The Matter of Chance: Auditing Web Search Results Related to the 2020 U.S. Presidential Primary Elections Across Six Search Engines

Urman, Aleksandra ; Makhortykh, Mykola ; Ulloa, Roberto

Abstract: We examine how six search engines filter and rank information in relation to the queries on the U.S. 2020 presidential primary elections under the default—that is nonpersonalized—conditions. For that, we utilize an algorithmic auditing methodology that uses virtual agents to conduct large-scale analysis of algorithmic information curation in a controlled environment. Specifically, we look at the text search results for “us elections,” “donald trump,” “joe biden,” “bernie sanders” queries on Google, Baidu, Bing, DuckDuckGo, Yahoo, and Yandex, during the 2020 primaries. Our findings indicate substantial differences in the search results between search engines and multiple discrepancies within the results generated for different agents using the same search engine. It highlights that whether users see certain information is decided by chance due to the inherent randomization of search results. We also find that some search engines prioritize different categories of information sources with respect to specific candidates. These observations demonstrate that algorithmic curation of political information can create information inequalities between the search engine users even under nonpersonalized conditions. Such inequalities are particularly troubling considering that search results are highly trusted by the public and can shift the opinions of undecided voters as demonstrated by previous research.

DOI: https://doi.org/10.1177/08944393211006863

Originally published at:
Urman, Aleksandra; Makhortykh, Mykola; Ulloa, Roberto (2021). The Matter of Chance: Auditing Web Search Results Related to the 2020 U.S. Presidential Primary Elections Across Six Search Engines. Social Science Computer Review:Epub ahead of print.
DOI: https://doi.org/10.1177/08944393211006863

Original Manuscript

Social Science Computer Review
1-17
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/08944393211006863
journals.sagepub.com/home/ssc
SAGE

The Matter of Chance: Auditing Web Search Results Related to the 2020 U.S. Presidential Primary Elections Across Six Search Engines

Aleksandra Urman[1,2], Mykola Makhortykh[1], and Roberto Ulloa[3]

Abstract
We examine how six search engines filter and rank information in relation to the queries on the U.S. 2020 presidential primary elections under the default—that is nonpersonalized—conditions. For that, we utilize an algorithmic auditing methodology that uses virtual agents to conduct large-scale analysis of algorithmic information curation in a controlled environment. Specifically, we look at the text search results for "us elections," "donald trump," "joe biden," "bernie sanders" queries on Google, Baidu, Bing, DuckDuckGo, Yahoo, and Yandex, during the 2020 primaries. Our findings indicate substantial differences in the search results between search engines and multiple discrepancies within the results generated for different agents using the same search engine. It highlights that whether users see certain information is decided by chance due to the inherent randomization of search results. We also find that some search engines prioritize different categories of information sources with respect to specific candidates. These observations demonstrate that algorithmic curation of political information can create information inequalities between the search engine users even under nonpersonalized conditions. Such inequalities are particularly troubling considering that search results are highly trusted by the public and can shift the opinions of undecided voters as demonstrated by previous research.

Keywords
search engines, web search elections, U.S. elections, algorithmic auditing

Search engines play a crucial role in today's high-choice media environment (Van Aelst et al., 2017). The rapid growth of the volume of available information dictates the need for large-scale filtering and ranking of information sources. Without automated mechanisms for prioritizing

[1] University of Bern, Switzerland
[2] University of Zurich, Switzerland
[3] GESIS, Cologne, Germany

Corresponding Author:
Aleksandra Urman, University of Bern, Fabrikstrasse 8, Bern 3012, Switzerland.
Email: urman@ifi.uzh.ch

reliable and relevant sources, users would be overwhelmed by the abundance of information. Thus, search engines turn into major information gatekeepers (Laidlaw, 2010; Schulz et al., 2005; Wallace, 2018) with their ranking and filtering mechanisms directing the information that is received by the users. By doing so, these mechanisms of algorithmic curation can influence users' beliefs and decisions and, in some cases, also reinforce their existing biases (Noble, 2018; White & Horvitz, 2015).

The leverage of search engine algorithms on information filtering and ranking is of particular concern in the context of political campaigns. Previous research has shown that merely the way the results are ranked can shift the voting preferences of undecided voters by 20% or more with the potential shift being as high as 80% for some demographic groups (Epstein & Robertson, 2015). While the effect size depends on the share and demographics of undecided voters as well as the level of Internet penetration in the country, it highlights the potential influence that search engines have over election outcomes. Hence, search engine corporations are increasingly called to take responsibility for preventing biases in search results and providing citizens with consistent and reliable information (Elgesem, 2008; Hinman, 2008).

The present study builds on the previous research on diversity, biases, and discrepancies in political web search results (i.e., Diakopoulos et al., 2018; Haim et al., 2018; Puschmann, 2019; Robertson, Jiang, et al., 2018; Steiner et al., 2020). Using an algorithmic auditing methodology (Mittelstadt, 2016) that builds on that proposed by Haim and colleagues (Haim et al., 2017; Haim, 2020), we investigate the curation of information on search engines under the default—that is, nonpersonalized—fileting and ranking conditions. Our methodological approach helps to further advance the field of algorithmic impact auditing as it allows tracing the effects of randomization of search results at scale. In-built randomization is a factor that can lead to major differences in search output (Makhortykh et al., 2020), yet it has been largely overlooked in the previous search engine auditing studies (i.e., Haim et al., 2018; Hannak et al., 2013; Puschmann, 2019; Robertson, Jiang, et al., 2018; Robertson, Lazer, et al., 2018; Trielli & Diakopoulos, 2019). The details on the methodology are outlined in the relevant section.

We contribute to the existing research on the representation of political topics in search results by examining how search engines distribute information about the candidates for the U.S. 2020 presidential elections during the primaries using a set of the following search queries: "us elections," "joe biden," "donald trump," and "bernie sanders." We compare our observations to those of the previous studies conducted in the context of the 2016 U.S. presidential elections (Kulshrestha et al., 2019; Trielli & Diakopoulos, 2019) and discuss potential implications of our observations.

In the current study, we explore the differences in the search results provided by six major search engines (Google, Yahoo!, Bing, DuckDuckGo, Baidu, and Yandex) for the queries related to the 2020 U.S. presidential elections during the early stage of the election campaign. Specifically, we scrutinize the results mentioning the U.S. elections, incumbent president Donald Trump, Bernie Sanders, and Joe Biden. We included queries on both, Sanders and Biden, as they were the two major contenders for the Democratic presidential nomination at the time when the experiment was conducted—1 week before the so-called Super Tuesday (the day when the biggest number of the U.S. states hold primaries).

In the context of politics, search engines are of utmost importance since, at least in the Western democracies, they are the first place where people look for political information (Dutton et al., 2017). At the same time, the general public tends to highly trust web search output (Pan et al., 2007; Schultheiß et al., 2018) despite the fact that several studies have shown that search results can exhibit racial and gender biases (Kay et al., 2015; Noble, 2018; Singh et al., 2020). Because the way search results are ranked can shift political opinions of undecided voters (Epstein & Robertson, 2015), it is important to investigate how political information is curated by search engines. To do so, we examine the differences in search outputs for the queries related to the 2020 U.S. presidential elections and aim to answer the following research questions:

> **Research Question 1:** How large are the differences in the results provided by various search engines under the default (i.e., nonpersonalized) selection and prioritization of information related to the 2020 U.S. presidential elections?
>
> **Research Question 2:** Are there differences in the results provided by the same search engine to identical users under the default conditions?
>
> **Research Question 3:** Do the levels of discrepancies vary between searches about different political actors in relation to the 2020 U.S. presidential elections?
>
> **Research Question 4:** Are there differences in the types of information sources prioritized by search engines for queries about different political candidates?

Related Work: Algorithmic Impact Auditing and Political Search Results

Algorithmic systems are essential elements of digital platform infrastructure. The need to assess their performance led to the formation of the set of methods collectively known as algorithmic auditing that is "a process of investigating the functionality and impact of decision-making algorithms" (Mittelstadt, 2016, P. 4994). While functionality auditing examines how algorithms arrive at certain decisions and outputs, impact auditing aims to find out which algorithmic outputs are prevalent and infer whether these outputs are biased in some way (Kroll et al., 2017; Sandvig et al., 2014). Algorithmic impact auditing of search engines is of paramount importance because they influence citizens' political information-seeking behavior by filtering and ranking politics-related information (Trevisan et al., 2018).

Since differences in search output can shift the opinions of undecided voters (Epstein & Robertson, 2015), biases in political search results can affect election outcomes and the general political landscape. In recent years, a number of studies that used algorithmic auditing in the context of political searches was conducted. Methodologically, such studies fall into three categories: those that rely on manually generated data (i.e., the ones collected from individual users or generated by the researchers themselves by manually querying search engines), those that rely on virtual agents simulating users' browsing behavior to generate and collect the data, and those that combine these two approaches.

The studies, which use manually generated data, primarily look on the effects of search personalization in the context of information behavior. Two of these studies investigated political filter bubbles on Google using crowd-sourced search results and found no evidence of bubbles' presence (Courtois et al., 2018; Robertson, Jiang, et al., 2018). Still, using a similar methodology another study found significant differences in personalized search results related to the U.S. presidential elections of 2016 (Robertson, Lazer, et al., 2018). Finally, a study that relied on the manual collection of the data by the researchers has assessed the diversity of search results in response to politically salient queries in the German context (Steiner et al., 2020). According to the findings, a certain degree of diversity is present even for the top results (depending on the query), but diversity generally increases for the long tail of search results.

The growing number of studies uses virtual agents to audit algorithmic content curation by search engines. One of the earliest studies (Feuz et al., 2011) on personalization of search results simulated browsing behavior of three different information-seeking personas on Google. The researchers found that results are affected by personalization and the effect increases overtime; the longer the virtual personas used the search engine, the more different were the results. Another study examined Google search results in the context of 2017 federal elections in Germany (Unkel & Haim, 2019). Specifically, the study simulated browsing activity of five information-oriented personas and showed the prevalence of general news websites and resources controlled by political parties in the results. Another study used a single virtual agent to query Google for a set of political queries that are germane to different ideological groups (Democrats vs. Republicans as the study was conducted in the U.S. context) and assess whether search engine results can be biased by the searcher's political orientation (Trielli & Diakopoulos, 2019). Another study that used a single virtual agent (Kulshrestha et al., 2019) investigated bias in Google's search results during the 2016 U.S. presidential election primaries. The authors found that Google's results tend to be biased in the direction of a specific candidate's political leaning (i.e., those related to "Donald Trump" exhibit a slight conservative bias and those related to "Hillary Clinton"—a slight liberal bias).

Besides studies relying exclusively on manual or agent-based data collection, there is some research combining the two approaches. The first study that combined virtual agent-based testing with crowd-sourced data for search engine auditing examined effects of different factors on search personalization (Hannak et al., 2013). The authors used virtual agents to generate a set of nonpersonalized results and compared them with the personalized results obtained from actual users. The study found that personalization significantly affects search results on both Bing and Google that were examined. In another study (Puschmann, 2019), the author asked the users to install a plug-in that queried Google for political searches at regular time intervals thus mimicking users' behavior and isolating potential bias related to the differences in the time when the searches were performed. The analysis revealed discrepancies in the ways different German parties were represented in Google Search and Google News in the run-up to the 2017 German federal elections.

The mentioned studies, with the exception of the one by Hannak and colleagues (2013) and the one by Steiner and colleagues (2020), have focused on one search engine—Google—and did not compare potential differences in algorithmic information curation between the search engines. This is understandable because Google currently dominates the global search market with around 90% of the market share (Statcounter, 2020) and is the engine that is the most commonly used by the majority of Western users. However, other search engines should not be overlooked because they are still used by millions of users across the globe and in some cases dominate regional search markets (i.e., Baidu is the leader on the Chinese market, and Yandex has around 50% of the market share in Russia; Statcounter, 2020). Furthermore, including other engines in the analysis allows testing whether some of them exhibit more biases than others and check whether the choice of a search engine itself affects the quality of information a user is exposed. Therefore, the first contribution to the existing scholarship that we aim to do is to compare politics-related results obtained through the six most popular search engines worldwide (Statcounter, 2020).

Apart from the lack of comparative research on search engine performance, aforementioned studies tend to look at the effects of personalization on search results and potential biases stemming from different variables (i.e., ideological bias of the searchers). None of them, however, has explored for the inherent randomization and volatility of search results. As search engines constantly and continuously update the results, the results inevitably change all the time. Hannak and colleagues (2013) have acknowledged the existence of this effect and attempted to control for it in their study by adding a control virtual agent. However, it is unclear if adding a single control agent is enough—that is, if noise affects all identical results equally. In addition, the scope of the differences in search results due to continuous search updates and inherent randomization has not been extensively examined to date. The only evidence on the level influence of these effects on search outputs comes from a commercial tool that tracks the volatility of search results for the same user throughout the day (SEMrush, n.d.) and from a study that found significant differences in the results for a singular "coronavirus" query when executed by several identical users at the exact same time under the same default filtering and ranking conditions (Makhortykh et al., 2020). With the present study, we aim to partially address this gap by examining the effects of the continuous search updates on the results through a systematic comparison of the results across several search queries and engines.

Method

Data Collection

Using automated agents to simulate browsing behavior of Internet users, we collected the HTML search results from the six most popular search engines according to Statcounter (2020): Google, Bing, Yahoo, Baidu, Yandex, and DuckDuckGo. Extending the methodology adopted by Haim et al. (2017), we built a cloud-based infrastructure to set up a controlled environment that allowed us to isolate external factors (e.g., time or location) and block the effects of search engine's in-built randomization (Makhortykh et al., 2020). Thus, our methodology addresses a potential limitation of earlier algorithmic auditing studies that did not account for the randomization effects (Kulshrestha et al., 2019; Puschmann, 2019; Steiner et al., 2020; Trielli & Diakopoulos, 2019; Unkel & Haim, 2019). Although one study looked at this effect for a singular query ("coronavirus") in the context of COVID-19 (Makhortykh et al., 2020), none, to date, examined the influence of randomization in the context of political search results, which is a gap we aim to address.

To implement the study, we used a cloud-based infrastructure made of 100 CentOS virtual machines deployed via Amazon Elastic Compute Cloud (EC2) and located in the Frankfurt EC2 region. We chose this particular region outside of the United States because (1) we considered that the usage of any region inside the United States might introduce biases in search results due to geolocation-based personalization as Republicans and Democrats are not evenly distributed across states; (2) we did not have the resources to afford more than one EC2 geographic region to counteract this potential effect (e.g., by selecting one pro-Republican and one pro-Democratic region), further, at the time of the analysis, EC2 had no clusters available in pro-Republican states; and (3) we selected Frankfurt because it serves as a base for many international companies and has a high share of English-speaking population.

All the machines were t3a.medium Amazon EC2 instances based on AMD EPYC 7,000 series processors. Each machine had two CPUs, four gigabyte (GB) RAM, and 20 GB hard drive. Because the machines were generated using the same Centos-based Amazon machine image with the same set of software installed (e.g., same Centos packages and browser versions), they had the same hardware and software specifications. Besides, all machines were located in the same range of Internet protocol (IPs) performed identical searches at the same time. Hence, the searches were conducted in a fully controlled environment that accounted for potential factors that could have led to the discrepancies in search results (e.g., due to personalization). The only difference between the machines related to their unique IP addresses—though they all belonged to the same IP range provided by EC2 and should not have affected the results due to, that is, location-based personalization. We do acknowledge, however, that this is a limitation of the present study, and future research should investigate the potential effects of the said discrepancy.

Each virtual machine hosted two browsers: Firefox and Chrome. In each browser ("agent"), we installed two extensions: a tracker and a bot. The tracker collected metadata (e.g., time stamps) and the full HTML of all pages that were visited within the browser that were sent to an external storage server. The bot emulated user browsing behavior by searching query terms from the predefined list (which included terms "us elections," "joe biden," "donald trump," and "bernie sanders") and navigating through the search results. The queries were selected based on the event that the search was centered on. We opted for generic actor names (e.g., instead of actor names accompanied by descriptions) to retrieve the least biased results about the actors. The focus on the three aforementioned actors is explained by the fact that ahead of the primary elections, Joe Biden and Bernie Sanders were the major contenders for the Democratic nomination, and Donald Trump was then-incumbent President running for the reelection. We added a generic "us elections" term to get a broader overview of search results at the time, which would not be biased toward one of the

Table 1. The Total Number of Agents That Completed the Task per Search Engine and Browser.

Browser	Baidu	Bing	DDG	Google	Yahoo!	Yandex
Firefox	15	16	17	17	15	16/6 (*)
Chrome	16	15	17	16	16	16/13 (*)

Note. The first row displays the name of the search engine (DDG is an abbreviation for DuckDuckGo), and the first column shows the name of the browser. (*) = We obtained fewer results for Yandex for the "U.S. elections" query because it triggered the bot detection algorithm of Yandex which blocked some of the agents.

candidates. In the present study, we entered identical queries into search engines without accounting for potential differences in the ways search engine algorithms handle multiword queries. We opted for this to achieve maximum consistency between the searches which we deemed necessary as our study is focused on impact auditing. Future studies that focus on functionality auditing of web search algorithms might investigate, however, how multiword queries are handled by different algorithms.

The navigation through the retrieved search results was organized in browser sessions, which consisted of three steps: (1) visiting the main landing page of a search engine, (2) inputting a query from the predefined list into the search text box, and "clicking" on the search button, and (3) navigating through the search results.

Each agent collected at least the top 50 results by visiting multiple result pages or by scrolling down the page (in case of infinite scrolling configuration of the search result page, such as in the case of DuckDuckGo). Immediately after each search session, the browsers were cleaned to prevent previous searches from affecting the following sessions. The bot removed both the data accessed by the browser (i.e., browsing history and cache) and the browser data that can be retrieved by the search engines' algorithms (i.e., local storage, session storage, and cookies). At the time of the data collection, none of the search engines was forcing the users to accept or reject their cookie policies. Hence, none of the agents accepted engine-specific policies.

Regardless of the search engine, each search session lasted less than 3 min. Each subsequent session started 7 min after the beginning of the previous one to guarantee at least a 4-min gap between sessions. Therefore, the agents were always synchronized at the beginning of all sessions to isolate the potential effect of time on search results.

The 200 agents were deployed on February 26, 2020, 1 day after the Democratic debate and almost a week before Super Tuesday, when 14 states hold democratic primary elections. The search engines (Baidu, Bing, DuckDuckGo, Google, Yandex, and Yahoo) selected for this study were equally distributed among the agents, so that 32 of 33 agents (15 of 16 from each browser group) were assigned to each search engine. During our collection, the expected amount of agents was slightly decreased because of the issues: (1) bot detection in Yandex via occasionally appearing captchas, and (2) a few browser crashes due to the limited volume of RAM available on the machines. The total number of agents providing data for each browser–engine combination is provided below (Table 1).

Data Analysis

After collecting the data, we used BeautifulSoup (Python; Richardson, 2020.) and rvest (R; Wickham & RStudio, 2019) packages to extract search results from the HTML for each query and filter out the URLs not related to the search results (e.g., ads). The latter decision is explained by our implicit interest in the default mechanisms for search filtering and ranking, not in the ads displayed by the engines. Then, for each query, we compared the URLs of the search results obtained by each possible pair of agents. We used two similarity metrics—Jaccard Index (JI) and Rank Biased Overlap (RBO).

JI measures the overlap between two sets of results and shows the size of the intersection between the sets over the union. JI has been used to measure similarities in search results by previous studies personalization of web search (Hannak et al., 2013; Kliman-Silver et al., 2015; Puschmann, 2019). The values of the JI vary from 0 to 1, with 1 indicating that the compared sets are identical, and 0 that they are completely different.

Although JI is valuable for assessing the similarity between two sets of results, it does not take into account their ranking. Yet, the latter feature is especially relevant for the present study due to the proven effect of search ranking on voting preferences (Epstein & Robertson, 2015). For this reason, we also used RBO metric that accounts for the order in which results are presented and is frequently utilized in the studies on search engines (Cardoso & Magalhães, 2011; Robertson, Jiang, et al., 2018; Robertson, Lazer, et al., 2018). Specifically, RBO takes into consideration three important characteristics of web search: incompleteness (there are too many search results so it is not possible to scrape all of them), indefiniteness (chosen result range is arbitrary), and top-weightedness (variation between the top results is more important than the one between the lower ones) of the results (Webber et al., 2010). The formula for RBO is as follows:

$$\mathrm{RBO}(S, T, p) = (1 - p)\sum_{d=1}^{\infty} p^{d-1} \cdot A_d,$$

where S and T are two infinite rankings, d is the depth to which their agreement is computed, A is the level of agreement (which is equal to a Jaccard similarity of top d results), and the persistence parameter p determines the importance of top results: The lower the value of p, the more weight is assigned to the top results.

For each of the four queries, we calculated JI for the overall set of results and for top 10 results, and RBO ($p = .95$ and $p = .8$) for all the result pairs. Setting p to .95 allowed us to conduct a more systemic analysis of the differences between result pairs, whereas $p = .8$ enabled us to put more emphasis on the first few results (Webber et al., 2010).

After calculating JI and RBO, we aggregated the data for each search engine examined in the study. To make sure that the agents' browsers did not cause the discrepancies between the search results, we aggregated data separately for Chrome and for Firefox. This also allowed us to check whether search results differ between the two browsers for otherwise identical agents. Afterward, we calculated the mean values for JI and RBO between the sets of agents with different combinations of search engines and browsers they were produced by.

We produced a linear mixed effect model using the lme4 and lmerTest R packages (Bates et al., 2020; Kuznetsova et al., 2020) to fit the data and calculate the statistical significance for our main independent variables: browser and search engine. The model allows us to control for (1) the effects of multiple comparisons of each agent, that is, the search results of an agent are used several times, one per comparison against the search results of the other agents, and (2) the effects of the machine combination, that is, since each machine contains two agents (one per browser), we need to control for pairing the same machines several times.

To assess whether there are qualitative differences in the types of content prioritized by the search engines in response to different queries, we have first aggregated data about the domains that appeared most frequently in the top 20 results for each search engine–query combination. Then, we have manually coded the results based on the following categories:

- think tank/academic websites (i.e., academic articles/think tank reports),
- social media sites (i.e., Facebook, Twitter),
- reference work (i.e., dictionaries, encyclopedic notes, Wikipedia),
- news aggregators (i.e., *Google News*),
- legacy media (i.e., *New York Times*),
- infotainment (i.e., soft news websites such as Buzzfeed),
- government (i.e., White House website),
- fact-checking websites (i.e., PolitiFact),
- commerce (i.e., online shops),
- campaign (i.e., official candidate–affiliated campaign websites),
- alternative media (i.e., digital-born partisan outlets such as Conservapedia), and
- not available (i.e., the link points to a site/page that is no longer available at the time of the analysis).

The coding was performed by one of the authors and then thoroughly checked by the two other authors to ensure agreement between them. All the disagreements arising from the checks were resolved via consensus-coding in a series of group discussions.

To answer our research questions, we first looked at the differences in the results obtained for the "us elections" query via different search engines (e.g., Google vs. Yahoo; Research Question 1), then by different agents using the same browser (e.g., when both agents used Google; Research Question 2). Then, we repeated these two steps for the queries related to specific politicians (i.e., "bernie sanders," "donald trump," "joe biden") and checked the discrepancies between the results obtained for each query (Research Question 3). Finally, we have qualitatively examined and categorized the types of domains in top 20 results for different search engines and analyzed the differences in the types of sources prioritized by each engine for specific politicians (Research Question 4).

Results

Differences in Search Results on "us elections" Query

In response to Research Question 1, we find significant discrepancies between filtering and ranking mechanisms utilized by different search engines for the "us elections" query on both, Chrome and Firefox browsers (see Supplemental Material for the complete statistical summary). We observe that 115 of 120 (95.8%) similarity values between different search engines are lower than .35, including all the JI values for the top 10 search results (Figure 1). The ranking of the results is also highly volatile that means even in the nonpersonalized setting, users of the same search engine are unlikely to see the same results. While some discrepancies in the results provided by different search engines are expected, given that they utilize different algorithms to filter and rank results, the magnitude of discrepancies suggests that users of these platforms get fundamentally different sets of information.

The most similar results are provided by DuckDuckGo and Yahoo search engines. However, in this case, the two engines share just under half of all the results ($n \sim 50$) results (measured by JI overall) and around a third of the top 10 results (JI for top 10). The similarities are even lower when the ranking is taken into account (RBO) with the ordering of top results (RBO, $p = .8$) in most cases being more volatile than the overall ordering of the results (RBO, $p = .95$). This finding echoes that of Steiner and colleagues (2020) who found that the differences in search results are higher for the lower positioned results. Still, for the second most similar pair, Bing and Yandex, the top results are more similar in terms of both, order and composition as indicated by higher top 10 JI and RBO with $p = .8$. Hence, the findings regarding the volatility of search results with different rankings are contextual.

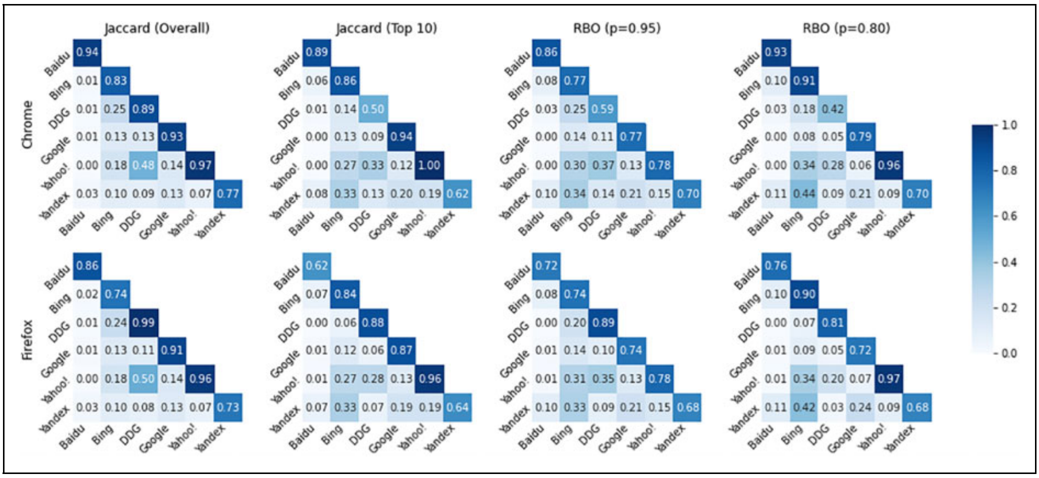

Figure 1. Cross-engine and cross-browser similarities in search results for the "us elections" search query. The columns show the different similarity measures used, and the rows show the results for Chrome (top) and Firefox (bottom) browsers.

The discrepancies in the results between different search engines can have important consequences for the public sphere because the users of different engines get different (political) information. However, most Western markets, including the United States, are currently dominated by Google. The Google's share of the U.S. search market is estimated to be just under 90%, slightly lower than Google's market share worldwide (Statcounter, 2020). This means that even if the results provided by Google are very different from those on other platforms, it does not affect about 90% of the U.S. public. However, what does affect the Western public is the high degree of randomization that creates discrepancies in the information curation even in the nonpersonalized context.

Concerning Research Question 2, we observed variations in search results within the same search engine for both Chrome and Firefox browsers (diagonal values in the plots in Figure 1). The only search engine in our sample did not randomize the selection of the top 10 results for the "us elections" query was Yahoo accessed from Chrome (but not from Firefox), and even in this case, the ordering of the results showed some variation between the agents. Since such variation happens under the nonpersonalized conditions, it is seemingly random and, most likely, attributed to the fact that search engine algorithms constantly adapt their output to provide the results viewed as the most relevant to the users at the given time. This constant output adaptation means that users of the same search engine are likely to receive different results even when they conduct searches at the same time and no personalization is involved. Even though the within-engine discrepancies are not as high as cross-engine ones, their effect on the public opinion is still important because the ranking of the search results can shift voters' opinion (Epstein & Robertson, 2015).

In terms of the browser differences, we only found significant differences between Firefox and Chrome for DuckDuckGo (the statistical table is reported in Supplemental Material, Table A1). In this case, the search results obtained in Firefox are more consistent than those obtained in Chrome for all our response variables. One potential explanation of this is that DuckDuckGo's search algorithm takes a user's browser into account when making curation decisions. If true, this is problematic since browser is a semantically nonmeaningful signal and its influence on search results can increase information inequalities between users of different browsers.

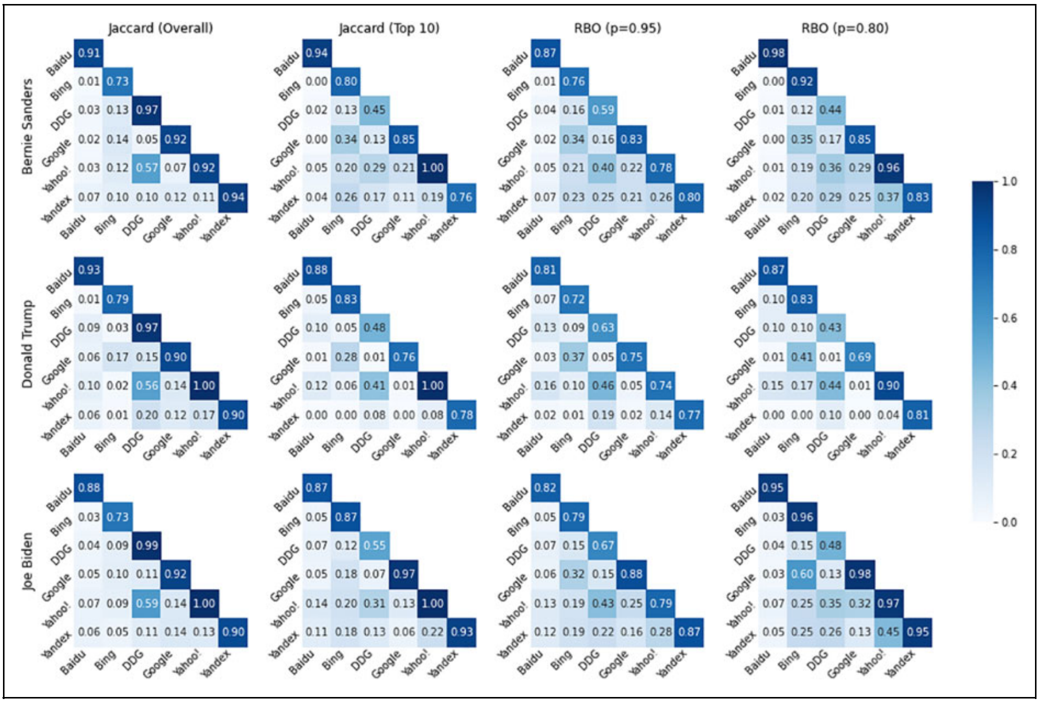

Figure 2. Similarities in search results across engines by relevant political candidate query, Chrome browser.

Discrepancies in the Stability of Search Results for Different Political Candidates

To answer Research Question 3, we compared the differences in the results for different political candidate queries. Similarly to the search results for the “us elections” query, there are large discrepancies for the queries related to specific political candidates (see Figure 2 for Chrome and Figure 3 for Firefox). On average, cross-engine dissimilarities for all the political candidates–related queries are slightly higher than those for the “us elections” query.

We found significant differences between the browsers depending on the query (statistically significant values are reported in Table 2; for the complete statistical tables see Supplemental Material, see Sections B.1, C.1, D.1). The results are in line with the findings for the “us elections” query: DuckDuckGo shuffles the results for Chrome users more than it does for Firefox users. For the “bernie sanders” query, we also observed some browser differences for Bing and Yahoo.

Looking only at the results for the same-engine comparisons and controlling for different browsers, we found statistically significant differences in terms of consistency of search results between the three queries (see Supplemental Material, Section E). These differences depend on the search engine that is being used, but overall “joe biden”- and “bernie sanders”-related results are less volatile than “donald trump” ones in terms of JI (top 10), RBO ($p = .8$) and RBO ($p = .95$; see Supplemental Material, Section F).

Prioritization of Source Types

In order to infer qualitative differences between the results for queries on different political candidates and, thus, answer Research Question 4, we examined the top 20 results most frequently obtained

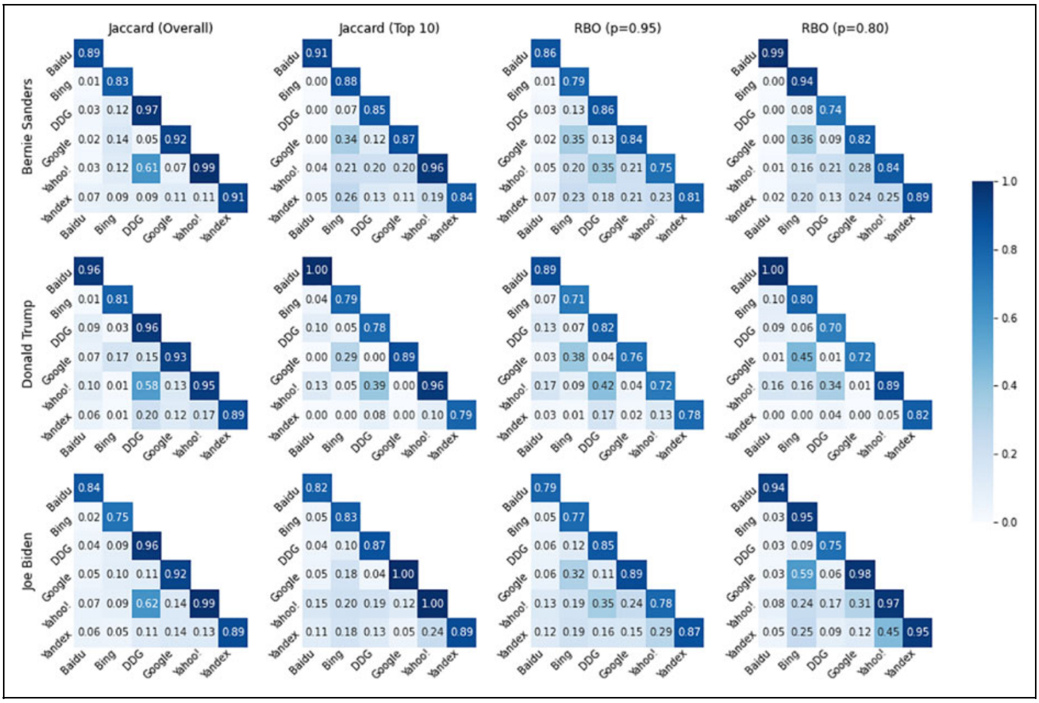

Figure 3. Similarities in search results across engines by relevant political candidate query, Firefox browser.

Table 2. *p* Values for Statistically Significant Effects of Browser in Politician Queries.

Search query	Engine	JI (Overall)	JI (Top 10)	RBO ($p = .8$)	RBO ($p = .95$)
bernie sanders	Bing	.016	—	—	—
	DDG	—	<.0001	.0296	.0001
	Yahoo	—	—	.0303	
joe biden	DDG		.0035	.0002	.0036
donald trump	—	—	—	—	—

Note. The first column shows the query term used. The second column refers to the search engine. JI = Jaccard Index; RBO = Rank Biased Overlap.

through each engine for each of the three candidate-related queries (see Figures 4–6). We found that search engines, in general, prioritize different categories of search results, and in some cases (i.e., Baidu, Yahoo, and Yandex), there are large discrepancies between different candidate queries.

On Google, the prioritization of results is consistent for all three candidates with legacy media dominating search outputs. This is similar to what was observed in a study on Google search results in relation to the 2017 German federal elections (Unkel & Haim, 2019). Overall, legacy media results were more prevalent in the outputs of Google and Bing than those of the other search engines. The only major difference we observed on Google for different queries is that the results for “bernie sanders” did not contain a link to Sanders’ campaign website, unlike those for “donald trump” and “joe biden.” In addition, we found that the candidates-controlled campaign websites were less prevalent in Google results during the 2020 primaries than during the 2016 primaries when almost a quarter of top 10 results were made of candidate-affiliated websites (Kulshrestha et al., 2019). Still,

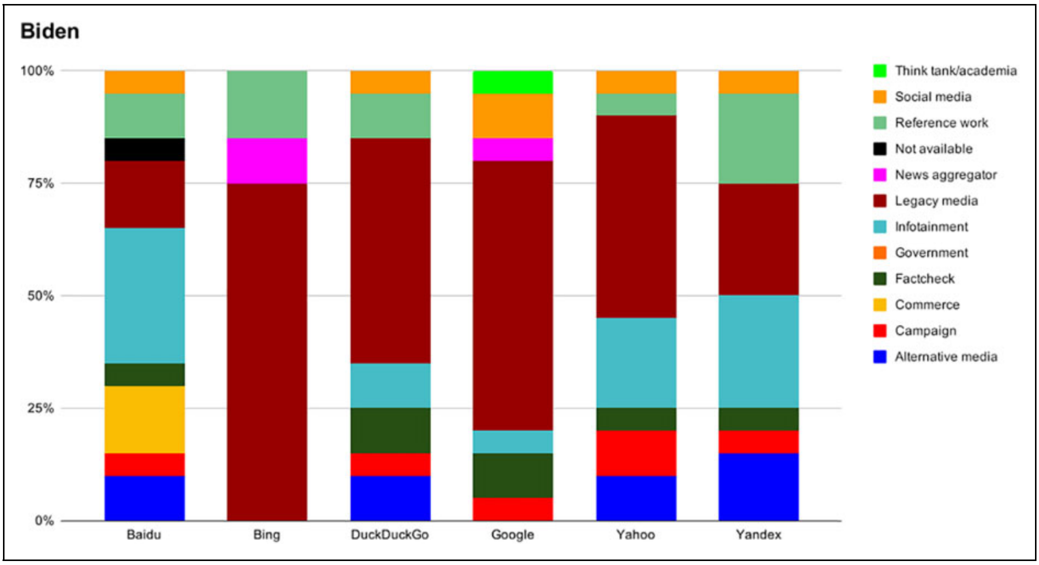

Figure 4. Information sources referenced in top 20 search results for “joe biden.”

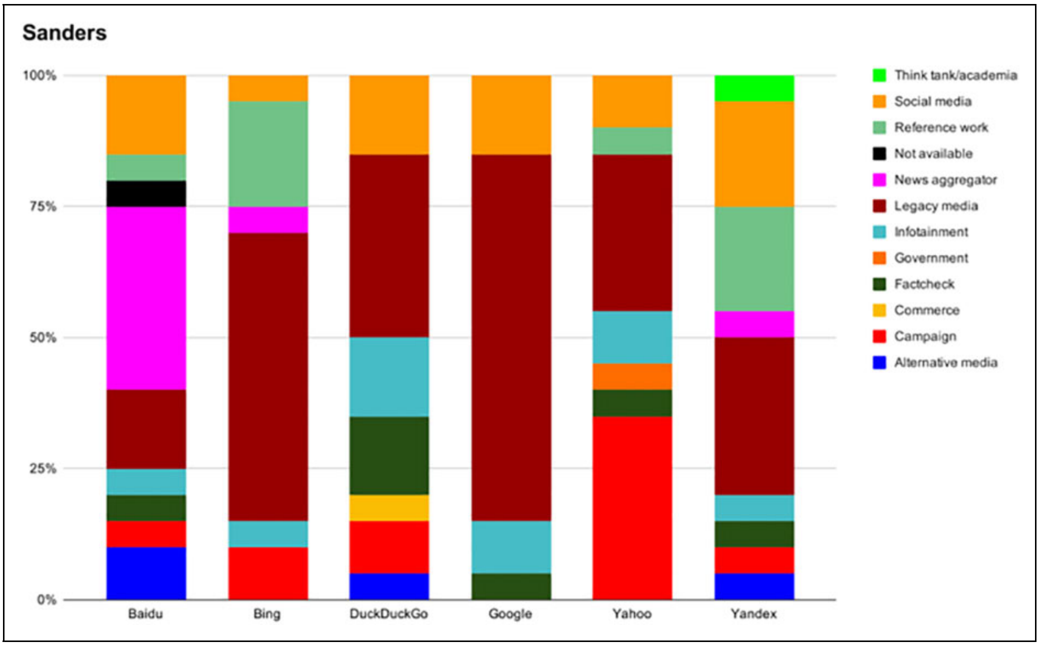

Figure 5. Information sources referenced in top 20 search results for “bernie sanders.”

further longitudinal studies are necessary to properly verify this claim. As we conducted a snapshot experiment, we cannot state whether how persistent and systematic this observation is.

More pronounced differences across candidates were observed on Bing and Yahoo, the second and third most popular engines in the United States with 6.55% and 3.65% of the search market, respectively (Statcounter, 2020). In terms of potential biases, Yahoo displayed a high ratio of pro-Sanders results with around 40% of the top 20 results linking to outlets related to his campaign,

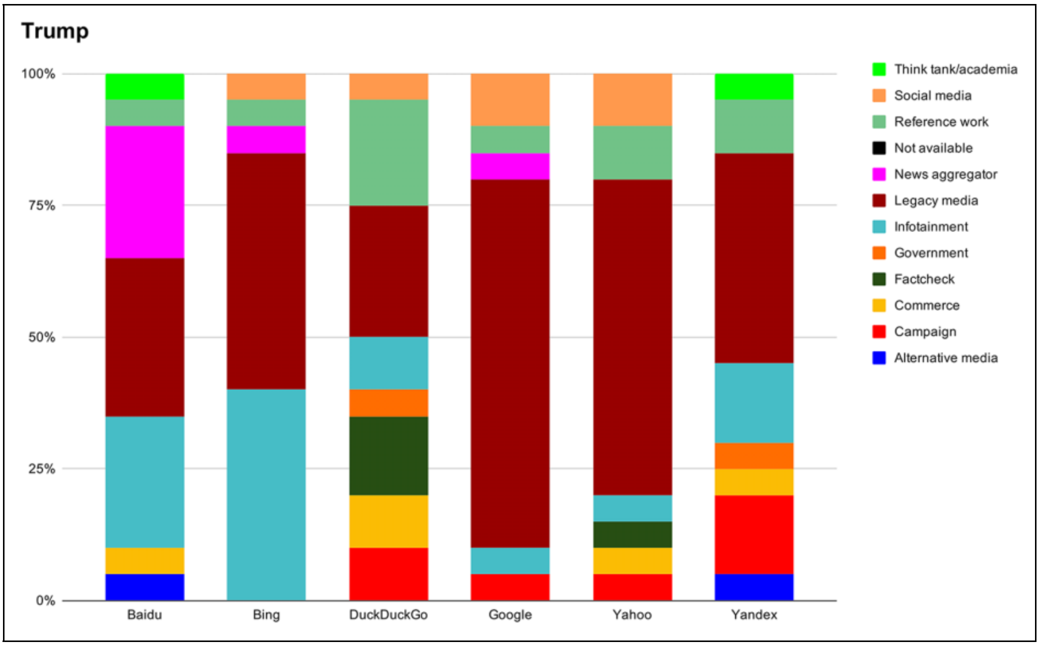

Figure 6. Information sources referenced in top 20 search results for "donald trump."

while for Biden and Trump the ratio was 10% and 5%, respectively. However, as we conducted a snapshot experiment, we cannot say how stable the observed effect is overtime.

Discussion

Our findings highlight two major issues related to the ways search engine filter and rank political information. The first issue is a number of differences in the search outputs produced by algorithmic curation mechanisms of different search engines (Research Question 1). While some variation in the selection of information sources is expected as the engines clearly employ different algorithms to retrieve and rank the results, our study indicates that even under the nonpersonalized conditions, search results show varying degrees of volatility and prioritize different types of sources depending on the query. These discrepancies, even if they are to be expected due to the differences in algorithms, can lead to information inequalities between the individuals who use different search engines, in particular, as some of the engines seem to prioritize sources which are more supportive (e.g., Yahoo for Sanders) or critical (e.g., Yandex for Biden) of specific candidates. We suggest that this observation warrants further studies into how usage of different search engines affects the populations from the social science perspective.

While the effects of these inequalities might be somehow limited in the United States where over 90% of the public are using Google as their default engine (Statcounter, 2020), in the countries where the search market is not dominated by a single engine, the cross-engine discrepancies can have a larger effect on the public sphere. Such contexts include, for instance, East Asian states such as China, Japan, and South Korea, as well as post-Soviet countries such as Russia, Kazakhstan, and Belarus, where local search corporations serve as major competitors for Western tech giants (Statcounter, 2020).

The second troubling issue is the volatility of search results within the same search engine (RQ2). The randomization of search results is not necessarily a negative phenomenon, because it allows the engines to present the most relevant information by updating the ranking of sources and can potentially diversify users' information diets (Helberger et al., 2018). On the other hand, such volatility makes search outputs less predictable and might lead to information inequalities between the users of the same engine by randomizing their access to information. We also find that the volatility of search results differs across different candidate queries (RQ3), with the results related to the two Democratic candidates being more stable than those for "donald trump" query.

In addition, our analysis has revealed qualitative differences in the composition of top results across the three political candidate queries (RQ4). For instance, we observed that Yahoo contained a much higher share of campaign websites for the "bernie sanders" query compared to other engines and queries. Such discrepancy might indicate a potential pro-Sanders bias in the output, but without a longitudinal study, it is not possible to verify how systematic this bias is. Further longitudinal research utilizing similar methodology is required to enhance our understanding of how resilient the observations coming from the current study are as we conducted a snapshot experiment and cannot state whether our observations indicate a presence of a systematic bias. Still, the observed differences in search results across political queries, engines and browsers are already troubling, because the ranking of political search results can affect voters' decisions (Epstein & Robertson, 2015).

In contrast to earlier research focusing on the effects of personalization on political information dissemination via search engines (i.e. Hannak et al., 2013; Puschmann, 2019; Unkel & Haim, 2019), our study highlights the need for taking into account search results' volatility that is present on all search engines we audited. Whereas personalization does not significantly alter election-related search results, at least on Google in the context of the German Federal elections (Unkel & Haim, 2019), our findings show that built-in randomization can strongly affect the composition and the ranking of results. It prompts the need to go beyond the current scholarship's focus on search personalization and its influence on promotion of specific biases (e.g., the ones related to gender and race; Noble, 2018) and discuss to what degree inherent volatility of the results can create informational inequalities which make users receive different information under identical conditions. Similar to the search queries related to the emergencies like the COVID-19 pandemic (Makhortykh et al., 2020), in the case of political queries such randomization, can result in some part of the population being less informed or even misinformed about important societal developments. Whether the users get to see certain information or not becomes, thus, a matter of chance that is in stark contradiction with the public's general perception of search results as accurate and trustworthy ("2020 Edelman Trust Barometer," n.d.; Pan et al., 2007) as well as the framing of the search process as unbiased and scientific by the search companies (Sweeney, 2013).

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This paper has been written within the project "Reciprocal relations between populist radical-right attitudes and political information behaviour: A longitudinal study of attitude development in high-choice information environments" lead by Silke Adam (U of Bern) and Michaela Maier (U of Koblenz-Landau) and sponsored by the SNF (100001CL_182630/1) and DFG (MA 2244/9-1).

Supplemental Material

The supplemental material is available in the online version of the article.

References

2020 Edelman Trust Barometer. (n.d.). *Edelman*. Retrieved October 7, 2020, from https://www.edelman.com/trustbarometer

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., & Fox, J. (2020). *Lme4: Linear mixed-effects models using "eigen" and S4 (1. 1-23)* [Computer software]. https://cran.r-project.org/package=lme4

Cardoso, B., & Magalhães, J. (2011). Google, Bing and a new perspective on ranking similarity. In *Proceedings of the 20th ACM International Conference on information and knowledge management* (pp. 1933–1936). https://doi.org/10.1145/2063576.2063858

Courtois, C., Slechten, L., & Coenen, L. (2018). Challenging Google search filter bubbles in social and political information: Disconforming evidence from a digital methods case study. *Telematics and Informatics*, *35*(7), 2006–2015. https://doi.org/10.1016/j.tele.2018.07.004

Diakopoulos, N., Trielli, D., Stark, J., & Mussenden, S. (2018). I vote for—How search informs our choice of candidate. In D. Tambini & M. Moore (Eds.), *Digital dominance: The power of Google, Amazon, Facebook, and Apple* (p. 22). Oxford University Press.

Dutton, W. H., Reisdorf, B. C., Dubois, E., & Blank, G. (2017). *Search and politics: The uses and impacts of search in Britain, France, Germany, Italy, Poland, Spain, and the United States* [Quello Center Working Paper No. 2944191]. https://ora.ox.ac.uk/objects/uuid:2cec8e9b-cce1-4339-9916-84715a62066c

Elgesem, D. (2008). Search engines and the public use of reason. *Ethics and Information Technology*, *10*(4), 233–242. https://doi.org/10.1007/s10676-008-9177-3

Epstein, R., & Robertson, R. E. (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), E4512–E4521. https://doi.org/10.1073/pnas.1419828112

Feuz, M., Fuller, M., & Stalder, F. (2011). Personal web searching in the age of semantic capitalism: Diagnosing the mechanisms of personalisation. *First Monday*. https://doi.org/10.5210/fm.v16i2.3344

Haim, M. (2020). Agent-based testing: An automated approach toward artificial reactions to human behavior. *Journalism Studies*. https://www.tandfonline.com/doi/pdf/10.1080/1461670X.2019.1702892

Haim, M., Arendt, F., & Scherr, S. (2017). Abyss or shelter? On the relevance of web search engines' search results when people Google for suicide. *Health Communication*, *32*(2), 253–258. https://doi.org/10.1080/10410236.2015.1113484

Haim, M., Graefe, A., & Brosius, S. (2018). Burst of the filter bubble? *Digital Journalism*, *6*(3), 330–343. https://doi.org/10.1080/21670811.2017.1338145

Hannak, A., Sapiezynski, P., Molavi Kakhki, A., Krishnamurthy, B., Lazer, D., Mislove, A., & Wilson, C. (2013). Measuring personalization of web search. *Proceedings of the 22nd International Conference on World Wide Web*, 527–538. https://doi.org/10.1145/2488388.2488435

Helberger, N., Karppinen, K., & D'Acunto, L. (2018). Exposure diversity as a design principle for recommender systems. *Information, Communication & Society*, *21*(2), 191–207. https://doi.org/10.1080/1369118X.2016.1271900

Hinman, L. M. (2008). Searching ethics: The role of search engines in the construction and distribution of knowledge. In A. Spink & M. Zimmer (Eds.), *Web search: Multidisciplinary perspectives* (pp. 67–76). Springer. https://doi.org/10.1007/978-3-540-75829-7_5

Kay, M., Matuszek, C., & Munson, S. A. (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* (pp. 3819–3828). https://doi.org/10.1145/2702123.2702520

Kliman-Silver, C., Hannak, A., Lazer, D., Wilson, C., & Mislove, A. (2015). Location, location, location: The impact of geolocation on web search personalization. In *Proceedings of the 2015 internet measurement conference* (pp. 121–127). https://doi.org/10.1145/2815675.2815714

Kroll, J., Huey, J., Barocas, S., Felten, E., Reidenberg, J., Robinson, D., & Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review*, *165*(3), 633.

Kulshrestha, J., Eslami, M., Messias, J., Zafar, M. B., Ghosh, S., Gummadi, K. P., & Karahalios, K. (2019). Search bias quantification: Investigating political bias in social media and web search. *Information Retrieval Journal*, *22*(1), 188–227. https://doi.org/10.1007/s10791-018-9341-2

Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B., & Jensen, S. P. (2020). *Lmertest: Tests in linear mixed effects models (3.1-2)* [Computer software]. https://cran.r-project.org/package=lmertest

Laidlaw, E. B. (2010). A framework for identifying Internet information gatekeepers. *International Review of Law, Computers & Technology*, *24*(3), 263–276. https://doi.org/10.1080/13600869.2010.522334

Makhortykh, M., Urman, A., & Ulloa, R. (2020). *How search engines disseminate information about COVID-19 and why they should do better*. Harvard Kennedy School Misinformation Review, 1 (COVID-19 and misinformation). https://doi.org/10.37016/mr-2020-017

Mittelstadt, B. (2016). Automation, algorithms, and politics\textbar auditing for transparency in content personalization systems. *International Journal of Communication*, *10*(0), 12.

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.

Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., & Granka, L. (2007). In Google we trust: Users' decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication*, *12*(3), 801–823. https://doi.org/10.1111/j.1083-6101.2007.00351.x

Puschmann, C. (2019). Beyond the bubble: Assessing the diversity of political search results. *Digital Journalism*, *7*(6), 824–843. https://doi.org/10.1080/21670811.2018.1539626

Richardson, L. (2020). beautifulsoup4: Screen-scraping library (4.9.3) [Python]. Retrieved March 26, 2021, from http://www.crummy.com/software/BeautifulSoup/bs4/

Robertson, R. E., Jiang, S., Joseph, K., Friedland, L., Lazer, D., & Wilson, C. (2018). Auditing partisan audience bias within Google search. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–148. https://doi.org/10.1145/3274417

Robertson, R. E., Lazer, D., & Wilson, C. (2018). Auditing the personalization and composition of politically-related search engine results pages. In *Proceedings of the 2018 World Wide Web conference* (pp. 955–965). https://doi.org/10.1145/3178876.3186143

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*, *22*, 4349–4357.

Schultheiß, S., Sünkler, S., & Lewandowski, D. (2018). We still trust in Google, but less than 10 years ago: An eye-tracking study. Information research. *An International Electronic Journal*, *23*(3). https://eric.ed.gov/?id=EJ1196314

Schulz, W., Held, T., & Laudien, A. (2005). Search engines as gatekeepers of public communication: Analysis of the German framework applicable to internet search engines including media law and anti trust law. *German Law Journal*, *6*(10), 1419–1431. https://doi.org/10.1017/S2071832200014401

SEMrush. (n.d.). *SEMrush Sensor – Google's rank and algorithm tracking tool*. Semrush. Retrieved March 26, 2021, from https://www.semrush.com/sensor/

Singh, V. K., Chayko, M., Inamdar, R., & Floegel, D. (2020). Female librarians and male computer programmers? Gender bias in occupational images on digital media platforms. *Journal of the Association for Information Science and Technology*, asi.24335. https://doi.org/10.1002/asi.24335

Statcounter. (2020). *Search engine market share worldwide. Statcounter global stats*. https://gs.statcounter.com/search-engine-market-share

Steiner, M., Magin, M., Stark, B., & Geiß, S. (2020). Seek and you shall find? A content analysis on the diversity of five search engines' results on political queries. *Information, Communication & Society*, *0*(0), 1–25. https://doi.org/10.1080/1369118X.2020.1776367

Sweeney, M. (2013). Not just a pretty (inter)face: A critical analysis of Microsoft's "Ms. Dewey" [University of Illinois at Urbana-Champaign]. http://hdl.handle.net/2142/46617

Trevisan, F., Hoskins, A., Oates, S., & Mahlouly, D. (2018). The Google voter: Search engines and elections in the new media ecology. *Information, Communication & Society*, *21*(1), 111–128. https://doi.org/10.1080/1369118X.2016.1261171

Trielli, D., & Diakopoulos, N. (2019). Search as news curator: The role of Google in shaping attention to news information. In *Proceedings of the 2019 CHI Conference on human factors in computing systems* (pp. 1–15). https://doi.org/10.1145/3290605.3300683

Unkel, J., & Haim, M. (2019). Googling politics: Parties, sources, and issue ownerships on Google in the 2017 German federal election campaign. *Social Science Computer Review*. https://doi.org/10.1177/0894439319881634

Van Aelst, P., Strömbäck, J., Aalberg, T., Esser, F., de Vreese, C., Matthes, J., Hopmann, D., Salgado, S., Hubé, N., Stępińska, A., Papathanassopoulos, S., Berganza, R., Legnante, G., Reinemann, C., Sheafer, T., & Stanyer, J. (2017). Political communication in a high-choice media environment: A challenge for democracy? *Annals of the International Communication Association*, *41*(1), 3–27. https://doi.org/10.1080/23808985.2017.1288551

Wallace, J. (2018). Modelling contemporary gatekeeping. *Digital Journalism*, *6*(3), 274–293. https://doi.org/10.1080/21670811.2017.1343648

Webber, W., Moffat, A., & Zobel, J. (2010). A similarity measure for indefinite rankings. *ACM Transactions on Information Systems*, *28*(4), 1–38. https://doi.org/10.1145/1852102.1852106

White, R. W., & Horvitz, E. (2015). Belief dynamics and biases in web search. *ACM Transactions on Information Systems*, *33*(4), 18:1–18:46. https://doi.org/10.1145/2746229

Wickham, H., & RStudio. (2019). *Rvest: Easily harvest (scrape) web pages*. https://cran.r-project.org/package=rvest

Author Biographies

Aleksandra Urman is a postdoctoral researcher at the Institute of Communication and Media Studies, University of Bern, and Social Computing Group, University of Zurich. Her PhD dissertation defended in May 2020 examines polarization on social media from a comparative perspective. Her research interests include political communication on social media, algorithmic biases, and computational research methods.

Mykola Makhortykh is a postdoctoral researcher at the University of Bern, where he studies information behavior in online environments. Before moving to Bern, he defended his PhD dissertation at the University of Amsterdam on the relationship between digital platforms and war remembrance in Eastern Europe and worked as a postdoctoral researcher in Data Science at the Amsterdam School of Communication Research, where he investigated the effects of algorithmic biases on digital news consumption.

Roberto Ulloa is a postdoctoral researcher at the Computational Social Science Department of GESIS—Leibniz Institute for the Social Sciences. His research interests include the role of institutions in polarization and homogenization of opinion. He participates in different investigation projects related to search engine auditing and biases, digital traces, social media platforms, and web tracking.