Auditing the Personalization and Composition of Politically-Related Search Engine Results Pages

Ronald E. Robertson
Northeastern University
rer@ccs.neu.edu

David Lazer
Northeastern University
d.lazer@neu.edu

Christo Wilson
Northeastern University
cbw@ccs.neu.edu

ABSTRACT

Search engines are a primary means through which people obtain information in today's connected world. Yet, apart from the search engine companies themselves, little is known about how their algorithms filter, rank, and present the web to users. This question is especially pertinent with respect to political queries, given growing concerns about filter bubbles, and the recent finding that bias or favoritism in search rankings can influence voting behavior. In this study, we conduct a targeted algorithm audit of Google Search using a dynamic set of political queries. We designed a Chrome extension to survey participants and collect the Search Engine Results Pages (SERPs) and autocomplete suggestions that they would have been exposed to while searching our set of political queries during the month after Donald Trump's Presidential inauguration. Using this data, we found significant differences in the composition and personalization of politically-related SERPs by query type, subjects' characteristics, and date.

CCS CONCEPTS

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

KEYWORDS

Search engine results; search ranking bias; autocomplete search suggestions; political personalization; filter bubble

ACM Reference Format:
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, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3178876.3186143

1 INTRODUCTION

Recent concerns surrounding political polarization, fake news, and the impact of media on public opinion have largely focused on social media platforms like Facebook and Twitter. Yet, recent surveys suggest that more news is sought through search engines than social media [2, 57], and that search engines are the second-most likely news gateway to inspire follow-up actions, such as further searching, online sharing, or talking about the news with others [50]. Search engines are also reportedly the most trusted source of news [7], and the majority of search engine users perceive search engine rankings to be unbiased, accurate, and fair [66].

One behavioral corollary of this deeply rooted trust is the persistent, predictable top-to-bottom browsing pattern of search engine users [60]. Long-term studies of Click-through Rates (CTRs) on search engines have consistently shown that the top three search results receive over 50% of clicks, and 75% of clicks are made on the first Search Engine Results Page (SERP) [67]. Similar browsing patterns appear to occur on virtually any online platform where content is ranked [4, 19, 23, 31, 40]. These heuristic-driven behavioral patterns, known as *order effects*, are among the most robust effects ever discovered in the psychological and behavioral sciences, and imbue the entity ranking the content with the power to change attitudes, beliefs, and behavior [3, 20, 21, 34, 55].

Scholars and regulators have raised concerns about the potential negative effects that search engines can have on users, especially with respect to political information. One pertinent issue is that favoritism in politically-related search rankings can shift voting decisions [21, 22]. Another concern is that personalized rankings confine users in "filter bubbles" where the information that they are exposed to reflects and entrenches their existing beliefs [4, 61]. Jointly, these concerns revolve around the filtering, positioning, and display of information, and how these factors might systematically vary among users.

Algorithms that filter, rank, and shape information undoubtedly serve a crucial role in our ability to effectively navigate the internet, but given the empherality of their output and the aforementioned concerns, it is of critical importance to develop methods for preserving and quantifying their presentation of information [32, 36]. In this paper, we focus on politically-related searches conducted on Google Search and report the results of a controlled *algorithm audit* [68]. On Trump's inauguration, and for the four weeks that followed, we recruited between 14 and 46 subjects once a week (187 total) to complete a survey and install a Chrome extension that enabled us to conduct searches from within their browsers.

We seeded our extension with a set of 21 *root queries*, and designed it to obtain the Google autocomplete suggestions for each root query. For all 105 resulting queries (*i.e.,* each root and its four children) the extension simulataneously retrieved a pair of Google SERPs, one from a standard and one from an incognito browser window. These paired sets of results allowed us to isolate the impact of users' cookies (which are not used in the incognito window) on Google's ranking algorithms, and ask: given our set of root queries, how was the information available to Google searchers presented and personalized?

WWW 2018, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5639-8/18/04.
https://doi.org/10.1145/3178876.3186143

After parsing the autocomplete search suggestions and SERPs we collected through our extension, we found variance in the diversity of suggestions by root query, and identified a diverse set of SERP *components* that affected the filtering, positioning, and display of information. We also found that personalization, defined as a rank-weighted difference between the two sets of URLs collected from the standard and incognito paried SERPs, varied as a function of root query, political preferences, Alphabet service usage, and date.

Overall, our work makes the following contributions:

- We provide the first audit of desktop Google Search that considers the ranking and composition of the entire ranked column portion of the interface, and introduce a framework for quantifying patterns in the ranking (personalization) and display (composition) of information that may generalize across other platforms with ranked lists, like Facebook, Reddit, and Twitter.
- We confirmed that individuals logged in to their Google accounts receive greater personalization, presented evidence of temporal variance in the overall magnitude of personalization following an event, and combined search queries and their autocomplete suggestions into a novel root and children structure, demonstrating that these structures varied in their diversity by root.
- Our audit reveals substantial variance in personalization and composition by query and rank, sheds first light on previously unidentified components (*e.g.,* embedded Twitter results), and reveals the prominence of two component types at the top search rankings (knowledge and news-card), paving the way for future research on their featured content and its impact on users.

Outline. The rest of the study is organized as follows. We first review several previous algorithm audits conducted on Google Search (§ 2) and then introduce our own auditing methodology (§ 3). We then provide an overview of the survey, search, and suggestion data that we collected (§ 4), explore differences in SERP composition (§ 5), and measure how personalization varies on Google Search (§ 6). We conclude with a discussion of our results (§ 7) and limitations (§ 8).

2 BACKGROUND

The research technique known as the *algorithm audit* provides a useful framework for investigating the output of an algorithm and auditing it for potential biases [49, 68]. The logic of auditing techniques was borne out of social science research designed to identify discriminatory hiring practices [49], where researchers systematically varied an input (*e.g.,* the race of the applicant) and examined its influence on the output (*e.g.,* the likelihood of receiving a call). Utilizing this paradigm, researchers have conducted audits on algorithms in online marketplaces [12, 13, 33, 47, 48], search ranking bias on Twitter [38], user awareness of Facebook's NewsFeed algorithm [24], localization of online maps [69], and, most relevant here, search engine rankings [32, 36, 43, 45].

Google Search Audits. To the best of our knowledge the first Google Search audit was conducted in 2013 utilizing a survey and proxy-based data collection to measure search personalization [32]. Overall, Hannak *et al.* found (1) evidence of general personalization above a noise floor, (2) greater stability for highly ranked results, and (3) a lack of personalization based on a users' past browsing and search behaviors. They also reported greater personalization when their manufactured user accounts were logged-in and when political queries were used, and some personalization based on IP address geolocation [32].

Silver *et al.* conducted a similar study on geolocation-based personalization in the mobile version of Google Search [36]. However, unlike previous work, Silver *et al.* were the first to engage with the complex presentation of results in Google SERPs by specifically examining `general`, `news-triplet`, and `map` results. Silver *et al.*'s `general` results are the standard blue links (possibly with some accompanying explanatory text) that are shared by most search engines, while the `news-triplet` features one primary link to a news article with an image and two smaller links to news articles below it. The `map` result embeds Google Maps in the page. In this work, we refer to these discrete types of results as *components*.

More recently, researchers exploring localized versions of Google found that the extent to which a geographical region has developed publishing and scientific industries is strongly correlated with the locality of search results returned [5]. Another recent audit was conducted on the search engines of Google, Yahoo, and Bing during the 2016 US Congressional Elections [45]. Using politicians' names for queries, these researchers found that Google provided the most stability in terms of which domains occupied highly ranked positions, and concluded that such stability, compared to the relative instability of the same metric on Yahoo, demonstrated a robustness to outside attempts by spammers and marketers to game search rankings. However, these findings are limited because the searches appear to have been conducted without controls or real users, and no distinctions were made among component types.

One type of component whose presence and impact have been under-explored are the `knowledge` components that appear at the top of Google SERPs and attempt to directly provide an answer by drawing from Google's "Knowledge Graph" [46]. Unfortunately, a recent audit of these `knowledge` components revealed that their presence reduced the amount of traffic for websites that would have otherwise occupied the first ranking, most notably Wikipedia [43]. The presence of the `knowledge` components also made users more likely to attribute their discovery of the information to Google rather than to Wikipedia. In addition to concerns around traffic loss, there have been several documented cases of `knowledge` components highlighting controversial or untrue information. Examples include a list of US Presidents who were active members of the Ku Klux Klan and the answer to whether Obama was planning a coup [35, 52, 71].

While informative, the findings from these audits are limited due to their reliance on proxy-based or otherwise indirect data collection methodologies, a lack of controls for isolating personalization, or a focus on a subset of the components that appear in modern Google SERPs [5, 32, 36, 43, 45]. Furthermore, the measures that have been previously employed to quantify ranking similarity, like Jaccard index or edit distance, are limited in their ability to distinguish between ranking differences that occur towards the top of a ranked list and differences that occur towards the bottom. A

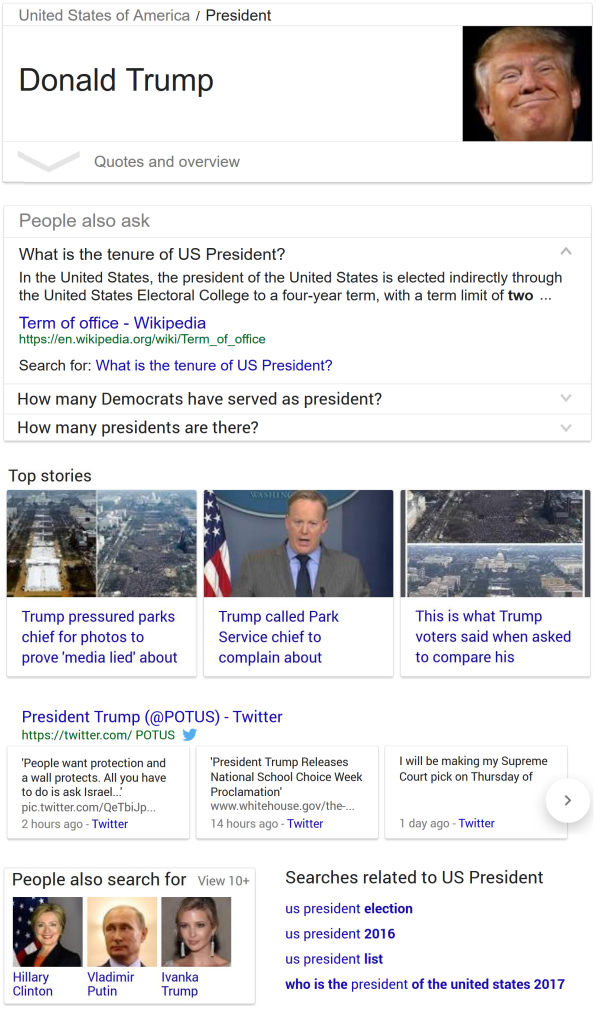

Figure 1: Examples of Google Search components. From top to bottom `knowledge`, `people-ask`, `news-card`, `twitter`, `people-search`, and `related-search` components. All components appear in their own row in SERPs but we have compressed them here.

more accurate measure of ranking similarity should incorporate the predictable browsing patterns of search engine users to weight changes in highly ranked items more than changes in lower ranked items [51, 74].

Google Search Suggestions. The algorithms that curate search suggestions could potentially wield control over the content that users' consume by leveraging heuristics like order effects in the ranking of suggestions, negativity bias in the valence of suggestion terms, and HTML effects like bolding to draw attention. Most of the existing literature on Google's autocomplete suggestions comes from informal publications aimed at manipulating the suggestions for a given query [70, 75] and identifying censorship and defamation in the suggestions of a set of targeted queries [16, 17]. Google states that their search predictions are based on factors including the terms you type, the popularity and freshness of those terms, your search and browsing histories, and trending topics in your area [29].

3 METHODOLOGY

We approached our audit of Google Search by considering not just the ranking of URLs [32, 36, 45] or the content of certain components [43], but the overall composition of the SERPs the platform produces, the URLs it provides, and the factors that shape each[1].

We conducted our experiment over the course of the five weeks following the inauguration of Donald Trump on January 19, 2017. On the inauguration, the day after, and once a week following, we posted recruitment ads on CrowdFlower (http://crowdflower.com) and Prolific Academic (http://prolific.ac), online subject pools that are comparable to the widely used Amazon Mechanical Turk (AMT; http://mturk.com) [6, 9, 65]. Recent research suggests that participants recruited from these platforms are more naïve and less dishonest than participants recruited on AMT [64]. On both websites we utilized built in features to restrict the visibility of our recruitment ads to participants within the US.

After participants provided informed consent, we asked them to complete a survey measuring their characteristics. Upon completing the survey, we asked them to install a Chrome browser extension we built and gave them a unique token. The token allowed them to launch the extension and enabled us to pair their search results to their survey responses. Below we describe the details of our survey, browser extension, and controls for isolating personalization.

Survey. Our survey included questions on demographics, Internet usage, and political preferences. Specifically, we asked about usage of Alphabet[2] services, political leaning and party affiliation, and ratings of the newly elected US president, Donald Trump. We asked participants "Is your overall opinion of Donald Trump positive or negative?" on both an 11-point Likert scale (which ranged from -5 to +5) as well as a binary rating (negative or positive). Both scales were counterbalanced.

Browser Extension. We built a custom Chrome extension that enabled us to automatically retrieve and preserve SERPs from within participants' browsers. After participants installed the extension and gave it the token from the survey, the extension opened two new browser windows, one standard and one incognito, and began conducting searches in the two windows in parallel from a predefined list of queries. The list contained 21 names of people, locations, and countries or groups that were potentially related to Donald Trump's inauguration (Table 1). For each query, each browser window opened a new tab, conducted the search, took a snapshot of the DOM, and closed the tab.

As each query in the extension's queue was conducted, the extension also retrieved the Google search suggestions for that query and appended them to the end of the search queue. This process was repeated for each root, generating a total of 105 queries and

[1]This study was IRB approved (Northeastern IRB #16-11-23) and summary data and code are available at http://personalization.ccs.neu.edu/
[2]The parent company of Chrome, Gmail, Google Search, YouTube, and other services, formerly known as just 'Google.'

Table 1: The root search queries we used.

Category	Root Query
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
Foreign Entities	China, Russia, Putin, UN

standard-incognito SERP pairs for each subject who completed all of the searches.

This approach allowed us to (1) utilize individuals' browsers – with their current cookies, logins, and search history intact – as a proxy through which to collect real-world personalized search data, (2) reduce potential noise due to temporal changes in Google's search index, and (3) pair each SERP we collected with an unpersonalized control SERP from the incognito window. While previous audits identified carry-over effects, in which previous queries (*e.g.,* "Hillary Clinton") can influence the SERP returned for new queres (*e.g.,* "Email") [32], Google's documentation indicates that carry-over effects should occur in both standard and incognito browser windows [28]. Thus, if these effects occur in our data, they should not affect our ability to isolate personalization.

4 DATA OVERVIEW

Here we provide an overview of the data that we collected. We begin with a description of our participants and then describe the composition of their SERPs.

Survey Data. In total, we recruited a demographically diverse sample of 187 participants from Prolific (74%) and Crowdflower (26%). Our sample was 46% female, predominantly White (66%) and Asian (17%), and 44% had a bachelor's degree or higher. Participants reported a median household income of $50,000 to $74,999 and their mean age was 32 (σ = 12.3). Politically, our sample leaned liberal (50%) and Democratic (47%). The mean rating of Donald Trump on the bipolar scale was -2.4 and 22% of participants gave him a "positive" rating on the binary scale. For comparison, President Trump's average approval rating during the period of time we conducted our study was 42.3% [25].

We asked subjects whether they were regular users of various Alphabet products [3], and 90% or more of participants reported using Gmail or YouTube regularly. The median number of Alphabet products that subjects reported regularly using was 4, and cumulatively, 96% of our participants were regular users of two or more products.

Subjects reported conducting a mean of 14.2 searches per day (σ = 16.9), and Google was by far the preferred search engine (88%), a result which is consistent with other surveys of search engine usage and preferences [15, 22]. 82% of our sample reported that Chrome was their preferred browser, with the closest competitor, Firefox, accounting for 12%. Google Search is the default search engine in both browsers, although Google may collect more information from Chrome users due to the browser's tight integration with Google services.

Search and Suggestion Data. In total, we collected the autocomplete search suggestions and standard-incognito SERP pairs for 15,337 queries. Among these queries, 3,624 were from our list of fixed roots and 11,713 were suggested by Google. The average query was 2.2 words long, which is comparable to previous findings [14, 37], given that 14 of our 21 root queries were one word long (Table 1).

Interestingly, there were large differences in the variability of suggestions among our root queries. We quantified this variability by normalizing the number of unique suggestions each root produced relative to the minimum number possible (4) and the maximum number we observed (69 for "China"). Thus a score of 0 (held only by "2017 US President") indicates that a root produced the exact same four suggestions for all subjects, and scores between 0 and 1 indicate the relative variability in the suggestions Google offered for that root (Figure 2, left-most column).

Utilizing the SERP pairs obtained from participants, we identified 14 unique result types and extracted 456,923 components and subcomponents (*e.g.,* the horizontal cards in a `news-card`) from these SERP pairs (Figure 1). Among the components we identified but have not yet mentioned are: `twitter` components that consist of a header linking to a Twitter account followed by three or more horizontal subcomponents that contain tweets from that account; `video` components that featured an embedded Youtube video; and slight variations of the `general` component that featured either a thumbnail of a Youtube video (`general-video`) or links to subdomains of the primary URL (`general-extra`). Collages of Google Image Search results appeared in `image` components, while results from Google Maps were featured in `map` components [36] and `map-card` components (which are similar to `news-card` components but feature locations instead of news). Finally, both the `people-search` and `related-search` components featured a set of suggested queries, though their formatting was different (Figure 1).

On average, the standard and incognito windows both returned a nearly identical number of components and subcomponents per SERP (both $\mu = 14.9, \sigma = 3.5$). We found small differences in the number of components between paired SERPs, with 15.3% of pairs differing by at least one component and 6.5% of pairs differing by four or more components, but a Wilcoxon signed-rank test between the paired SERPs was not significant ($V = 1.33 \cdot 10^6, P = 0.07$). The smallest SERP we found had 7 components or subcomponents and the largest had 25.

While not the focus of this paper, we note that Google's side bar (as identified in [43]) appeared in approximately 69% of the SERPs we collected. Using McNemar's paired samples χ^2 test we found a small but significant number of SERP pairs (1.3%) in which the side bar appeared in one but not both of standard-incognito paired SERPs ($\chi^2 = 78.000, p < 0.01$). The presence of the side bar varied widely by root query, with "Trump inauguration" and "Donald" producing the lowest and highest percentage of SERPs that contained a side bar, respectively (Figure 2, second column from the left). However, when grouped by root query, none of the paired differences were significant, which suggests that the presence or

[3]We defined regular usage as once a week or more, and asked about usage of Android, Gmail, Google Calendar, Google Docs, Google Drive, Google+, Google Groups, Google Maps, and YouTube.

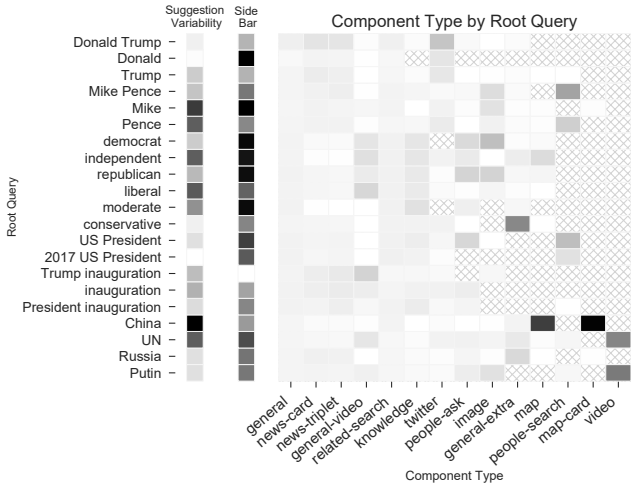

Figure 2: Composition of standard SERPs by root query. Components along the x-axis are sorted by their ubiquity (the number of root queries they appeared in). Cells filled with a gray crisscross pattern indicate that a component never appeared for that root query.

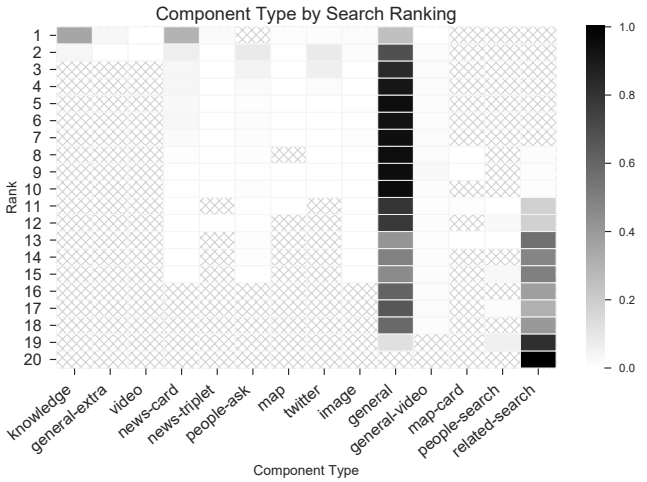

Figure 3: Composition of standard SERPs by rank. Components along the x-axis are sorted by their median rank. Cells filled with a gray crisscross pattern indicate that a component never appeared at that rank.

absence of the side bar is not heavily personalized. A study focused on the content within the side bar could reveal additional aspects of personalization (*e.g.,* [43]).

5 SERP COMPOSITION

In this section we provide an analysis of the trends that shaped the composition of the SERPs we collected. For now, we focus on the subset of standard SERPs and leave comparative analysis of the standard and incognito SERPs to § 6. Overall, we found that the probability of receiving different component types varies by the root query that produced it, and that components differ by their typical rank position.

Components by Root. Using the subset of standard SERPs, we calculated the probability distribution of each component type across our root queries. We found substantial differences in SERP composition by root query, with `twitter` components appearing the most frequently in SERPs for queries that stemmed from the root "Donald Trump," `knowledge` components occurring most frequently for queries stemming from political party and ideology name roots, and `people-search` components occurring the most frequently for queries stemming from "Donald Trump" and "Mike Pence" (Figure 2).

Components by Rank. To assess how components varied by rank we utilized a position probability matrix to find the probabilty of finding each component type given each rank (Figure 3). Using this matrix we found that `knowledge` and `news-card` were the most highly ranked components, accounting for over 60% of all components appearing at the first rank (composing 35% and 29.8%, respectively). Most other rank positions were dominated by `general` results, with the `people-search` and `related-search` components almost exclusively occupying the SERP footer. Given the disproportionate amount of clicks and attention that go to the first result [60], our finding with respect to the prominence of `knowledge` and `news-card` components at the first rank suggests that the reported cases of untrue or controversial information likely had high exposure to searchers examining those topics [35, 52, 71].

Domains. A 2016 report on sources of website traffic found that Google accounted for 39.5% of all referrals to publishers' websites [62]. Given the high rate of traffic referrals and the previous findings with respect to Wikipedia [43], we extracted [4] from each SERP the URLs present in the titles of Google's direct answer (`knowledge` and `people-ask`), `general`, and media components (`news-card`, `news-triplet`, and `twitter`) and grouped them by their respective 2^{nd}-level domain names (*e.g.,* cnn.com).

Among all component types we found that the top 20% of the domains (n = 724) accounted for 96.1% of all domains we found. This inequality in domain presence was also present among individual components as well (Table 2). Though both the `news-card` and `news-triplet` components serve similar functions–displaying news articles–the top domains they surfaced were substantially different, suggesting differences in their underlying algorithms. This finding is bolstered by our observation that `news-card`s and `news-triplet`s were shown in response to similar root queries (Figure 2), meaning that the differences in top domains were not caused by differing queries.

[4] We extracted the URLs featured in the title of most results (*e.g.,* `general`, `news-cards`, `news-triplets`, and `knowledge` components). For `twitter` components, we extracted the account URL from the component header, and the first URL found in each subcomponent (tweets). For `people-ask` components we extracted only the first URL.

Table 2: The top ten most frequently occurring domains for six of the component and subcomponent types.

knowledge (n = 9, 029)	%	people-ask (n = 7, 050)	%	general (n = 307, 275)	%	news-card (n = 49, 188)	%	news-triplet (n = 3, 637)	%	twitter-card (n = 32, 429)	%
no URL	65.5	no URL	80.8	en.wikipedia.org	8.7	nytimes.com	6.5	nytimes.com	10.5	twitter.com	79.8
en.wikipedia.org	12.2	en.wikipedia.org	5.5	nytimes.com	4.2	cnn.com	6.4	cnn.com	5.8	ind.pn	4.0
books.google.com	5.2	enkivillage.com	1.2	twitter.com	3.0	foxnews.com	3.7	npr.org	4.6	buff.ly	1.6
dictionary.com	2.4	infoplease.com	1.0	cnn.com	2.6	washingtonpost.com	2.8	telegraph.co.uk	4.3	bit.ly	1.5
depressionet.org.au	2.3	timeanddate.com	0.9	theatlantic.com	2.6	nbcnews.com	2.8	abcnews.go.com	4.1	conservativereview.com	1.5
grammar-monster.com	2.0	al.com	0.8	facebook.com	2.2	usatoday.com	2.6	usatoday.com	3.6	en.kremlin.ru	1.1
careers.un.org	1.6	nationalistpartyamerica.com	0.8	washingtonpost.com	1.8	bbc.com	2.4	foxnews.com	3.2	conservativetribune.com	1.1
historyinpieces.com	0.9	globalpolicy.org	0.7	time.com	1.4	telegraph.co.uk	2.3	cbsnews.com	3.1	miamiherald.com	1.1
owl.english.purdue.edu	0.8	indy100.independent.co.uk	0.6	usatoday.com	1.4	politico.com	1.9	bbc.com	2.9	45.wh.gov	1.0
factmonster.com	0.8	aims.edu	0.6	merriam-webster.com	1.2	npr.org	1.9	reuters.com	2.2	instagram.com	0.9

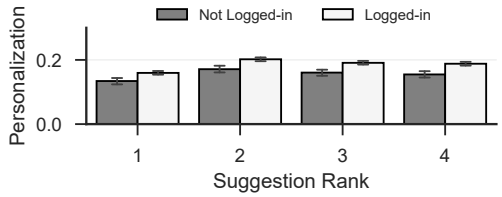

Figure 4: Personalization averaged over all suggestion ranks. Error bars represent 95% CIs.

6 PERSONALIZATION

To measure personalization, we used the list of URLs we extracted from each SERP and then utilized a ranking similarity metric called Rank-Biased Overlap (RBO) [74] to compare the lists found on each standard-incognito SERP pair.

We utilized RBO because it provides an *indefinite rank similiarity measure* that is particularly suitable for the comparison of search engine rankings. It accounts for several important aspects of rank comparisons that other commonly employed rank similarity measures, including Kendall's τ and Spearman's ρ, do not [41, 51, 74]. Specifically, RBO accounts for (1) *top-weightedness*, by imposing a stronger penalty for differences at the top of the rankings, (2) *incompleteness*, by handling lists containing different items without assuming underlying conjointness, and (3) *indefiniteness*, by limiting the weight of unseen items in the conceptually infinite tail [74].

RBO takes a parameter p that determines the top-weightedness of the metric. If $p = 0.9$, then the first 10 ranks account for 86% of the evaluation [74]. To set p we first obtained CTR data from January 2017 and found that the first 13 ranks (the average number of components in our dataset) accounted for 80.4% of clicks [67]. We used this to find a $p = 0.938$ that attributed 80.4% of the weight of the RBO evaluation to the first 13 ranks[5].

Given that RBO is a measure of rank similarity, we used it to define personalization for a pair of SERPs as:

$$1 - RBO(URLs_{incognito}, URLs_{standard}). \quad (1)$$

Thus personalization can range between 0 and 1, with 0 indicating that the given standard and incognito SERPs are identical[6].

[5]See Equation 21 in the original Webber *et al.* paper [74] for additional details on *RBO* and how to set p.

[6]We found that RBO was fairly strongly correlated with personalization metrics operationalized in previous work, including jaccard index ($\rho = 0.739$, $p < 0.001$), a 0 to 1 bounded measure of similarity (the intersection over the union of two sets), and edit distance ($\rho = -0.656$, $p < 0.001$), an integer measure of ranking differences (the number of swaps, insertions, and deletions).

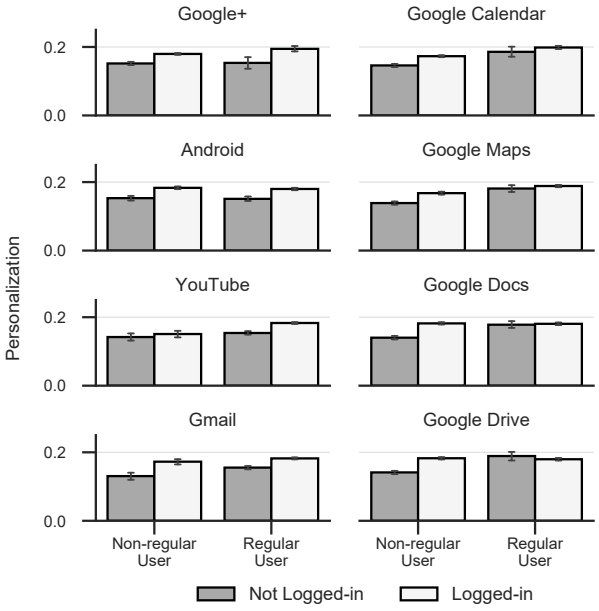

Figure 5: Average personalization with 95% CIs by participants' Google account login status (colored bars) and their regular usage of individual Alphabet services (x-axes). Only four participants reported regularly using Google Groups, and all were logged-in during the audit, so we omitted it from the figure. For the bottom three services in the right column, regular users received the same level of personalization regardless of their login status.

Personalization by Search Suggestion. Given that autocomplete is personalized [29], we checked whether personalization varies by the rank ordering of the search suggestions that produced the paired SERPs (Figure 4). Using a nonparametric Kruskal-Wallis χ^2 test we found significant differences in personalization by suggestion rank, with SERPs produced at the second rank 26.7% higher than the SERPs produced at the first rank. However, the number of words composing a query was not significantly correlated with personalization and the number of characters was only weakly correlated with personalization ($\rho = 0.05, p < 0.001$). This may

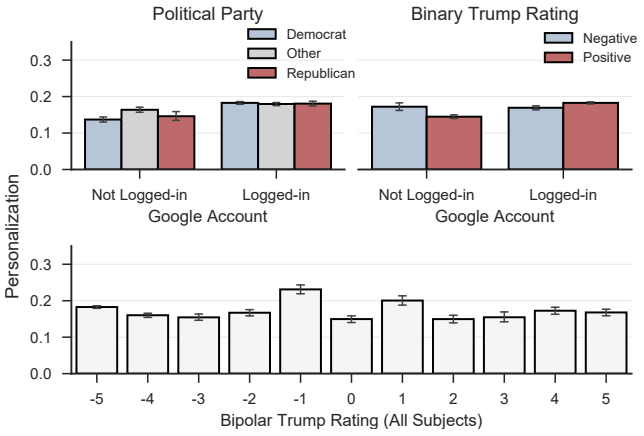

Figure 6: Average personalization and 95% CIs by participants' ratings of Donald Trump and their political party.

suggest some importance of the second search suggestion for the autocomplete algorithm, but data on how people browse and select search suggestions are largely unavailable.

Personalization by Google Usage. We found that personalization on Google Search increased with the amount of Alphabet services that participants reported regularly using ($\rho = 0.07, p < 0.001$) and was 19.3% higher for participants who were logged-in to their Google accounts than for those who were not ($U = 1.52 \cdot 10^7, p < 0.001$).

Among the Alphabet services that we asked participants to indicate if they regularly used or not, we found that regular usage of a subset of services increased personalization regardless of the participants' Google account login status (Figure 5). While non-regular users who were logged-in to their Google account received significantly greater personalization for all services than those who were not (all $p < 0.001$ except for regular users of Youtube where $p < 0.05$), but among regular users, the difference in the magnitude of personalization for logged-in and logged-out users was not significant for Google Docs ($U = 2.43 \cdot 10^6, p = 0.29$), Drive ($U = 1.69 \cdot 10^6, p = 0.13$), or Maps ($U = 3.38 \cdot 10^6, p = 0.07$). This finding suggests that regular usage of services in which highly personal data can be mined is associated with greater personalization.

Personalization by Political Preference. We found significant differences in personalization by participants' bipolar ($\chi^2 = 180.739, p < 0.001$), and binary ($U = 1.85 \cdot 10^7, p < 0.001$) ratings of Trump, but not by their political party ($\chi^2 = 0.269, p = 0.87$). Participants who provided low-strength ratings of Trump on the bipolar scale, in either direction (-1 or 1), received significantly more personalization in their rankings than any other category on our bipolar scale (Figure 6). This finding is especially intriguing given that previous studies deploying the same type of rating scale to measure political opinions found that people who made these low-preference ratings (-1 or 1) were the most susceptible to having their preferences influenced by biased search rankings (*e.g.,* the Search Engine Manipulation Effect [SEME]) [21, 22].

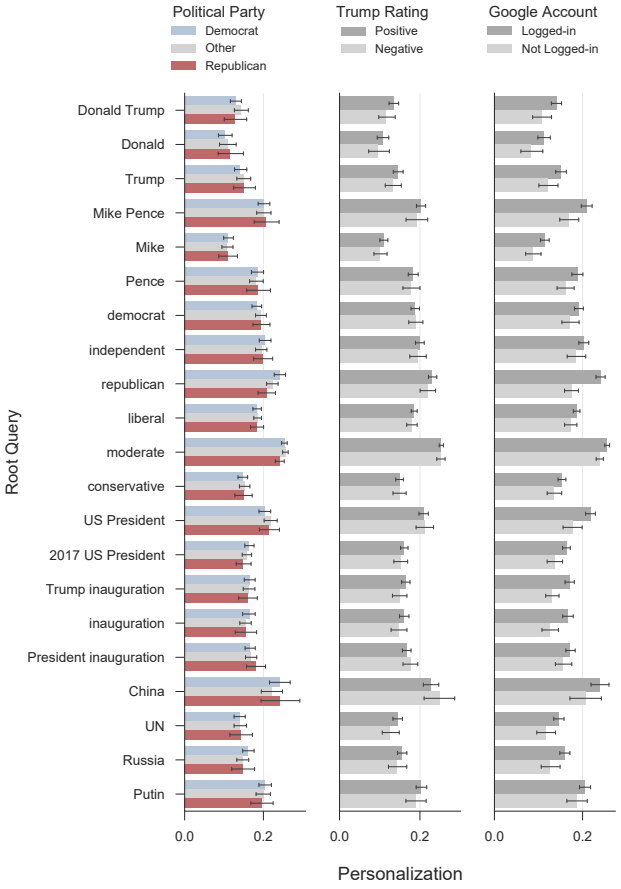

Figure 7: Average personalization with 95% CIs for each root query by participants' characteristics.

Personalization by Root Query. After aggregating the SERP pairs for each root and its child suggestions, we found substantial differences in personalization by root (Figure 7), with the highest personalization–among all participants–occurring for the roots "republican," "moderate," and "China." Higher personalization for searches stemming from the root "China" might be explained in part by the variability we found in Google's search suggestions for the root "China," and the disproportionate prevalence of map components in the SERPs produced by the root "China" (Figure 2).

We examined how personalization by root query might vary by participants' political characteristics (political party and Trump rating) but found no significant differences (Figure 7). Whether or not a participant was logged-in to a Google account appears to be the biggest driver of personalization, regardless of the root query, a finding consistent with past Google audits [32]).

Temporal Dynamics. We found a negative correlation between personalization and date ($\rho = -0.12, p < 0.001$) and significantly more personalization on January 20 than any other day

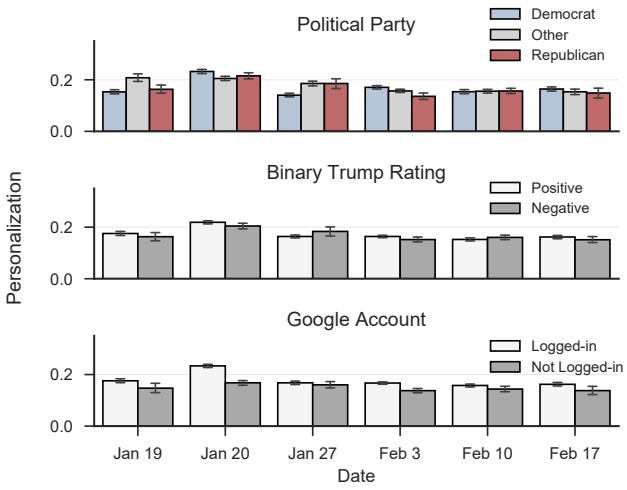

Figure 8: Personalization by data collection date, political view, political party, binary rating of President Trump, and Google account login.

in our dataset ($\chi^2 = 364.368, p < 0.001$) suggesting that personalization was highest the day immediately following the inauguration. Compared to the last day that we collected data, February 17, participants experienced 37.6% greater personalization on January 20. This finding suggests that the amount of personalization varies temporally, possibly as a result of algorithmic changes or changes in Google's information corpus as breaking news stories emerge.

Among political party affiliations we found significant differences in the amount of personalization for each date except February 10 and February 17 (Figure 8). The level of personalization was more negatively correlated with the date of data collection for Republicans ($\rho = -0.142, p < 0.001$) and Others ($\rho = -0.135, p < 0.001$) than it was for Democrats ($\rho = -0.101, p < 0.001$), suggesting that the increased personalization observed following the inauguration was highest for Republicans. With respect to participants' binary ratings of Trump, we only found significant differences in personalization on January 20, where participants who gave Trump a negative rating received marginally more personalization than participants who gave him a positive rating ($U = 1.04 \cdot 10^6, p < 0.05$).

We again found the largest differences when we examined personalization by date and participants' login status. Participants who were logged-in to their Google accounts during the audit received significantly greater levels of personalization than participants who were not logged-in on all days except January 27 (Figure 8). Personalization significantly decreased over time for logged-in participants ($\rho = -0.131, p < 0.001$) but less so for participants who were not logged-in ($\rho = -0.077, p < 0.001$). Similar to the binary Trump ratings, the greatest difference in personalization by login status occurred on the day of the inauguration, where participants who were logged-in received 39% more personalization compared to participants who were not logged-in ($U = 9.48 \cdot 10^5, p < 0.001$).

7 DISCUSSION

We conducted a targeted algorithm audit of Google's ranking interface by utilizing a browser extension to obtain real search data relevant to a major US event (the inauguration). The results of our audit shed light on the largely ignored diversity in the presence and absence of SERP components (Figure 2), their differential prominence in the search rankings (Figure 3), and their domain filtering (Table 2). Using the inverse of the RBO rank similarity metric [74] to quantify personalization, we showed that personalization in politically-related searches conducted on Google is (1) relatively low, (2) dependent on query selection, (3) higher for subjects who are logged-in to a Google account, and (4) fluctuates substantially over time, perhaps in response to events that generate press, like the inauguration. Such personalization, in combination with the influence that search engines wield over users [1, 22, 42, 59], could have important implications for policy-makers concerned with algorithmic accountability and transparency [18, 58, 63].

Among all subjects, we found that personalization was substantially higher for subjects who used more Alphabet products, and it appears that regular usage of certain Alphabet products (Google Drive, Google Docs, and Google Maps in particular) are associated with heightened personalization. We also found that low-strength ratings of Trump on the bipolar scale in either direction (-1 or 1) received significantly more personalization in their rankings than any other category (Figure 6). This finding has potentially interesting implications given that a recent experiment using a similar scale demonstrated how subjects with low-strength political opinions were the most susceptible to the influence of search ranking bias [22].

Examining personalization by date, we found that personalization decreased over the course of our data collection window. It is possible that the large increase in personalization we measured on the day following the inauguration (Figure 8) was due to interactions between the information produced by the event and Google's algorithms, but because of our limited data collection window we cannot say anything about the patterns that might have occurred before the event or after our data collection had ended.

In terms of composition, we found that Google's `knowledge` and `news-card` components accounted for 64.8% of all components seen at the first search ranking. Given the trust associated with, and clicks accrued by the first ranked search result [60], this finding suggests that the recent cases of untrue or controversial information surfacing in these components likely increased their dissemination [35, 52, 71]. While researchers interested in the study of misinformation or fake news have thus far focused primarily on social media, our results point out a diverse ecosystem of information presentation that, in combination with the trust placed in search engines [7, 60], could increase exposure to and consumption of misinformation. Future work on misinformation may want to consider how search engines and highly ranked components featuring direct answers (`knowledge` and `people-ask`) and news or social media (`twitter`, `news-card`, and `news-triplets`) might contribute to the spread of untrue information [2, 30, 39, 44, 73]. Similarly, future behavioral experiments on the influence of search rankings on beliefs and behavior should explore how the diversity of components

we identified (*e.g.,* `knowledge`, `news-card`, and `people-ask`) could be used to enhance effects like SEME [21, 22, 59].

In terms of URL domains, we found that `knowledge` components typically do not feature a URL (Table 2), and when they do it is often a link to Wikipedia, confirming a finding from a previous audit [43]. Given the concerns that `knowledge` components can reduce traffic to Wikipedia, or other sites that might have otherwise occupied the top rank, it is concerning that we were unable to extract a featured URL for 65.5% of `knowledge` components in our data set, even though it occupied the first rank in 35% of all SERPs.

We found that Twitter components appeared more frequently for searches of "Donald Trump" than for any other query (Figure 2), and when these components did appear in a SERP, they typically appeared within the first three rankings (Figure 3). It is possible that the President's tweeting habits are responsible for this association, which was less prevalent for the root "Mike Pence." Further research is needed to understand the dynamics between social media components, the web domains they surface, and user information consumption and decision-making. While Google is able to enforce a measure of quality upon the results they present in other components, but the factors they use to filter Tweets in search is unknown, though the embedded Tweets appear to be sorted temporally.

Our results direct attention to the dearth of research on ranking interfaces like search engines that are used by billions of individuals every day, and highlight the need to consider not just how information is ranked but also how it is formatted. *Primacy effects*, which guide user attention and behavior toward items placed at the top of the screen [10, 19, 26, 27, 56] may only be part of the puzzle, and future work is needed to understand how people use modern search engine interfaces and interact with their various components.

Our focus was on the display and ranking of information, and therefore we did not explore many promising avenues of leveraging the text content of SERPs and the pages they link to. Text features on a SERP, such as HTML bolding (*e.g.,* `<b>text</b>`) or any other such text accentuation may shed additional light on the content and importance of text in search results. Mining and modeling the linguistic characteristics of search results and their corresponding webpages to obtain bias scores or develop topic models are also interesting future directions. Given the constant changes in the composition and ranking factors that produce SERPs, audits such as the one we conducted here will need to be conducted with some regularity in order to keep our understanding up to date.

8 LIMITATIONS

The findings of our audit are limited to searches conducted on the desktop version of Google Search. Unlike previous research [32], we did not attempt to avoid carry-over effects. For the research question we sought to answer – the extent to which individual and group factors influence the algorithmic ranking and composition of politically-related search results – we considered such differences to be a part of an individual's personalization experience and only measured the differences between individuals' search results in the standard and incognito browser windows. We therefore rely on Chrome's incognito mode to de-personalize web search by withholding users' cookies [8, 11, 53]. There is support for this notion on Google's Chrome privacy page, where it states that "Chrome won't share existing cookies with sites you visit in incognito or guest mode" [28], but it is unclear what this precisely means.

We quantified personalization at the individual level, measuring the ranking diferences between the lists of URLs collected from the SERPs generated by the standard and incognito browser windows. That is, our controls were paired within the individual, enabling us to isolate the impact that their browser mode had on their search rankings for each query we searched. Therefore we had to make the assumption that our results could be reasonably aggregated across days in order to compare groups. We recommend that future research focus on controlled compairisons between paired groups. For example, pairing political searches across party members and simulating the searches in their browsers in parallel. Such an investigation would also need to pair participants by both party membership and geographical region, raising recruitment and coordination challenges.

Our usage of autocomplete was useful for generating queries related to our root queries, but it is possible that not every query generated for each root was unambiguously related to our topic of interest. Furthermore, in our comparisons by root query, we assumed that the SERPs generated by a root query and its suggestions can be meaningfully aggregated. This makes sense given our focus on the information pathways that were available to a user at the time of the audit, but differences in the suggestions returned for each user and each query could have impacted our results. Our results demonstrated that query choice does not only impact content, but also impacts composition and personalization in seemingly systematic ways, and future audits should develop methods for investigating the suggestion structures we introduced.

Our audit was designed to survey the ways in which Google Search constructs information pathways, consisting of query suggestions and SERPs, that steer users towards certain pieces of information given a starting ngram query. However, we did not capture how users might have interacted with these pathways. Although it is well established that the highest rankings receive a disproportionate amount of traffic and attention [60], it is unclear whether this aggregate behavioral pattern varies among individual users or groups, or how the various component types might nteract with order effects. It is also unclear how these SERPs might have affected users' decision-making [21, 22]. Future research could incorporate various types of user studies with our data collection method to potentially answer these questions [1, 21, 59, 60, 72], and should investigate possibilities for suppressing unwanted influences with design interventions [22, 42, 54].

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their helpful comments and Piotr Sapieżyński, Devin Gaffney, Brendan Nyhan, Andrew Guess, Luke Horgan, Anikó Hannák and others for invaluable discussions 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] Ahmed Allam, Peter Johannes Schulz, and Kent Nakamoto. 2014. The impact of search engine selection and sorting criteria on vaccination beliefs and attitudes: two experiments manipulating Google output. *Journal of Medical Internet Research* 16, 4 (2014), e100.

[2] 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

[3] Solomon E Asch. 1946. Forming impressions of personality. *The Journal of Abnormal and Social Psychology* 41, 3 (1946), 258–290.

[4] Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132.

[5] Andrea Ballatore, Mark Graham, and Shilad Sen. 2017. Digital hegemonies: the localness of search engine results. *Annals of the American Association of Geographers* (2017), 1–22.

[6] Adam J Berinsky, Gregory A Huber, and Gabriel S Lenz. 2012. Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk. *Political Analysis* 20, 3 (2012), 351–368.

[7] Edelman Berland. 2017. 2017 Edelman Trust Barometer. http://www.edelman.com/trust2017/. (2017). Accessed: 2017-03-07.

[8] SEO Blog. 2014. How to Turn Off Google Personalized Search Results. http://www.seoblog.com/2014/09/turn-google-personalized-search-results/. (2014). Accessed: 2017-10-01.

[9] 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.

[10] Georg Buscher, Edward Cutrell, and Meredith Ringel Morris. 2009. What do you see when you're surfing?: using eye tracking to predict salient regions of web pages. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 21–30.

[11] Traffic Generation Cafe. 2017. Google Incognito: How to Search Google Without Being Tracked. https://trafficgenerationcafe.com/search-engine-ranking-tip-incognito-search/. (2017). Accessed: 2017-10-01.

[12] Le Chen, Alan Mislove, and Christo Wilson. 2015. Peeking Beneath the Hood of Uber. In *Proceedings of the 2015 ACM Conference on Internet Measurement*.

[13] Le Chen, Alan Mislove, and Christo Wilson. 2016. An Empirical Analysis of Algorithmic Pricing on Amazon Marketplace. In *Proceedings of the 25th International World Wide Web Conference*.

[14] Chitika. 2012. *Ask.com Has The Most Long-Winded Searchers, Report Says*. Technical Report. Chitika. http://searchengineland.com/ask-com-has-the-most-long-winded-searchers-report-says-109202

[15] Inc comScore. 2017. comScore Explicit Core Search Query Report (Desktop Only). https://www.comscore.com/Insights/Rankings. (2017). Accessed: 2017-02-12.

[16] Nick Diakopoulos. 2013. Algorithmic defamation: The case of the shameless autocomplete. http://www.nickdiakopoulos.com/2013/08/06/algorithmic-defamation-the-case-of-the-shameless-autocomplete/. (2013).

[17] Nick Diakopoulos. 2013. Sex, Violence, and Autocomplete Algorithms: Methods and Context. http://www.nickdiakopoulos.com/2013/08/01/sex-violence-and-autocomplete-algorithms-methods-and-context/. (2013).

[18] Nicholas Diakopoulos. 2014. Algorithmic accountability reporting: On the investigation of black boxes. *Tow Center for Digital Journalism, Columbia University* (2014).

[19] Dimitar Dimitrov, Philipp Singer, Florian Lemmerich, and Markus Strohmaier. 2016. Visual positions of links and clicks on wikipedia. In *Proceedings of the 25th International Conference Companion on World Wide Web*. International World Wide Web Conferences Steering Committee, 27–28.

[20] Hermann Ebbinghaus. 1913. *Memory: A contribution to experimental psychology*. Number 3. University Microfilms.

[21] 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.

[22] 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 2.

[23] Eyal Ert and Aliza Fleischer. 2016. Mere Position Effect in Booking Hotels Online. *Journal of Travel Research* 55, 3 (2016), 311–321.

[24] Motahhare Eslami, Amirhossein Aleyasen, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. Feedvis: A path for exploring news feed curation algorithms. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing*.

[25] Gallup. 2017. Presidential Approval Ratings – Donald Trump. http://news.gallup.com/poll/203198/presidential-approval-ratings-donald-trump.aspx. (2017). Accessed: 2017-10-08.

[26] Florian Geigl, Kristina Lerman, Simon Walk, Markus Strohmaier, and Denis Helic. 2016. Assessing the Navigational Effects of Click Biases and Link Insertion on the Web. In *Proceedings of the 27th ACM Conference on Hypertext and Social Media*. ACM, 37–47.

[27] David F Gleich, Paul G Constantine, Abraham D Flaxman, and Asela Gunawardana. 2010. Tracking the random surfer: empirically measured teleportation parameters in PageRank. In *Proceedings of the 19th International Conference on World Wide Web*. ACM, 381–390.

[28] Google. 2017. Google Chrome Privacy Notice. https://www.google.com/chrome/browser/privacy/#browser-modes. (2017). Accessed: 2017-04-01.

[29] Google. 2017. Search using autocomplete. https://support.google.com/websearch/answer/106230. (2017). Accessed: 2017-04-01.

[30] Andrew Guess, Brendan Nyhan, and Jason Reifler. 2018. Selective Exposure to Misinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign. (2018).

[31] Xunhua Guo, Mingyue Zhang, Chenyue Yang, et al. 2016. Order Effects in Online Product Recommendation: A Scenario-based Analysis. In *Proceedings of the 22nd Americas Conference on Information Systems*.

[32] 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*.

[33] Aniko Hannak, Gary Soeller, David Lazer, Alan Mislove, and Christo Wilson. 2014. Measuring Price Discrimination and Steering on E-commerce Web Sites. In *Proceedings of the 2014 ACM Conference on Internet Measurement*.

[34] Robin M Hogarth and Hillel J Einhorn. 1992. Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology* 24, 1 (1992), 1–55.

[35] Adrianne Jeffries. 2017. Google's featured snippets are worse than fake news. https://theoutline.com/post/1192/google-s-featured-snippets-are-worse-than-fake-news. (2017). Accessed: 2017-10-08.

[36] 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*.

[37] Certified Knowledge. 2012. *Ask.com Has The Most Long-Winded Searchers, Report Says*. Technical Report. Certified Knowledge. http://certifiedknowledge.org/blog/are-search-queries-becoming-even-more-unique-statistics-from-google/

[38] 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)*.

[39] 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

[40] Kristina Lerman and Tad Hogg. 2014. Leveraging position bias to improve peer recommendation. *PloS one* 9, 6 (2014), e98914.

[41] Xiaolu Lu, Alistair Moffat, and J Shane Culpepper. 2016. The effect of pooling and evaluation depth on IR metrics. *Information Retrieval Journal* 19, 4 (2016), 416–445.

[42] Ramona Ludolph, Ahmed Allam, and Peter J Schulz. 2016. Manipulating Google's Knowledge Graph box to counter biased information processing during an online search on vaccination: application of a technological debiasing strategy. *Journal of Medical Internet research* 18, 6 (2016).

[43] 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).

[44] Nicco Mele, David Lazer, Matthew Baum, Nir Grinberg, Lisa Friedland, Kenneth Joseph, Will Hobbs, and Carolina Mattsson. 2017. Combating fake news: An agenda for research and action. (2017).

[45] P Takis Metaxas and Yada Pruksachatkun. 2017. Manipulation of Search Engine Results during the 2016 US Congressional Elections. In *ICIW*.

[46] Peter J. Meyers. 2013. 101 Google Answer Boxes: A Journey into the Knowledge Graph. https://moz.com/blog/101-google-answer-boxes-a-journey-into-the-knowledge-graph. (2013). Accessed: 2017-10-08.

[47] Jakub Mikians, László Gyarmati, Vijay Erramilli, and Nikolaos Laoutaris. 2012. Detecting price and search discrimination on the Internet. In *Proceedings of the 11th ACM Workshop on Hot Topics in Networks*.

[48] Jakub Mikians, László Gyarmati, Vijay Erramilli, and Nikolaos Laoutaris. 2013. Crowd-assisted search for price discrimination in e-commerce: First results. In *Proceedings of the Ninth ACM Conference on Emerging Networking Experiments and Technologies*.

[49] 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.

[50] Amy Mitchell, Jeffrey Gottfried, Elisa Shearer, and Kristine Lu. 2017. *How Americans Encounter, Recall and Act Upon Digital News*. Technical Report. Pew Research Center.

[51] Alistair Moffat and Justin Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)* 27, 1 (2008), 2.

[52] Motherboard. 2017. Go Ahead, Ask Google 'What Happened to the Dinosaurs'. https://motherboard.vice.com/en_us/article/pga4wg/go-ahead-ask-google-what-happened-to-the-dinosaurs. (2017). Accessed: 2017-10-08.

[53] Moz. 2012. Face-off - 4 Ways to De-personalize Google. https://moz.com/blog/face-off-4-ways-to-de-personalize-google. (2012). Accessed: 2017-10-01.

[54] Sean A Munson, Stephanie Y Lee, and Paul Resnick. 2013. Encouraging Reading of Diverse Political Viewpoints with a Browser Widget.. In *ICWSM*.

[55] Bennet B Murdock. 1962. The serial position effect of free recall. *Journal of Experimental Psychology* 64, 5 (1962), 482–488.

[56] Jamie Murphy, Charles Hofacker, and Richard Mizerski. 2006. Primacy and recency effects on clicking behavior. *Journal of Computer-Mediated Communication* 11, 2 (2006), 522–535.

[57] Nic Newman, Richard Fletcher, Antonis Kalogeropoulos, David A. L. Levy, and Rasmus Kleis Nielsen. 2017. *Reuters Institute Digital News Report 2017*. Technical Report. Reuters Institute for the Study of Journalism.

[58] Helen Nissenbaum. 1996. Accountability in a computerized society. *Science and Engineering Ethics* 2, 1 (1996), 25–42.

[59] Alamir Novin and Eric Meyers. 2017. Making sense of conflicting science information: Exploring bias in the search engine result page. In *Proceedings of the 2017 Conference on Human Information Interaction and Retrieval*. ACM, 175–184.

[60] Bing Pan, Helene Hembrooke, Thorsten Joachims, Lori Lorigo, Geri Gay, and Laura Granka. 2007. In google we trust: Users' decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication* 12, 3 (2007), 801–823.

[61] Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you*. Penguin UK.

[62] Parse.ly. 2017. *The authority report: How audiences find articles, by topic*. Technical Report. Parse.ly. http://learn.parsely.com/rs/314-EBB-255/images/authority-report-13.pdf

[63] Frank Pasquale. 2015. *The black box society: The secret algorithms that control money and information*. Harvard University Press, Cambridge, MA.

[64] 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.

[65] Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. 2014. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior research methods* 46, 4 (2014), 1023–1031.

[66] Kristen Purcell, Joanna Brenner, and Lee Rainie. 2012. *Search engine use 2012*. Technical Report. Pew Research Center's Internet and American Life Project.

[67] Advanced Web Ranking. 2017. CTR Study February 2017. https://www.advancedwebranking.com/cloud/ctrstudy. (2017). Accessed: 2017-04-01.

[68] 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*.

[69] Gary Soeller, Karrie Karahalios, Christian Sandvig, and Christo Wilson. 2016. MapWatch: Detecting and Monitoring International Border Personalization on Online Maps. In *Proceedings of the 25th International World Wide Web Conference*.

[70] Lauren Starling. 2013. How to remove a word from Google autocomplete. (2013). http://www.laurenstarling.org/how-to-remove-a-word-from-google-autocomplete/

[71] Danny Sullivan. 2017. Google's "One True Answer" problem âĂŤ when featured snippets go bad. https://searchengineland.com/googles-one-true-answer-problem-featured-snippets-270549. (2017). Accessed: 2017-10-08.

[72] Sander van der Linden, Anthony Leiserowitz, Seth Rosenthal, and Edward Maibach. 2017. Inoculating the public against misinformation about climate change. *Global Challenges* 1, 2 (2017).

[73] 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

[74] William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)* 28, 4 (2010), 20.

[75] Wiideman. 2010. Beat the autocomplete - A study of Google auto-suggest. (2010). https://www.wiideman.com/research/google-autocomplete/study-results