of consistency in terms of website popularity.

We also compared our scores to scores adapted from a 2014 Pew survey on trust in media sources [68]. Following the procedure used in a recent study [81], we multiplied the proportion of self-reported political identities by their respective scores (-1, -0.5, 0, 0.5, 1) to convert the Pew numbers into partisan bias scores and found that they correlated well with our metric ($r = 0.78^{***}$) (Figure 3c). Although the scores developed by Ribeiro *et al.*, who used the Facebook marketing API to gather demographic data on the audiences of news outlets with a Facebook presence, were more correlated with the adapted Pew scores ($r = 0.97$), our measure had stronger correlations with the three other metrics that Ribeiro *et al.* also used for validation (Facebook [3], AllSides [2], and Budak *et al.* (2016) [12]). It is unclear why our scores have a weaker correlation with the adapted Pew scores, but it is also unclear exactly how the adapted trust metric translates to partisan bias.

To further validate our scores, we also collected domain-level ratings from human-raters. After splitting our scored domains into 20 quantiles and taking the top 10 most shared domains in each quantile (400 domains in total), we utilized the same 5-point Likert scales and participant pool (MTurk) as Budak *et al.* to collect three ratings for each domain.[14] Using this rater-based method, we found a relatively weak correlation with our audience-based measure ($\rho = 0.54^{***}$) (Figure 3e), suggesting that collecting rater-based bias at the domain level, while perhaps more practical, may be less accurate than those collected at the article-level and then aggregated post-hoc [12].

Aside from our comparisons to existing bias measures, we also checked the construct validity of our partisan audience bias scores in two additional ways. First, as a form of retrospective validation,

[14]Mirroring Budak *et al.*'s method, we had each domain rated on two 5-point Likert scales (ranging from "Very Negative" [-1] to "Very Positive" [1]), and asked participants to rate how the domains "Depicted the Democratic [Republican] Party." We then combined these scores by inverting the Democratic score and taking the average of the two favorability scores for each domain. We also used most of the same qualification requirements used by Budak *et al.* (over 1,000 HITs, 98% success rate, located in the US), but did not require raters to pass a test of political knowledge. Another important difference is that Budak *et al.* had people rate individual news articles rather than domains, which likely evokes a different evaluation process in raters.

we examined whether the scores of the domains shared by our panel matched expectations. We found that they largely did, with Democrats primarily sharing domains with scores to the left of zero and vice versa (Figure 4). Second, we extracted the bias scores of well-known media websites and they largely matched expectations, providing some degree of face-validity (Figure 5). Of particular note is the Wall Street Journal, which was the only news source that a 2014 Pew study found to be "More trusted than distrusted" by people from across the ideological spectrum [68], and which had a score close to zero in our data (shared by a similar proportion of Democrats and Republicans).

In terms of domains that did not correspond to the mass media, we found that both the desktop and mobile versions of Wikipedia were shared more by Democrats than Republicans (bias scores of -0.22 and -0.10, respectively). Similarly, we found that YouTube and WebMD were shared more by Republicans (bias scores of 0.13 and 0.19, respectively). In these cases, where there may be little or no expectation of political lean, it is possible that our scores pick up on non-obvious latent traits of the domains that surface through preferential sharing patterns.[15]

Although the validation we have presented here enables comparisons between domains (*e.g.,* domain A is farther right than B), our metric does not have a valid zero point because its location is dependent on the methodological decisions we have laid out here. Thus, we caution that scores from our metric should not be interpreted in absolute terms, only relative terms.

Scoring SERPs. To assess the bias of each SERP, we adapted metrics developed for auditing partisan ranking bias in Twitter search, which often returns lists of tweets and retweets that are orders of magnitude larger than the number of results returned on a single SERP [57]. In web search, users often do not browse past the first page of search results before reformulating a query, and the first SERP typically contains about 15 components [84]. According to an ongoing study, the average desktop click-through rate for the first 15 links during our audit accounted for the vast majority (80.2%) of clicks [79]. Therefore, we only collected the first SERP returned for each query and adjusted the *input bias* and *output bias* metrics developed in prior work [57] to an *average bias* and *weighted bias*, respectively, to distinguish our focus on a smaller set of rankings (the first SERP). We also adopted their *ranking bias* metric, giving us a total of three metrics for each SERP:

(1) The *average bias* of a SERP is the unweighted average of the bias of all items on the page. This captures the overall bias of the SERP irrespective of item ordering, and provides a measure of bias for Google's filtering algorithm.
(2) The *weighted bias* of a SERP is the weighted average of the bias of all items on the page, where weight decreases with rank.[16]
(3) The *ranking bias* of a SERP is the average bias minus the weighted bias. This captures the impact of Google's ranking algorithm, given the results of its filtering algorithm. That is, given Google's filtering, how did their ranking algorithm further alter the bias of the SERP?

4 RESULTS

In total, we were able to match our bias scores to 83.9% of the domains that appeared in our search data (Figure 6).[17] We were able to match domains in `knowledge` (78.4%), `general-extra` (75.9%),

[15] We present more domain-level bias scores in SM.
[16] Weighted bias is calculated using the same formula as output bias (OB) [57], but limited to the first page of results. One can calculate OB by finding the bias (B) until reaching rank r for a query q given each score $s_i : B(q, r) = \sum_{i=1}^{r} s_i / r$, and then taking the normalized sum of this over all ranks: $OB(q, r) = \sum_{i=1}^{r} B(q, i)/r$.
[17] This excludes components and subcomponents that did not contain a domain-level URL, such as internal Google links.

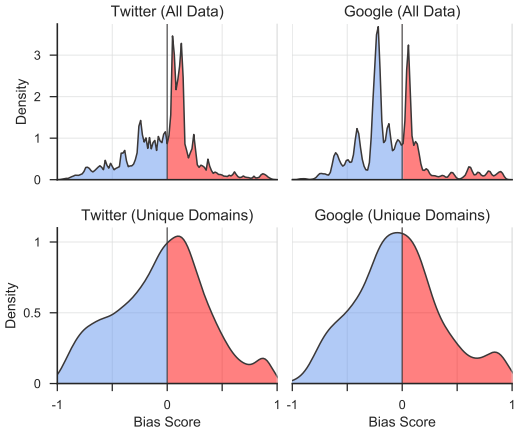


Fig. 6. Distribution of bias scores in the Twitter and Google data for all domains and unique domains.

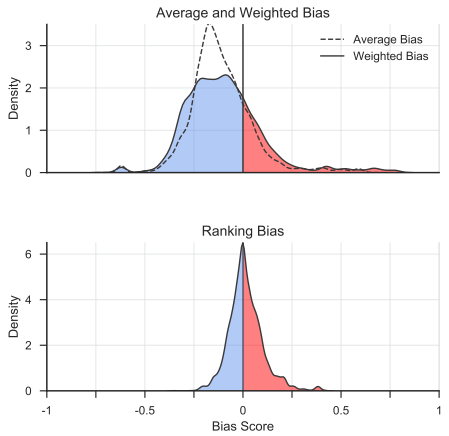


Fig. 7. Overall distributions of SERP bias metrics: average and weighted bias and ranking bias.

`video` (100%), `news-card` (98.7%), `news-triplet` (98.6%), `people-ask` (66.2%), `general` (80.1%), and `general-video` (100%) components.[18]

Below we report how average bias, weighted bias, and ranking bias varied overall, by personalization factors, root query, user characteristics, and time.

4.1 Overall Results

Comparing the mean ranking bias of the standard and incognito SERPs within participants, we found little support for the "filter bubble" hypothesis that personalization increases the partisan bias of web search. However, we did find a small but significant ranking bias among the standard SERPs collected from participants' computers ($\mu = 0.02, t = -21.4^{***}$), suggesting that Google's ranking algorithm has a minimal impact on the partisan lean of SERPs, but not due to personalization (Figure 7). This result held even after removing links to wikipedia.org, which accounted for 14% of all domains appearing at the first rank, and which had a relative left-lean of -0.22. Removing these links shifted the the ranking bias significantly but unsubstantially to the right (shift: 0.0056; $t = 43.87^{***}$; see SM for more details). It is unclear whether a ranking bias of this magnitude would translate to substantial effects on voter behavior, such as those found in lab-based experiments using maximally biased results, though the effects of subtle biases could potentially compound over time [28].

By Personalization. We found negligible or non-significant differences between the SERPs returned in the standard and incognito windows, suggesting that, relative to the nonpersonalized results, Google's personalization had little impact on the partisan audience bias that our participants would have been exposed to. More specifically, we used a paired samples t-test to compare the average and weighted bias present in each standard-incognito SERP pair and found a small but significant difference in average bias, with standard windows producing slightly less left-leaning

[18]There was a large degree of variance in our coverage of the domains that appeared in each component for two reasons. First, some components, such as `map`, `map-card`, and `people-search`, linked only to internal Google URLs. Second, some components, such as `knowledge` and `people-ask` often did not provide a URL, or presented one in a dynamic format that did not fully load during data collection. For `general-video` and `video` components, coverage was perfect, but the number of unique domains produced by each was low ($n = 12$ and $n = 1$ [youtube.com], respectively).

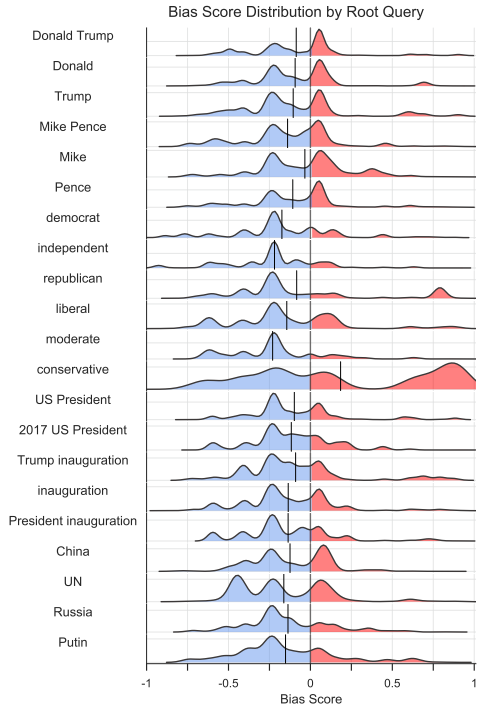


Fig. 8. Domain bias score distribution aggregated by root query. Vertical lines represent the mean of the distribution. Relative to the other roots, only "conservative" and its suggestions produced a right-leaning distribution.

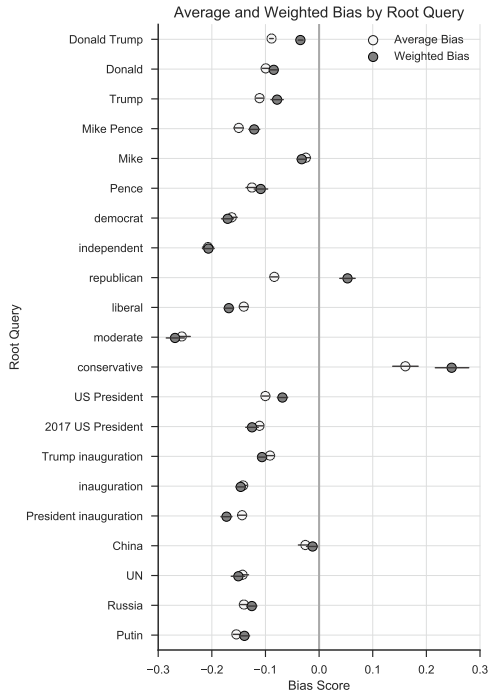


Fig. 9. Relative to the other roots and their suggestions, only "republican" and "conservative" produced SERPs with a mean right-leaning weighted bias, and only "conservative" produced SERPs with a mean right-leaning average bias. Horizontal lines represent 95% confidence intervals.

SERPs than incognito windows, on average (Difference $\mu = 0.001, t = 2.87^{**}$). However, the difference in weighted bias between standard-incognito paired SERPs was not significant ($\mu = 0.0001, t = -0.29, P = 0.77$).[19] To avoid inflating our sample size by considering both SERPs in each pair, which were not meaningfully different, we consider only the standard SERPs in the rest of our analysis.

By Root Query. We found substantial differences in the distribution of partisan bias scores by root query (Figure 8) and examined these differences with respect to the average, weighted, and ranking bias of the standard SERPs they produced. To do this, we plotted the mean average and weighted bias of the SERPs produced for each root query (Figure 9), and compared them as before. After applying a Boneferroni correction for multiple hypothesis testing, we found that Google's rankings created significant differences between the mean average and weighted bias for 18 of the 21 root queries (non-significant roots: "Mike", "independent", and "inauguration").

Among the root queries that had significant differences between their mean average and weighted bias, only two were not significant at the $P < 0.001$ level: "UN" (ranking bias $\mu = -0.01, t = 3.217^{*}$)

[19] We examined this more in depth by fitting a two-way ANOVA using participants' political party and Google account login status, which has previously been found to be a key feature used to personalize search results [45, 84], to predict average and ranking bias but did not find significant differences (see SM).

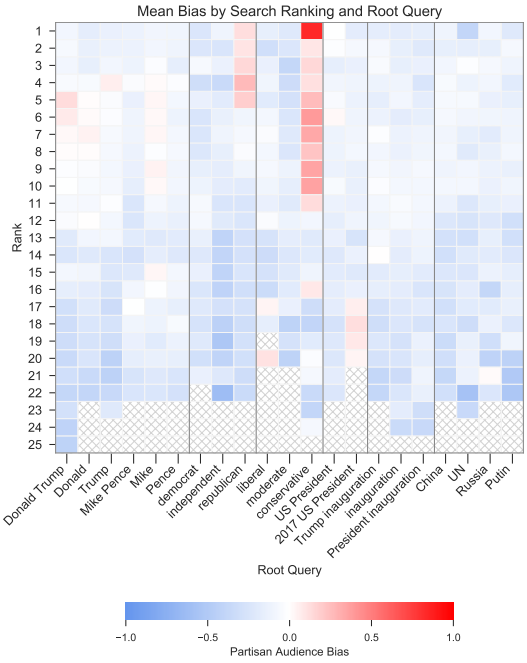


Fig. 10. Heatmap of the bias score given each root and rank. Only the root queries specifically mentioning the ideological right ("conservative" and "republican") return SERPS with right-leaning domains at prominent search rankings. The root "Donald Trump" had right-leaning domains at ranks five through eight due to the `twitter-card` subcomponents linking to his account.

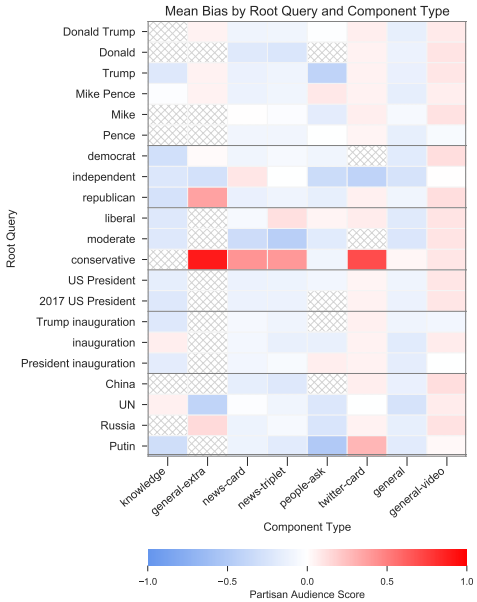


Fig. 11. Mean bias by root query and component type for the nine component types that we could match our bias scores to. Here we can see that the relatively right-leaning average and weighted bias for the root "conservative" (Figure 9) may be due to the relative right-lean of the domains appearing in highly ranked components for this root, such as the `general-extra`, `news-card`, `news-triplet`, and `twitter-card` components.

and "democrat" (ranking bias $\mu = -0.01, t = 3.747^{**}$). Among the 16 other roots, the mean ranking bias ranged from −0.03 ("President inauguration") to 0.14 ("republican"). Although the root "conservative" produced the average bias that was the farthest to the right relative to the other roots, we saw the largest right-leaning shift in ranking bias for the root query "republican" (Figure 9).

To obtain a more detailed picture of how SERP bias varied by root, we constructed a heatmap detailing the mean bias at each search ranking for all SERPs produced by each root query and its suggestions (Figure 10). Here we unpack the subcomponents present on each SERP, increasing the length of some SERPs (max rank is 25 compared to a max rank of 20 in Figure 2, where subcomponents are merged). This figure shows us that the top rank of SERPs produced by the root "conservative" was, on average, substantially right leaning (0.84) relative to the average top rank produced by other roots. A similar pattern can be seen for the root "republican." In contrast, the root "Donald Trump" produced a mean right-leaning bias only for ranks five through eight, where the `twitter-card` subcomponents were typically located.

4.2 Component Bias

Given that different root queries produce substantial differences in SERP composition (Figure 2) [84], we analyzed how bias varied between and within each component type. We utilized the average

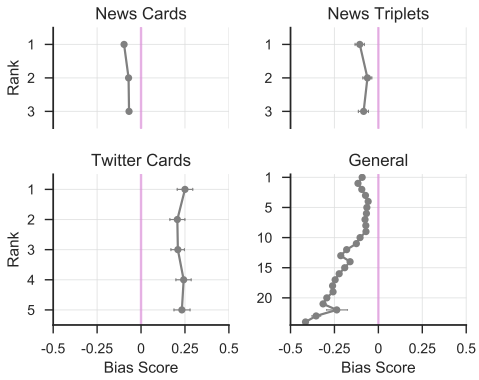


Fig. 12. Mean bias by rank for each component, and where applicable, its subcomponents.

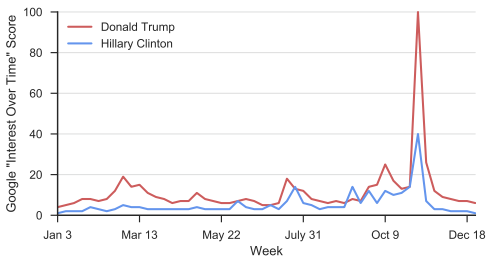


Fig. 13. Google's Interest over time score for the search queries "Donald Trump" and "Hillary Clinton" for each week in 2016. In aggregate, Donald Trump received 2.1 times as much search activity as Hillary Clinton [40].

bias (unweighted average) of the results within each component type as our primary metric because we were no longer analyzing whole SERPs. Excluding the components for which we did not have any bias score coverage, we found significant differences in mean bias by component type ($F(10, 345, 141) = 2,284.9^{***}$). Comparing among the components types that we could match our bias scores to ($N = 8$; ignoring `video` comoponents which solely linked to youtube.com), we found three component types that produced right-leaning mean bias scores relative to the other components (Figure 11). However, among these, only `twitter-card` components constituted a substantial proportion of the total components seen (7.1% compared to 0.3% for `general-extra`, and 1.1% for `general-video` components).

By Rank. Extending upon prior work [21, 29], we examined the differences in mean bias by search ranking *within* the subcomponents of `news-card`, `news-triplet`, and `twitter-card` components (Figure 12). For example, given a `news-card`, which consists of three rank ordered subcomponents (Figure 2), did bias significantly and systematically vary by rank within the component? We also show the mean bias of the ubiquitous `general` components by rank, but for these we consider the entire SERP since these are not composed of ranked subcomponents.

We found significant differences in mean bias by rank for the domains present in `news-card` ($F(2, 24, 455) = 17.2^{***}$) subcomponents and `general` components ($F(24, 173, 099) = 225.8^{***}$), but not for `news-triplet` ($F(2, 1, 581) = 2.45, P = 0.09$) or `twitter-card` subcomponents ($F(2, 2, 097) = 0.81, P = 0.52$). In terms of correlations between bias score and rank, we found small but significant relationships for `general` ($\rho = -0.12^{***}$), `news-card` ($\rho = 0.06^{***}$), and `news-triplet` ($\rho = 0.02^{*}$) subcomponents, but not for `twitter-cards` ($\rho = -0.04, P = 0.06$). These results resemble those from prior work [29], and can have many interpretations [21], but without access to Google's data, we can only speculate. Among possible interpretations are (1) that by some measure of popularity or "quality," Google's algorithms favor left-leaning news domains, (2) left-leaning domains were somehow more "relevant" to the set of queries we used, or (3) left-leaning news sources were producing more news at the time of our audit, and this "freshness" led them to be highly ranked.[20]

[20]We found some evidence for the popularity interpretation by merging our scores with public web domain rankings from Alexa ($\rho = -0.26^{***}$), and the focus of our search queries on the inauguration of a Republican president cast doubt on (2), but a deeper examination of these is beyond the scope of this paper.

By Root Query. Relative to the other roots, we found that the root "conservative" returned the most right-leaning mean bias in the `general-extra`, `news-card`, `news-triplet`, and `twitter-card` components (Figure 11). Similarly, compared to the mean bias by root in other components, `twitter-card` subcomponents produced a right-leaning bias for 17 roots, `general-video` components produced an average right-leaning bias for 18 roots, and `general-extra` components produced an average right-leaning bias for 7 roots.

5 DISCUSSION

Here we presented a study in which we audited partisan audience bias within Google Search, a platform that is used trillions of times a year [92], and which research suggests has the power to sway democratic elections [28, 30]. After collecting Google Search data through the browsers of real users over the course of several weeks around a major political event, we developed partisan audience bias metric based on the sharing propensities of registered Democrats and Republicans from a large virtual Twitter panel and applied these scores to the search data we collected. Compared to systems relying on human raters, our scoring method offers a more practical and responsive system for obtaining and updating domain bias scores in real-time.

After matching the bias scores we developed to the domains present in the search results with fairly high coverage (83.9%), we grouped them by SERP and quantified two types of bias for each. Adapting from prior work on social media search engine bias [57], we operationalized *average bias*, a simple unweighted average, as the partisan bias of Google's web corpus and filtering algorithm, and *weighted bias*, a rank-weighted average, as the partisan bias of a SERP after Google's ranking algorithm sorts it. Utilizing paired statistical tests between the average and weighted bias of each SERP, we measured the significance of *ranking bias*, the weighted bias minus the average bias.

Within this framework, we found little evidence for the "filter bubble" hypothesis. Instead, we found that across all participants the results placed toward the bottom of Google SERPs were more left-leaning than the results placed toward the top, which connects to prior findings [29], and that the direction and magnitude of overall lean varied widely by search query, component type, and other factors. Utilizing the rank-weighted metrics that we adapted, we also found that Google's ranking algorithm shifted the average lean of SERPs slightly to the right of their unweighted average.

One explanation for the correlation we found between lower search rankings and more left-leaning domains is that Democrat audiences may browse further down a page of search results than Republican audiences. Ancillary evidence for this hypothesis can be found in a number of recent studies, including (1) a 2018 ethnographic study on conservative search engine users which found that they have high levels of trust in Google's rankings [94], (2) a 2016 study which used poll data from Gallup to show that Republicans trust the mainstream media far less than Democrats [1], and (3) a 2014 Pew study which found that Democrats relied on a wider range of news outlets while Republicans tended to tightly cluster around a single news source (Fox News) [68], a finding that has been consistent over time [49, 65]. In a similar vein, the political psychology literature has consistently found an association between identifying as a liberal and the Big Five[21] personality trait "Openness to Experience" [15, 83]. However, we can only speculate on the cause of this consistent finding [29]; only Google has the behavioral data to answer it.

In our audit, we took into account an aspect of the search engine interface that has often been overlooked – the diverse ranking component types that Google Search employs [21, 84]. We found that Google's decision to embed tweets in their search results, which they began doing in

[21]The Big Five is a set of personality traits–Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness to Experience–that has long been studied in the psychological sciences, typically through survey-based methods.

August 2015 [39, 86], likely amplified the reach of Donald Trump's Twitter account because of its prominence near the top of search results (Figure 2). How many people were exposed to his tweets in search results? According to Google Trends data, "Donald Trump" was the only person's name in the top ten searches of 2016, and activity peaked during the week of the election [40] (Figure 13). Comparatively, searches for "Hillary Clinton" in 2016 had about half the volume of search activity.

It is possible that the relative right-leaning mean bias of `twitter-card` components extended beyond searches for "Donald Trump." An analysis of the most viral tweets during the final 68 days of the 2016 election found that 63% either supported Trump or attacked Clinton, and that tweets favoring Trump were retweeted more often than those favoring Clinton [19]. However, the specific factors that trigger a `twitter-card` component to appear in a Google SERP for a given query are unclear, as are the factors influencing which account will be featured and how tweets from that account are filtered and ranked. Unfortunately, it is beyond the scope of this study to determine the actual impact that this design decision had on the reach of Donald Trump's tweets or the outcome of the election. Future research should try to synthesize and assess the link between search engines and social media, as this link can have unexpected impacts on spreading processes [82].

Limitations. There are several reasons why our bias scores should not be taken as a ground truth purely reflecting population preferences. Given that our measure is audience-based, a value of 0 only means that an equal proportion of unique Democrats and Republicans in our panel shared that domain, and does not explicitly establish that the domain is nonpartisan or "neutral." Furthermore, a URL from a domain that typically publishes left-leaning content could be widely shared by Republicans due to the specifics of the article, and vice versa. Our measure is also coarse-grained because we consider tweets containing a domain to reflect agreement with the linked article rather than disagreement, which is unlikely to always be the case. Finally, we cannot control for the impact of Twitter's curation algorithms that likely affected which domains our panel members were exposed to and how often they were exposed to them, both of which could have affected their propensity to share those domains.

Despite these limitations, the overall agreement of our metric with prior bias metrics [2, 3, 12] and hand coding (Figure 3), as well as the expected leans of popular media domains (Figure 5) is encouraging. Similarly, we found high agreement between panelists' registered political party and the average bias of the domains they shared (Figure 4). Given the fairly strong validity we established for our scores, the scalability of our method offers a promising route forward compared to the alternatives. Given the challenges associated with constructing a virtual panel with voter registration records, future research should further investigate the possibility of creating a large scale panel by inferring political affiliations [46, 61, 98].

It is important to recognize that our results are aggregated across queries and dates, and that our participant sample was not balanced in terms of their political preferences or demographics, nor matched on the time of day that they participated in our audit (which precludes us from performing between-subjects comparisons). In terms of queries, we focused our root queries around Trump's inauguration and used Google's autocomplete functionality to expand them. This resulted in sets of queries that were popular and relevant to our root queries, but these sets did not reflect the nuanced nature of users' query formulation processes [94]. In terms of temporal findings, we found that average and ranking bias varied by date, but a detailed temporal analysis is hindered by our limited and uneven sample size per day. Across all dates, the mean average bias of SERPs was left-leaning, and Google's ranking algorithm significantly pushed the weighted bias slightly right (all $P < 0.001$), but otherwise, no clear trends emerged. A study similar to ours, but with a carefully matched sample, could provide a longitudinal picture of partisan bias in web search.

Conclusions. When billions of people cognitively depend on an online platform each day [90], every design decision made by the platform carries a broad impact. This impact is not only on the individual information seeker, determining what information they find and absorb, but on society at large, influencing our culture and politics by steering people toward certain answers and perspectives. In terms of studying this impact externally, algorithm audits provide a useful starting point. However, broader frameworks incorporating design elements like the components we identified may shed light not just on the algorithms, but the user interface as well.

As Google's search results and their components continue to evolve with other entities in the online information ecosystem, audits like the one we have conducted will continue to grow in importance and should be conducted regularly. Future simulation-based models of search engine users (*e.g.,* [31]) and behavioral experiments should incorporate our findings with respect to the variance in bias by component type and root query. Future audits should keep an eye out for the release of new or modified component types that may only appear during election seasons, such as the "Issue Guide" identified by Diakopoulos *et al.* [21]. Indeed, the Twitter components we studied have already evolved in response to Twitter's character limit increase, and now resemble the larger `news-card` components. How will this new design influence the amount of clicks they receive? And how will it amplify the reach of the accounts and Tweets they give prominence to? We leave this and other questions to future research.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their helpful comments and Stefan McCabe, Will Hobbs, Piotr Sapieżyński, Nir Grinberg, and others for invaluable discussions and comments on this work. This research was supported in part by NSF grants IIS-1408345 and IIS-1553088. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

REFERENCES

[1] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31, 2 (May 2017), 211–236. https://doi.org/10.1257/jep.31.2.211
[2] AllSides. 2018. Media Bias Ratings. AllSides. (2018). https://www.allsides.com/media-bias/media-bias-ratings
[3] Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132.
[4] Nicholas J. Belkin. 1980. Anamolous States of Knowledge as a Basis for Information Retrieval. *Canadian Journal of Information and Library Sciences* 5 (1980), 133–143.
[5] James R Beniger. 1987. Personalization of mass media and the growth of pseudo-community. *Communication research* 14, 3 (1987), 352–371.
[6] W. Lance Bennett and Jarol B. Manheim. 2006. The One-Step Flow of Communication. *The ANNALS of the American Academy of Political and Social Science* 608, 1 (2006), 213–232.
[7] Shlomo Berkovsky, Jill Freyne, and Harri Oinas-Kukkonen. 2012. Influencing individually: fusing personalization and persuasion. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2, 2 (2012), 9.
[8] Edelman Berland. 2017. 2017 Edelman Trust Barometer. (2017). http://www.edelman.com/trust2017/ Accessed: 2017-03-07.
[9] Edward L Bernays. 1928. *Propaganda.* Ig publishing.
[10] Robert M Bond, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle, and James H Fowler. 2012. A 61-million-person experiment in social influence and political mobilization. *Nature* 489, 7415 (2012), 295–298.
[11] Erik Borra and Ingmar Weber. 2012. Political Insights: Exploring Partisanship in Web Search Queries. *First Monday* 17, 7 (July 2012). https://doi.org/10.5210/fm.v17i7.4070
[12] Ceren Budak, Sharad Goel, and Justin M Rao. 2016. Fair and balanced? Quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly* 80, S1 (2016), 250–271.
[13] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2011. Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science* 6, 1 (2011), 3–5.

[14] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.
[15] Dana R. Carney, John T. Jost, Samuel D. Gosling, and Jeff Potter. 2008. The Secret Lives of Liberals and Conservatives: Personality Profiles, Interaction Styles, and the Things They Leave Behind. *Political Psychology* 29, 6 (Dec. 2008), 807–840. https://doi.org/10.1111/j.1467-9221.2008.00668.x
[16] Kalyani Chadha and Rob Wells. 2016. Journalistic Responses to Technological Innovation in Newsrooms: An Exploratory Study of Twitter Use. *Digital Journalism* 4, 8 (Nov. 2016), 1020–1035. https://doi.org/10.1080/21670811.2015.1123100
[17] Chun-Fang Chiang and Brian Knight. 2011. Media Bias and Influence: Evidence from Newspaper Endorsements. *The Review of Economic Studies* 78, 3 (2011), 795–820.
[18] Inc comScore. 2017. comScore Explicit Core Search Query Report (Desktop Only). (2017). https://www.comscore.com/Insights/Rankings Accessed: 2017-02-12.
[19] Kareem Darwish, Walid Magdy, and Tahar Zanouda. 2017. Trump vs. Hillary: What Went Viral During the 2016 US Presidential Election. In *International Conference on Social Informatics*. Springer, 143–161.
[20] Stefano DellaVigna and Ethan Kaplan. 2007. The Fox News effect: Media bias and voting. *The Quarterly Journal of Economics* 122, 3 (2007), 1187–1234.
[21] 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*, M. Moore and D. Tambini (Eds.). 22.
[22] James N Druckman and Michael Parkin. 2005. The impact of media bias: How editorial slant affects voters. *Journal of Politics* 67, 4 (2005), 1030–1049.
[23] William H. Dutton and Grant Blank. 2013. *Cultures of the Internet: The Internet in Britain: Oxford Internet Survey 2013 Report.* Technical Report. Oxford Internet Institute, University of Oxford. http://oxis.oii.ox.ac.uk/wp-content/uploads/2014/11/OxIS-2013.pdf
[24] William H. Dutton, Bianca Christin Reisdorf, Elizabeth Dubois, and Grant Blank. 2017. Search and Politics: A Cross-National Survey. (2017).
[25] William H. Dutton, Bianca Christin Reisdorf, Elizabeth Dubois, and Grant Blank. 2017. Search and Politics: The Uses and Impacts of Search in Britain, France, Germany, Italy, Poland, Spain, and the United States. (2017).
[26] Dina ElBoghdady. 2013-04-23T09:52-500. Market Quavers after Fake AP Tweet Says Obama Was Hurt in White House Explosions. *Washington Post* (2013-04-23T09:52-500).
[27] Robert Epstein. 2018. Taming Big Tech: The Case for Monitoring. Hacker Noon. (May 2018). https://hackernoon.com/taming-big-tech-5fef0df0f00d
[28] Robert Epstein and Ronald E Robertson. 2015. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences* 112, 33 (2015), E4512–E4521.
[29] Robert Epstein and Ronald E Robertson. 2017. *A Method for Detecting Bias in Search Rankings, with Evidence of Systematic Bias Related to the 2016 Presidential Election.* Technical Report White Paper no. WP-17-02. American Institute for Behavioral Research and Technology. 5 pages.
[30] Robert Epstein, Ronald E. Robertson, David Lazer, and Christo Wilson. 2017. Suppressing the search engine manipulation effect (SEME). *Proceedings of the ACM: Human-Computer Interaction* 1, 42 (2017). Issue CSCW.
[31] S. Fortunato, A. Flammini, F. Menczer, and A. Vespignani. 2006. Topical Interests and the Mitigation of Search Engine Bias. *Proceedings of the National Academy of Sciences* 103, 34 (Aug. 2006), 12684–12689. https://doi.org/10.1073/pnas.0605525103
[32] Adrien Friggeri, Lada A Adamic, Dean Eckles, and Justin Cheng. 2014. Rumor Cascades. In *ICWSM*.
[33] Gallup. 2017. Presidential Approval Ratings – Donald Trump. (2017). http://news.gallup.com/poll/203198/presidential-approval-ratings-donald-trump.aspx Accessed: 2017-10-08.
[34] Matthew Gentzkow and Jesse M Shapiro. 2010. What drives media slant? Evidence from US daily newspapers. *Econometrica* 78, 1 (2010), 35–71.
[35] Alan S Gerber, Dean Karlan, and Daniel Bergan. 2009. Does the media matter? A field experiment measuring the effect of newspapers on voting behavior and political opinions. *American Economic Journal: Applied Economics* 1, 2 (2009), 35–52.
[36] Tarleton Gillespie. 2014. The Relevance of Algorithms. In *Media Technologies: Essays on Communication, Materiality, and Society*, Tarleton Gillespie, Pablo J. Boczkowski, and Kirsten A. Foot (Eds.). The MIT Press, Cambridge, 167–194. https://doi.org/10.7551/mitpress/9780262525374.003.0009
[37] Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan J. Watts. 2015. The Structural Virality of Online Diffusion. *Management Science* 62, 1 (July 2015), 180–196. https://doi.org/10.1287/mnsc.2015.2158
[38] Sharad Goel, Duncan J. Watts, and Daniel G. Goldstein. 2012. The Structure of Online Diffusion Networks. ACM Press, 623. https://doi.org/10.1145/2229012.2229058

[39] Google. 2015. Tweets take flight in the Google app. Google Blog. (2015). https://googleblog.blogspot.com/2015/05/tweets-take-flight-in-google-app.html
[40] Google. 2016. Google Trends: Year in Search. Google Trends. (2016). https://trends.google.com/trends/yis/2016/GLOBAL
[41] Google. 2017. Google Chrome Privacy Notice. (2017). https://www.google.com/chrome/browser/privacy/#browser-modes Accessed: 2017-04-01.
[42] Jeffrey Gottfried and Elisa Shearer. 2016. News use across social media platforms. Pew Research Center. (2016). http://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/
[43] Tim Groseclose and Jeffrey Milyo. 2005. A measure of media bias. *The Quarterly Journal of Economics* 120, 4 (2005), 1191–1237.
[44] Andrew Guess, Brendan Nyhan, Benjamin Lyons, and Jason Reifler. 2018. *Avoiding the Echo Chamber About Echo Chambers*. Technical Report. Knight Foundation.
[45] Aniko Hannak, Piotr Sapieżyński, Arash Molavi Kakhki, Balachander Krishnamurthy, David Lazer, Alan Mislove, and Christo Wilson. 2013. Measuring Personalization of Web Search. In *Proceedings of the 22nd International World Wide Web Conference*.
[46] Itai Himelboim, Stephen McCreery, and Marc Smith. 2013. Birds of a Feather Tweet Together: Integrating Network and Content Analyses to Examine Cross-Ideology Exposure on Twitter. *Journal of Computer-Mediated Communication* 18, 2 (Jan. 2013), 40–60. https://doi.org/10.1111/jcc4.12001
[47] William Hobbs, Lisa Friedland, Kenneth Joseph, Oren Tsur, Stefan Wojcik, and David Lazer. 2017. "Voters of the Year": 19 Voters Who Were Unintentional Election Poll Sensors on Twitter. In *ICWSM*.
[48] Lucas D Introna and Helen Nissenbaum. 2000. Shaping the Web: Why the Politics of Search Engines Matters. *The Information Society* 16 (2000), 169–185.
[49] Shanto Iyengar and Kyu S Hahn. 2009. Red Media, Blue Media: Evidence of Ideological Selectivity in Media Use. *Journal of Communication* 59, 1 (March 2009), 19–39. https://doi.org/10.1111/j.1460-2466.2008.01402.x
[50] Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political Ideology Detection Using Recursive Neural Networks. In *Association for Computational Linguistics*.
[51] Shan Jiang and Christo Wilson. 2018. Linguistic Signals under Misinformation and Fact-Checking: Evidence from User Comments on Social Media. *Proceedings of the ACM: Human-Computer Interaction* 2 (2018). Issue CSCW.
[52] Kenneth Joseph, Lisa Friedland, William Hobbs, Oren Tsur, and David Lazer. 2017. ConStance: Modeling Annotation Contexts to Improve Stance Classification. *arXiv preprint arXiv:1708.06309* (2017).
[53] Cecilia Kang, Tiffany Hsu, Kevin Roose, Natasha Singer, and Matthew Rosenberg. 2018. Mark Zuckerberg Testimony: Day 2 Brings Tougher Questioning. *The New York Times* (April 2018).
[54] Elihu Katz and Paul Felix Lazarsfeld. 1955. *Personal influence, The part played by people in the flow of mass communications*. The Free Press.
[55] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. 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*. ACM Press, 3819–3828. https://doi.org/10.1145/2702123.2702520
[56] Chloe Kliman-Silver, Aniko Hannak, David Lazer, Christo Wilson, and Alan Mislove. 2015. Location, Location, Location: The Impact of Geolocation on Web Search Personalization. In *Proceedings of the 2015 ACM Conference on Internet Measurement*.
[57] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P Gummadi, and Karrie Karahalios. 2017. Quantifying search bias: Investigating sources of bias for political searches in social media. In *20th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW 2017)*.
[58] Harold D Lasswell. 1938. *Propaganda technique in the world war*. P. Smith, New York, NY.
[59] David Lazer, Alex (Sandy) Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. 2009. Life in the Network: The Coming Age of Computational Social Science. *Science (New York, N.Y.)* 323, 5915 (Feb. 2009), 721–723. https://doi.org/10.1126/science.1167742
[60] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The Science of Fake News. *Science* 359, 6380 (March 2018), 1094–1096. https://doi.org/10.1126/science.aao2998
[61] Huyen Le, Zubair Shafiq, and Padmini Srinivasan. 2017. Scalable News Slant Measurement Using Twitter. In *ICWSM*. 4.
[62] S. C. Matz, M. Kosinski, G. Nave, and D. J. Stillwell. 2017. Psychological Targeting as an Effective Approach to Digital Mass Persuasion. *Proceedings of the National Academy of Sciences* (Nov. 2017), 201710966. https://doi.org/10.1073/pnas.1710966114

[63] Connor McMahon, Isaac 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. (2017).
[64] Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology* 27, 1 (Aug. 2001), 415–444. https://doi.org/10.1146/annurev.soc.27.1.415
[65] Solomon Messing and Sean J. Westwood. 2014. Selective Exposure in the Age of Social Media: Endorsements Trump Partisan Source Affiliation When Selecting News Online. *Communication Research* 41, 8 (Dec. 2014), 1042–1063. https://doi.org/10.1177/0093650212466406
[66] P Takis Metaxas and Yada Pruksachatkun. 2017. Manipulation of Search Engine Results during the 2016 US Congressional Elections. In *ICIW*.
[67] Ronald B Mincy. 1993. The Urban Institute audit studies: their research and policy context. *Clear and Convincing Evidence: Measurement of Discrimination in America* (1993), 165–86.
[68] Amy Mitchell, Jeffrey Gottfried, Jocelyn Kiley, and Katerina Eva Matsa. 2014. Political Polarization & Media Habits. Pew Research Center's Journalism Project. (Oct. 2014).
[69] Nic Newman, David A. L. Levy, and Rasmus Kleis Nielsen. 2017. Reuters Institute Digital News Report 2017. *SSRN Electronic Journal* (2017). https://doi.org/10.2139/ssrn.2619576
[70] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism.* NYU Press.
[71] Brendan Nyhan. 2012. Does the US Media Have a Liberal Bias? *Perspectives on Politics* 10, 03 (Sept. 2012), 767–771. https://doi.org/10.1017/S1537592712001405
[72] Stefan Palan and Christian Schitter. 2017. Prolific.ac - subject pool for online experiments. *Journal of Behavioral and Experimental Finance* (2017). https://doi.org/10.1016/j.jbef.2017.12.004
[73] Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you.* Penguin UK.
[74] Souneil Park, Minsam Ko, Jungwoo Kim, Ying Liu, and Junehwa Song. 2011. The politics of comments: predicting political orientation of news stories with commenters' sentiment patterns. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. ACM, 113–122.
[75] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond the Turk: alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* 70 (2017), 153–163.
[76] Richard E Petty, Joseph R Priester, and Pablo Brinol. 2009. *Mass media attitude change: Implications of the elaboration likelihood model of persuasion* (3rd ed.). Vol. 2. Lawrence Erlbaum Associates, Mahwah, NJ, 155–198.
[77] Media Insight Project. 2014. The Personal News Cycle: How Americans Choose to Get Their News. Media Insight Project. (March 2014). https://www.americanpressinstitute.org/publications/reports/survey-research/how-americans-get-news/
[78] Kristen Purcell, Joanna Brenner, and Lee Rainie. 2012. *Search engine use 2012.* Technical Report. Pew Research Center's Internet and American Life Project.
[79] Advanced Web Ranking. 2017. CTR study. (2017). https://www.advancedwebranking.com/cloud/ctrstudy.
[80] Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic Models for Analyzing and Detecting Biased Language. In *51st Annual Meeting of the Association for Computational Linguistics*. ACL, 1650–1659.
[81] Filipe N Ribeiro, Lucas Henrique, Fabricio Benevenuto, Abhijnan Chakraborty, Juhi Kulshrestha, Mahmoudreza Babaei, and Krishna P Gummadi. 2018. Media Bias Monitor: Quantifying Biases of Social Media News Outlets at Large-Scale. (2018), 10.
[82] Christoph Riedl, Johannes Bjelland, Geoffrey Canright, Asif Iqbal, Kenth Engø-Monsen, Taimur Qureshi, Pål Roe Sundsøy, and David Lazer. 2018. Product Diffusion through On-Demand Information-Seeking Behaviour. *Journal of The Royal Society Interface* 15, 139 (Feb. 2018), 20170751. https://doi.org/10.1098/rsif.2017.0751
[83] Rainer Riemann, Claudia Grubich, Susanne Hempel, Susanne Mergl, and Manfred Richter. 1993. Personality and attitudes towards current political topics. *Personality and Individual Differences* 15, 3 (1993).
[84] Ronald E Robertson, David Lazer, and Christo Wilson. 2018. Auditing the Personalization and Composition of Politically-Related Search Engine Results Pages. In *WWW 2018: The 2018 Web Conference*. ACM, Lyon, France, 11. https://doi.org/10.1145/%00203178876.3186143
[85] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. In *Proceedings of "Data and Discrimination: Converting Critical Concerns into Productive Inquiry", a Productivereconference at the 64th Annual Meeting of the International Communication Association.*
[86] Barry Schwartz. 2015. Google Officially Expands Twitter Into Desktop Search Results. Search Engine Land. (2015). https://searchengineland.com/google-officially-expands-twitter-into-desktop-search-results-228723
[87] David O. Sears and Jonathan L. Freedman. 1967. Selective Exposure to Information: A Critical Review. *Public Opinion Quarterly* 31, 2 (1967), 194. https://doi.org/10.1086/267513
[88] Pawel Sobkowicz. 2017. Social Simulation Models at the Ethical Crossroads. *Science and Engineering Ethics* (Nov. 2017), 1–15. https://doi.org/10.1007/s11948-017-9993-0

[89] Ruihua Song, Zhenxiao Luo, Ji-Rong Wen, Yong Yu, and Hsiao-Wuen Hon. 2007. Identifying Ambiguous Queries in Web Search. In *Proceedings of the 16th International Conference on World Wide Web (WWW '07)*. ACM, New York, NY, USA, 1169–1170. https://doi.org/10.1145/1242572.1242749

[90] Betsy Sparrow, Jenny Liu, and Daniel M Wegner. 2011. Google effects on memory: Cognitive consequences of having information at our fingertips. *Science* 333, 6043 (2011), 776–778.

[91] Jennifer A Stark and Nicholas Diakopoulos. 2017. Using Baselines for Algorithm Audits. In *European Conference on Data and Computational Journalism*. 3.

[92] Danny Sullivan. 2016. Google Now Handles at Least 2 Trillion Searches per Year. Search Engine Land. (May 2016). https://searchengineland.com/google-now-handles-2-999-trillion-searches-per-year-250247

[93] New York Times. 1995. Edward Bernays, 'Father of Public Relations' And Leader in Opinion Making, Dies at 103. (1995).

[94] Francesca Tripodi. 2018. Searching for Alternative Facts: Analyzing Scriptural Inference in Conservative News Practices. Data & Society. (May 2018).

[95] Liwen Vaughan and Mike Thelwall. 2004. Search Engine Coverage Bias: Evidence and Possible Causes. *Information Processing & Management* 40, 4 (July 2004), 693–707. https://doi.org/10.1016/S0306-4573(03)00063-3

[96] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The Spread of True and False News Online. *Science* 359, 6380 (March 2018), 1146–1151. https://doi.org/10.1126/science.aap9559

[97] Duncan J. Watts and David M. Rothschild. 2017. Don't blame the election on fake news. Blame it on the media. Columbia Journalism Review. (2017). https://www.cjr.org/analysis/fake-news-media-election-trump.php

[98] Felix Ming Fai Wong, Chee Wei Tan, Soumya Sen, and Mung Chiang. 2016. Quantifying political leaning from tweets, retweets, and retweeters. *IEEE transactions on knowledge and data engineering* 28, 8 (2016), 2158–2172.

[99] Daniel Xiaodan Zhou, Paul Resnick, and Qiaozhu Mei. 2011. Classifying the Political Leaning of News Articles and Users from User Votes.. In *ICWSM*.

[100] Jonathan L Zittrain. 2014. Engineering an election. *Harvard Law Review Forum* 127, 8 (2014), 335–341.

Received April 2018; revised June 2018; accepted September 2018