




# Search bias quantification: investigating political bias in social media and web search

Juhi Kulshrestha<sup>1,2</sup>  · Motahhare Eslami<sup>3</sup> · Johnnatan Messias<sup>1</sup> · Muhammad Bilal Zafar<sup>1</sup> · Saptarshi Ghosh<sup>4</sup> · Krishna P. Gummadi<sup>1</sup> · Karrie Karahalios<sup>3</sup>

Received: 12 November 2017 / Accepted: 3 August 2018 / Published online: 21 August 2018  
© The Author(s) 2018

## Abstract

Users frequently use search systems on the Web as well as online social media to learn about ongoing events and public opinion on personalities. Prior studies have shown that the top-ranked results returned by these search engines can shape user opinion about the topic (e.g., event or person) being searched. In case of polarizing topics like politics, where multiple competing perspectives exist, the political bias in the top search results can play a significant role in shaping public opinion towards (or away from) certain perspectives. Given the considerable impact that search bias can have on the user, we propose a generalizable search bias quantification framework that not only measures the political bias in ranked list output by the search system but also decouples the bias introduced by the different sources—input data and ranking system. We apply our framework to study the political bias in searches related to 2016 US Presidential primaries in Twitter social media search and find that both input data and ranking system matter in determining the final search output bias seen by the users. And finally, we use the framework to compare the relative bias for two popular search systems—Twitter social media search and Google web search—for queries related to politicians and political events. We end by discussing some potential solutions to signal the bias in the search results to make the users more aware of them.

**Keywords** Search bias · Search bias quantification · Sources of search bias · Social media search · Web search · Political bias inference

---

This work is an extended version of the paper: Kulshrestha et al., Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media, ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17), ACM, New York, NY, USA, 417–432. <https://doi.org/10.1145/2998181.2998321>.

---

✉ Juhi Kulshrestha  
juhi@mpi-sws.org

<sup>1</sup> Max Planck Institute for Software Systems (MPI-SWS), Saarbrücken, Germany

<sup>2</sup> GESIS - Leibniz Institute for the Social Sciences, Mannheim, Germany

<sup>3</sup> University of Illinois at Urbana-Champaign, Champaign, USA

<sup>4</sup> Indian Institute of Technology Kharagpur, Kharagpur, India

## 1 Introduction

Algorithmic systems have become ubiquitous in our modern lives, and they exert great influence on many aspects of our daily lives, including shaping news and information we are exposed to via information retrieval algorithms. An important class of such retrieval algorithms is search systems. We all rely on search for a wide variety of goals in our day-to-day lives—ranging from finding specific website or content (navigational queries) to learning more broadly about entities, people, topics or events (informational queries) (Welch et al. 2011). For instance, during election season, people are known to make repeated queries about political candidates and events (e.g., “democratic debate”, “Donald Trump”, “climate change”) on the Web, as well as on social media sites like Facebook and Twitter (<http://tinyurl.com/OfficialGoogleBlog>; Teevan et al. 2011) to learn more about their queried terms.

While the goal of informational search queries is to provide users with greater knowledge about a topic, this knowledge is not necessarily always impartial. When a query is issued to a search system, a set of relevant items for the query are first extracted from the whole corpus of data items (e.g., web links or social media posts). This set of relevant items for the query are in turn fed to the ranking system which returns a ranked list of search results to the user who made the query. For polarizing topics like politics, many of these returned results can be biased towards one political perspective or the other, therefore by ranking items from one perspective higher than the other the ranking system could (possibly inadvertently) return a list of politically biased search results to the user. This bias in the search results could be introduced because of biased data that forms the input to the ranking system, or because of the ranking system itself.

The potential biases that search systems can introduce and users’ unquestionable trust in search results have lead to growing concerns about search systems’ impact on the behavior of users, especially in scenarios where they may potentially misinform or mislead the users. Prior field studies have shown that not only do the users place greater trust in highly ranked search results (Pan et al. 2007), but the opinions of undecided voters can be manipulated by biasing the search results about political candidates (Epstein and Robertson 2015). In such polarizing scenarios, where multiple different perspectives about the searched topic exist (e.g., political candidates or events), the bias in the top search results can influence the user’s opinion and shape public opinion towards (or away from) certain competing perspectives. However, such biases of search systems are challenging to detect and quantify, since multiple sources of bias exist (e.g., input data and ranking system) whose effects are hard to disentangle.

In this paper, we tackle this challenge by proposing a novel generalizable search bias quantification framework. This framework not only captures the bias in the search results output by a search system but is also capable of decoupling this output bias into different components to identify the sources of bias—the input data or the ranking system. For our chosen context of 2016 US Presidential primaries, we first apply our search bias quantification framework to political searches on social media (Twitter) to quantify and investigate its sources of political bias, and then we use our framework to quantify and compare relative bias for political searches on social media search (Twitter) and Web search (Google).

To apply our framework to study the sources of bias in political searches on Twitter social media, we first needed a methodology to measure the political bias of an individual search result, i.e., a tweet. We operationalized the political bias of a tweet as its source bias, i.e., the political bias of the author of the tweet, and developed a highly

scalable and accurate author based crowdsourced methodology for inferring the political bias of a Twitter user. We then utilized these inferred biases of tweets to quantify the sources of bias for Twitter search. Not only could we observe the search results output by the ranking system of Twitter search, but we were also able to gather the tweets containing a query which form the input to the ranking system. Armed with this data, we were able to disentangle the bias of different sources of bias for Twitter search, using our bias quantification framework. In our analyses of Twitter search results, we show that the bias in the search results does not only originate from the ranking system, but the bias of the input data (that is input to the ranking system) is also a significant contributor to the overall search bias. Moreover, we observe that the top Twitter search results display varying degrees of political bias that depends on several aspects, such as the topic (event/person) being searched for, the exact phrasing of the query (even for semantically similar queries), and also the time at which the query is issued.

After quantifying the bias in social media search, we proceed to use our quantification framework to compare the *relative bias* for political searches on two popular search systems - Twitter social media search and Google Web search. Our motivation for performing this comparison is to make the biases of different channels more visible and accessible to the users. Traditional media channels like Fox News or CNN have often been scrutinized by academics (Ribeiro et al. 2015; Babaei et al. 2018; Budak et al. 2016; Gentzkow and Shapiro 2010; Groseclose and Milyo 2005; Baron 2006; Munson et al. 2013b) as well as media watchdog groups (like FAIR (fair.org) and AIM (aim.org)) for fairness, accuracy and balance in the news they report. Additionally, tools have also been developed to mitigate or expose the media bias (Purple Feed 2018; Park et al. 2009; Munson et al. 2013b; <https://twitter-app.mpi-sws.org/media-bias-monitor/>; <https://mediabiasfactcheck.com>) to users. However, the relative biases of newer digital algorithmic channels like search systems are not as well studied and documented as yet, and thus users may not be taking their relative biases into account while selecting the channel to get their information from. In fact, many users believe that these algorithmically curated channels (as opposed to human editorial curation) are powerful, infallible and thus unbiased (Eslami et al. 2016; Springer et al. 2017), which is far from being true. This lack of awareness can result in “blind faith” in search systems (Pan et al. 2007), and impairs the users from making an informed choice of which search channel to use. With this study, we aim to highlight the differences in the political bias of these two popular search systems—Twitter social media search and Google Web search—and make their relative bias more visible.

Our comparison of relative bias of the two search systems reveals that the bias for political candidates is much more favorable to the candidates on Web search than on social media search. This difference is mainly due to multiple neutral or supportive (candidate-controlled) web-links (for instance candidate’s homepage or their social media profile links) that get included in the top results on Web search. We also observed that the bias in Web search results is less dynamic over time as compared to bias in social media search. Our findings show that search systems exhibit not only political bias in their search results but also different search systems exhibit different biases. It is important to highlight these differences in political bias of varying search systems, since the users currently may not be taking these biases into account when choosing one search system over the other to get information from.

Our research contributions in this work can be summarized as follows:

1. We propose a *novel generalizable search bias quantification framework* to measure not only the bias in the search output but also to discern the contribution of different sources—input data or ranking system (Sect. 3).
2. We apply the framework to *investigate the sources of bias* in political searches on Twitter social media search where we show that both input data and ranking system contribute to the final output bias seen by the users. We also observe that the bias varies with the topic being searched for, the exact phrasing of the query and the time at which the query is made. (Sect. 4).
3. We also utilize our framework to *compare the relative bias* for political queries on two popular search systems: Twitter social media search and Google Web search. As compared to social media search, we find that the political bias on Web search is a lot less dynamic, more favorable for the candidate queries, and has a higher fraction of top search results containing links to candidate-controlled sources, such as links to their website or social media profiles (Sect. 5).

Finally, a version of the present work has been published earlier at a conference (Kulshrestha et al. 2017). This paper extends and improves upon the earlier paper in the following manner: (i) We have strengthened the evaluation of bias inference for tweets by including a comparison of our source based scheme with a content based scheme in Sect. 4.3. Our results indicate that our bias inferred using our source based scheme has a higher (70% or more) match with the bias of the tweets (using human annotations), and performs better than content based schemes. (ii) We also included new results on the temporal variation of bias for political queries on Twitter social media search in Sect. 4.4.3. (iii) And finally, we have applied our bias quantification framework to study the relative bias of two popular search systems—Twitter social media search and Google web search. We report our findings about the comparison in Sect. 5.

Our work is aimed towards making social media users aware of the potential political biases of social media search and how it compare with the bias in web search and encouraging the development of novel information retrieval systems and mechanisms for presenting search results which could represent multiple competing perspectives on the same event or person.

## 2 Background

Today, algorithms that curate and present information on online platforms can affect users' experiences significantly. While powerful, these algorithms are not without flaws. Algorithms have been shown to create discriminatory ads based on gender (Datta et al. 2015) or race (Sweeney 2013), to show different prices for the same products/service to different users (Hannak et al. 2014), to skew users' ratings to benefit low-rated hotels (Eslami et al. 2017) and to mistakenly label a black man as an ape (Hern 2015). These issues have lead researchers, organizations and even governments towards a new avenue of research called "auditing algorithms", which endeavors to understand if and how an algorithmic system can cause biases, particularly when they are misleading or discriminatory to users (Sandvig et al. 2014; Executive Office of the President 2016).

Search engines are an important set of algorithms that users interact with on daily basis and these algorithms' susceptibility to bias has resulted in several audit studies in recent years. These audits cover a wide range of search platforms including Web search and social

media search. Next, we give an overview of prior work on examining the bias for search platforms, and discuss how our work adds to this line of existing research.

## 2.1 Bias in web search

In recent years, Web search engines and their potential biases have received a lot of scrutiny (Tavani 2014; Van Couvering 2010; Fortunato et al. 2006; Vaughan and Thelwall 2004; Mowshowitz and Kawaguchi 2005). This scrutiny has typically stemmed from the concern that dominant search engines like Google might favor certain websites over others when ranking relevant search results. For example, some argue that Google manipulates its search results to rank its services (such as Google Health links) higher than other competing services (Edelman 2010). In another example, Vaughan and Thelwall (2004) examined the geographical bias in Web search and observed that sites from certain countries like the US are covered more than sites from other countries.

Several studies have focused on the *political bias* of Web search results and queries during recent years. Weber et al. (2012) investigated the political leanings of search queries by linking the queries to political blogs. In another line of research, researchers conducted field studies to examine the influence of political bias seen in search results on users' voting decisions. For instance, Epstein and Robertson found that by manipulating the political bias in top search results they could impact the voting preferences of undecided voters by 20% or more (Epstein and Robertson 2015) and they termed this phenomenon as search engine manipulation effect. As a continuation of this line of research, Epstein et al. have shown that modifying the design of search engines to include alerts about the bias in the search results shown to the users can mitigate the aforementioned search engine manipulation effect significantly (Epstein et al. 2017a). Motivated by these findings, in this paper, we propose a generalizable search bias quantification framework and apply it to investigate the sources of bias in social media search as well as apply it to compare relative biases of social media and web search.

*Personalization in Web Search:* A complementary line of work has focussed on the personalization effects and studied the differences in the results seen by different users for the same query due to personalization. Various factors including geo-location of users has been found to lead to personalization of search results (Hannak et al. 2013; Kliman-Silver et al. 2015). On the other hand, in another study (Koutra et al. 2015) it was shown that during disruptive events such as shootings, the users tend to changes their information-seeking behavior and use the search engines to seek information that they agree with. In contrast, we study the bias in consistent, non-personalized search results for political queries shown to all users on social media search and the web search and we find that biases exist even for such non-personalized results. Addition of personalization is likely to add another source of bias for these search results and this is a potential direction of future research.

## 2.2 Bias in social media

With more and more users relying on social media platforms like Twitter and Facebook to receive news (Lichterman 2010) and information about on-going events and public figures (Teevan et al. 2011), there has been a debate about the impact these platforms are having on the news that users are consuming. While some have envisioned increased democratization with users from different political ideologies engaging with each other (Semaan

et al. 2014), others warned that use of social media platforms could encourage selective exposure by reinforcing users' existing biases (Liu and Weber 2014).

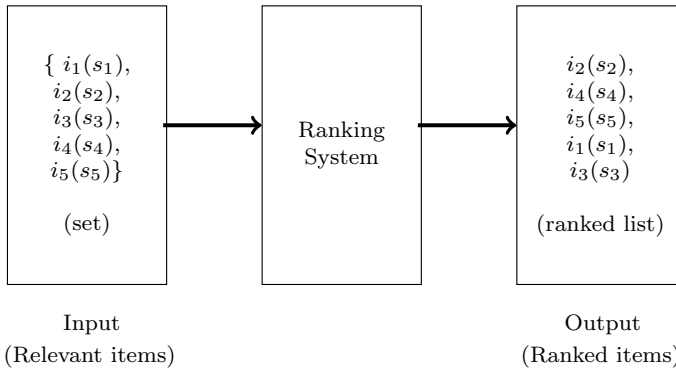
Further inspection of cross-ideological exposure (Himmelboim et al. 2013) revealed that political discourse on Twitter is highly partisan and users are unlikely to get exposed to cross-cutting content via their social neighborhood. These results have been reinforced by studies showing that not only are social media users more willing to communicate with other like-minded users (Liu and Weber 2014; Smith et al. 2013), they are also unable to engage in meaningful discussions with users with different beliefs than their own (Yardi and Boyd 2010). Therefore, political polarization on social media platforms has been an active area of research, with multiple different studies analyzing the behaviors of ideologically different groups of users. It has been shown that the retweeting network is highly partisan with users typically only retweeting other users who share their political ideology (Conover et al. 2011b).

Considerable research effort has also been dedicated to studying controversies and controversial topics online (Coletto et al. 2017; Garimella et al. 2016; Lu et al. 2015). Garimella et al. (2016) have proposed a method for quantifying controversy on social media using social media network and content. While BiasWatch is a system to discover and track bias themes from opposing sides of a topic in a semi-supervised manner (Lu et al. 2015).

While these studies give evidence of polarized and controversial content generation as well as sharing on social media platforms, it is unclear how this data impacts automated retrieval systems like search systems and the bias in their results. In this paper, we propose a search bias quantification framework which not only quantifies the bias in the output ranked list shown to the users, but it also discerns to what extent is this bias due to the ranking system of the search system, or the input data to the ranking system. We then apply this framework to study political bias in social media and web search.

### 2.3 Measuring political bias on social media and the web

The first step in quantifying the bias for political searches on social media and web search is measuring the political bias of an individual result (i.e., a tweet or a weblink). There have been some attempts to infer the political bias of blogs and news stories (Adamic and Glance 2005; Yano et al. 2010; Zhou et al. 2011) as well as hashtags on Twitter (Weber et al. 2013). However, there has been limited work on inferring the bias of content of short social media posts like tweets. Instead, researchers have inferred the bias of the users posting tweets by modeling how different polarity users use language (Purver and Karolina 2015; Makazhanov and Rafiei 2013; Fang et al. 2015), or by leveraging the linking behavior of users (Golbeck and Hansen 2011; Conover et al. 2011a, b), or by leveraging both textual and network features for political leaning classification (Pennacchiotti and Popescu 2011). Zafar et al. (2016), have quantified the impartiality of social media posts by measuring how easy it is to guess the political leaning of its author. Bond and Messing (2015) inferred the political leanings of Facebook users by observing the endorsements of Facebook Pages of known politicians, while Wong et al. (2016) measure the endorsements in terms of retweeting behavior of users to infer their political leanings. Cohen and Ruths (2013) used supervised methods to classify users into different groups of political activities and showed that it is hard to infer the political leaning of "normal" users. Most of these prior studies make the assumption that the leaning of the user is explicit in their language, social connections or endorsements, however this may not always be true. We build upon these prior studies to propose a methodology for inferring the bias of a Twitter user by



**Fig. 1 Overview of our search bias quantification framework.** For a given query  $q$ , a set of data items relevant to the query is first selected. Each individual data item (e.g.,  $i_1, i_2$ ) has an associated bias score (e.g.,  $s_1, s_2$ ). These set of relevant items is input to the ranking system which produces a ranked list of the items. Our framework includes metrics for measuring the bias in the set of relevant items input to the ranking system (input bias), and the bias in the ranked list output by the ranking system (output bias)

leveraging their interests, which are correlated to their political affiliation. And we use this proposed methodology to quantify the bias for political searches on Twitter social media and show that the method can be used to infer the political leaning of users with varied levels of political activities.

A number of prior studies have investigated political bias in traditional news media (Budak et al. 2016; Gentzkow and Shapiro 2010; Groseclose and Milyo 2005; Munson et al. 2013b). Budak et al. (2016) combined machine-learning and crowdsourcing techniques to study the selection and framing of political issues by news organizations. As online news sources have gained popularity, such studies have also been extended to them, as in the case of the Balance study (Munson et al. 2013b), which assigns political bias scores to many of the popular news websites based on the political leanings of the websites, blogs and Digg users that link to or vote for the news website. We utilize this prior work (Munson et al. 2013b) to quantify the bias of news search results on Twitter and the Web for comparing the relative biases of these two search systems.

### 3 Search bias quantification framework

The first research question we focus on pertains to quantifying the bias of a search system. In this work, we quantify the bias for political searches on Twitter social media search and Google web search, in the context of the US political scenario, which has two primary political parties: the Democratic party and the Republican party. In this section, we propose a bias quantification framework which captures the bias introduced at different stages of a search process, including metrics which measure the bias at each stage.

Figure 1 gives a high-level overview of the different stages of information retrieval via an algorithmic search system. The search system retrieves information from a corpus of data, where each individual data item (e.g.,  $i_1, i_2$ ) has an associated bias score (e.g.,  $s_1, s_2$ ). In the later sections (Sects. 4 and 5), we describe methodologies for computing the bias score for political searches on Twitter social media and Google web search platforms. When a user makes a query  $q$ , a set of data items relevant to the query is first selected out

of the whole corpus. Then, this set of retrieved relevant items forms the input data to the ranking system which produces a *ranked list of the relevant items*, which is shown as the search output to the users. The framework can also be generalized to modern-day IR systems which perform retrieval and ranking together, such as systems using topic modeling. We comment on this issue in Sect. 8.

Within our framework, we define three different components of the bias of a search system, each of which is quantified in terms of the biases of the individual data items: (i) *input bias*: the bias in the set of retrieved items relevant to the query that are filtered out of the whole corpus. This set of retrieved items serve as the input data to the ranking system, (ii) *ranking bias*: the bias introduced by the ranking system, and (iii) *output bias*: the cumulative bias in the ranked list output by the search system and shown to the users. In the rest of the section, we discuss the metrics we proposed to quantify these different components of bias of a search system.

### 3.1 Bias score of an individual data item

We are interested in quantifying the search bias for political queries in the context of US politics. Since there are two primary political parties in the US, each data item (e.g., a tweet or a web-link) can be positively biased (i.e., supportive), negatively biased (i.e., opposing), or neutral towards each of the parties. Therefore the bias score of each item must capture the extent to which it is biased with respect to the two parties.

To apply our bias quantification framework in the context of political searches on a search platform, we need a methodology for inferring the bias scores for each data item (indicated by  $s_i$  in Fig. 1). Later in the paper, we present methodologies for measuring the bias scores of individual items for our chosen scenario of political searches on Twitter social media (Sect. 4) and Google Web search platforms (Sect. 5).

Next, we use these bias scores of individual data items to define the metrics for the input bias, output bias, and ranking bias.

### 3.2 Input bias

Once a user issues a query, the search system retrieves a set of items from the whole corpus that are relevant to the query and provides them as an input to the ranking system. Since this input data captures the bias introduced due to the filtering of the relevant items from the data corpus according to the issued query, we measure the input bias for a query as the aggregate bias of all the items relevant to the query in this input data set. In other words, input bias gives a measure of the bias a user would observe if they were shown *random* items relevant to the query, instead of the output list ranked by the ranking system.

Specifically, the Input Bias  $IB(q)$  for query  $q$  is the average bias of the  $n$  data items that are relevant to  $q$

$$IB(q) = \frac{\sum_{i=1}^n s_i}{n} \quad (1)$$

where the summation is over all the bias scores ( $s_i$ ) of the  $n$  data items found relevant to  $q$ . For instance, for the query  $q$  shown in Fig. 1, the input bias is  $IB(q) = \frac{1}{5}(s_1 + s_2 + s_3 + s_4 + s_5)$ .



**Table 1** Explaining the bias metrics with reference to Fig. 1

Rank $r$	Bias till rank $r$	Value
1	$B(q, 1)$	$s_2$
2	$B(q, 2)$	$\frac{1}{2}(s_2 + s_4)$
3	$B(q, 3)$	$\frac{1}{3}(s_2 + s_4 + s_5)$
4	$B(q, 4)$	$\frac{1}{4}(s_2 + s_4 + s_5 + s_1)$
5	$B(q, 5)$	$\frac{1}{5}(s_2 + s_4 + s_5 + s_1 + s_3)$
Output bias at rank 5		$\frac{1}{5}[s_2(1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5})$ $+s_4(\frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5})$ $+s_5(\frac{1}{3} + \frac{1}{4} + \frac{1}{5})$ $+s_1(\frac{1}{4} + \frac{1}{5})$ $+s_3(\frac{1}{5})]$

### 3.3 Output bias

The output bias of a search system is the cumulative bias in the final ranked list of search results presented to the user who issued the search query. Prior studies have shown that not only are the users more likely to browse the top search results (Manning et al. 2008), but they also tend to put more trust in them (Pan et al. 2007). Therefore, we propose an output bias metric inspired from the well-known metric—mean average precision (Manning et al. 2008)—which gives more importance to higher ranked search results.

For a given search query  $q$ , we first define the bias till a particular rank  $r$  in the ranked results (i.e., the aggregate bias of the top  $r$  results). The bias  $B(q, r)$  till rank  $r$  of the output ranked list is defined as

$$B(q, r) = \frac{\sum_{i=1}^r s_i}{r} \tag{2}$$

where the summation is over the top  $r$  items in the ranked list.

As an example, the first five rows in Table 1 depict the bias till ranks 1, 2, ..., 5, for the sample search scenario shown in Fig. 1.

The Output Bias  $OB(q, r)$  for the query  $q$  at rank  $r$  is then defined by extending the above definition as follows,

$$OB(q, r) = \frac{\sum_{i=1}^r B(q, i)}{r} \tag{3}$$

The last row of Table 1 depicts  $OB(q, r)$  at rank  $r = 5$  with respect to Fig. 1. In this formulation, the bias score  $s_2$  of the top-ranked item  $i_2$  is given the highest weight, followed by the bias score  $s_4$  of the second-ranked item  $i_4$ , and so on, following the intuition that the bias in the higher ranked items is likely to influence the user more than bias in lower ranked items.<sup>1</sup>

<sup>1</sup> Similar to how missing relevance judgements are handled in the Information Retrieval literature (Yilmaz and Aslam 2006), in case there exists an item for which the bias score cannot be computed, we just ignore the item and compute the rankings.

### 3.4 Ranking bias

If the internal details of the deployed ranking system were known, then the ranking bias could be measured by auditing the exact features being used for ranking. However, for most of the real-world commercially deployed search engines, the internal details of the ranking system are not known publicly. Therefore, building on previous studies that have adopted a “black-box” view for an algorithmic system while auditing it (Eslami et al. 2015; Liao et al. 2016; Hannak et al. 2013, 2014; Chen et al. 2015), we treat the ranking system as a black-box, such that we only observe its inputs and outputs. In such a scenario, the ranking bias captures the *additional* bias introduced by the ranking system, over the bias that was already present in the input set of relevant items.

Therefore, we define the Ranking Bias  $RB(q, r)$  for the query  $q$  as simply the difference between the output bias and the input bias for  $q$  (as given by Eqs. 1 and 3).

$$RB(q, r) = OB(q, r) - IB(q) \quad (4)$$

### 3.5 Time-averaged bias

To capture the overall trend in the bias, we collect multiple snapshots of search results, compute the different bias metrics for each snapshot, and then compute the time-averaged values of the aforementioned metrics. For instance we compute the time-averaged output search bias  $TOB(q, r)$  as the average of the  $OB(q, r)$  (given by Eq. 3) values measured at various instants of time. Similarly, we define  $TIB(q)$  and  $TRB(q, r)$  as the time-averaged input bias and time-averaged ranking bias for query  $q$  respectively.

## 4 Investigating sources of bias for political searches on social media

Having described our search bias quantification framework, we next apply it to political searches on Twitter social media for queries related to 2016 US presidential primaries. With this study, we highlight an important application scenario of our framework, where not only can we observe the search system’s output results, but we also can observe the set of relevant items that form the input to the ranking system.

We begin by describing our selected queries and data set for Twitter search (Sect. 4.1), followed by the methodology we used for measuring the political bias of an individual Twitter search result (Sect. 4.2), and then we finally present our findings about how biased are the search results for political topics on Twitter and where does this bias in the search results comes from (Sect. 4.4).

### 4.1 Collecting Twitter search data

Here, we describe the queries we considered and the data gathered from Twitter for conducting the analyses.

### 4.1.1 Selecting search queries

In an ideal scenario, for studying the bias in political searches on Twitter, we would use the actual search queries that people are making on the platform for following news and information related to 2016 US presidential primaries. However, we did not have access to this proprietary data about the queries issued on Twitter. In the absence of the actual search queries issued on Twitter, we followed the methodology used in Koutra et al. (2015) of first identifying a seed set of queries and then expanding them to identify a larger set of potential queries. Our seed set consists of the queries *democratic debate*, and *republican debate*, and their shortened versions (*dem debate* and *rep debate*) popular on Twitter because of their short lengths.

We wanted our expanded set of queries to satisfy two properties: (i) they should be popular and be used by many users, and (ii) they should not be biased towards any particular party, candidate or organization in their formulation, i.e., the leaning of the user issuing the query should not be obvious from the query.

To satisfy the first property of selecting popular queries, we focused on hashtags for expanding the query set. This choice was bolstered by the knowledge that hashtags are used extensively on Twitter to tag and follow discussions about politics (Conover et al. 2011a). Additionally, every time a user clicks on a hashtag, a Twitter search page with the hashtag as the query opens up, making hashtags effectively act as recommended queries on Twitter. To identify such popular hashtags, we collected the Twitter search results for our four seed queries during the November 2015 Republican and Democratic debates. We then identified top 10 most frequently occurring hashtags for each of the debate’s dataset which contained the term “debate” in them (to ensure they are about the primary debates), resulting in a total of 15 distinct hashtags (*#debate*, *#demdebate*, *#democraticdebate*, *#republican-debate*, *#gopdebate*, *#debatewithbernie*, *#hillarycantdebate*, *#debatewithbe*, *#nprdebate*, *#cnndebate*, *#cnbcgopdebate*, *#fbngopdebate*, *#foxbusinessdebate*, *#gopdebatequestions*, *#gopdebate moderators*).<sup>2</sup>

Due to our second desirable property, we wanted to retain only the unbiased queries from the above 15 hashtags, to avoid over-estimating the bias in the search results. Doing so, we removed queries which were biased towards (or against) a candidate (*#debatewithbernie*, *#hillarycantdebate*, and *#debatewithbe*),<sup>3</sup> an organization (*#nprdebate*, *#cnndebate*, *#cnbcgopdebate*, *#fbngopdebate*, and *#foxbusinessdebate*), or a party (*#gopdebatequestions*, and *#gopdebate moderators*). Therefore, we were left with the expanded set of 8 queries which are popular and for whom it was hard to guess the political leaning of the user issuing the query—*democratic debate*, *dem debate*, *#democraticdebate*, *#demdebate*, *republican debate*, *rep debate*, *#republicandebate* and *#gopdebate*.

In addition, we also included the names of the 17 presidential candidates, resulting in a total of 25 queries, which we used to measure the bias for political searches on Twitter. Table 8 shows the exact phrasings of the 25 queries from our dataset.

<sup>2</sup> We did not include *#debate* in our selected query dataset because it was too generic and many tweets containing it were about topics unrelated to 2016 US Presidential Primaries.

<sup>3</sup> For example, we observed that the hashtag *#debatewithbernie* was biased towards Bernie Sanders (and the Democratic party), with *#FeelTheBern*, *#BernieSaidItFirst* and *#Bernie2016* being the hashtags which co-occurred with *#debatewithbernie* the most.

### 4.1.2 Data collection from Twitter

For applying our bias quantification framework to Twitter search, we needed to collect data about the output search results given out by Twitter’s ranking algorithm, as well as the set of tweets which were relevant to our selected queries that form the input to the ranking system. For performing our bias analysis, we collected the search data for a one week period in which both a Democratic debate (December 19, 2015) and a Republican debate (December 15, 2015) took place—14–21 December 2015.

Even though Twitter provides multiple different filters for their search functionality, we collected the search snapshots for our set of selected queries for the default filter of “top” search results (<https://twitter.com/search-home>). The “top” search results are the output of Twitter’s proprietary ranking system, which performs ranking based on a multitude of factors, including the number of users engaging with a tweet (<https://help.twitter.com/en/using-twitter/top-search-results-faqs>). During the one week period, search snapshots were collected at 10-min intervals for each query. Each snapshot consists of the top 20 results on the first page of search results, and we used these snapshots to compute the output bias for the queries. Across all queries, we collected a total of 28,800 snapshots which consisted of 34,904 distinct tweets made by 17,624 distinct users.

Finally, we used Twitter’s streaming API to collect the tweets containing our selected queries during this one week period, and this set of tweets formed the input to the ranking system and were used to compute the input bias for the queries.<sup>4</sup> Across all queries, we collected more than 8.2 million tweets posted by 1.88 million distinct users.

*Collecting non-personalized search results:* In this work, we focus on quantifying the bias in consistent, non-personalized search results shown to every user, therefore to mitigate the personalization effects we made all the search queries from the same IP subnet (in Germany), and without logging in to Twitter.

## 4.2 Measuring political bias of an individual search result

To apply our bias quantification framework to Twitter search for queries related to US presidential primaries, we need a methodology for inferring the political bias of an individual result—a tweet. The short length of tweets (140 characters) makes it very challenging to infer the bias of a tweet from its content (i.e., to measure its *content bias*). Instead in this work, we operationalize the bias of a tweet as its *source bias*, i.e., we approximate the bias of a tweet with the political bias of the author of the tweet.

In the rest of this section, we begin by presenting our methodology for inferring source bias of a tweet and then present our evaluation results. Finally, we end with a short analysis of how well source bias and content bias of a tweet match each other in practice for political searches on Twitter.

---

<sup>4</sup> We observed that 74.8% of tweets included in the search results were also included in the data that we collected via the streaming API. In comparison, prior work that compared (Morstatter et al. 2013) data collected using Twitter’s Streaming API with Twitter’s Firehose (full Twitter stream), found that on average, the Streaming API contained 43.5% of data available on the Firehose on any given day.

#### 4.2.1 Source bias: inferring political bias of Twitter users

Prior studies have shown that people's political affiliations are correlated with their personality attributes and responses to different stimuli (Carney et al. 2008; Shi et al. 2017; <http://2012election.procon.org/view.resource.php?resourceID=004818>). Based on this knowledge, we propose a methodology for inferring political leaning of Twitter users by leveraging their interests. Therefore, our methodology for inferring the political bias of a Twitter user  $u$ , is based on the following three steps:

1. *Generating representative sets of Democratic and Republican users:* We use the crowd-sourced methodology described in Ghosh et al. (2012) and Sharma et al. (2012), which infers the topical attributes of a user  $v$  by mining the Twitter Lists that the other users have included  $v$  in. By relying on what others are reporting about a user, rather than what the users are identifying themselves as, we avoid the self-reportage problem, as well as avoid biasing the sets towards the group of users who have self-reported. Following this methodology, we identified a seed set of 865 users labelled as “Democrats” and 1348 users labelled as “Republicans”. These seed sets include known politicians (e.g., Steny Hoyer, Matt Blunt), political organizations (e.g., DCCC, Homer Lkprt Tea-party) as well as regular users.
2. *Inferring topical interests of a user:* To infer the interests of a user  $u$  we rely on the methodology developed by Bhattacharya et al. (2014a, b), which for a user  $u$ , returns a list of topics of interest of  $u$  along with the number of users whom  $u$  follows who have been labeled with this topic using the methodology described in Ghosh et al. (2012) and Sharma et al. (2012). Therefore, our method leverages the network neighborhood of  $u$  to infer the interests and hence the political leaning of  $u$ . For instance, if a user  $u$  follows three users tagged with ‘politics’ and four users tagged with ‘entertainment’, then the returned list would be {politics: 3, entertainment: 4}. We convert this <topic, #users> list into a weighted  $tf\_idf$  vector for user  $u$  (where the  $idf$ -s are computed considering the interest lists of all the users in our dataset) and refer to it as the *interest-vector*  $I_u$  of the user  $u$ . We are not able to infer the topical interests of a user when either their accounts are protected, and we can not gather the users they are following or because they follow too few other users (less than 10). But in prior work, it has been shown that such cases are few, and this methodology infers the interests of a significant fraction of active users on Twitter (Bhattacharya et al. 2014a, b).
3. *Matching user's interests to interests of Democrats and Republicans:* We first compute the representative interest vectors for Democrats ( $I_D$ ) and Republicans ( $I_R$ ) by aggregating the interest vectors of users in each set and normalizing such that  $I_D$  and  $I_R$  vectors sum up to 1 each. These aggregate vectors not only capture the differences in the political interests of Democrats and Republicans (e.g., [progressive, democrats, obama, dems, liberals] and [patriots, conservative, tcot, right, gop] are the top terms in  $I_D$  and  $I_R$  respectively), but also the differences in their non-political interests (e.g.,  $I_R$  has higher weight for sports-related terms, while  $I_D$  has higher weight for technology and entertainment related terms). Therefore, even in the case of users who don't follow any politicians on Twitter or the ones who follow politicians from both parties, these representative vectors can be used to infer their likely political bias. Finally, the *bias score* of user  $u$  with interest vector  $I_u$  is given by the difference in the cosine similarities of  $I_u$  with  $I_D$  and  $I_R$ ,

$$Bias(u) = \cos\_sim(I_u, I_D) - \cos\_sim(I_u, I_R). \quad (5)$$

We max-min normalize the scores such that the bias score of a user lies in the range  $[-1.0, 1.0]$ , with a score closer to  $+1.0$  indicating more Democratic bias, while a score closer to  $-1.0$  indicating more Republican bias.

*Public deployment of the source bias inference methodology:* We have publicly deployed the aforementioned source bias inference methodology in the form of a Twitter application, at <http://twitter-app.mpi-sws.org/search-political-bias-of-users/>. One can login to the application using their Twitter credentials, and see their inferred political affiliation. One can also search for other Twitter users to check out their inferred political leaning.

#### 4.2.2 Evaluation of political bias inference methodology

To validate whether our bias inference method works well for a whole spectrum of politically interested users, we perform the evaluation over three test sets of Twitter users—(i) politically interested common users, selected randomly from the set of users who have retweeted the two parties' accounts on Twitter, (ii) the current US senators, and (iii) self-identified common users (with fewer than 1000 followers), who have identified their political ideology in their account bios. We use two metrics for evaluating the methodology: (i) *coverage*—for what fraction of users can the methodology infer the political bias, and (ii) *accuracy*—for what fraction of users is the inference correct.

We begin by using the set of politically interested common users to evaluate our inferred bias scores, followed by a description of how we discretize our bias score into three distinct categories—Republican, neutral and Democratic, and we end by presenting our methodology's performance in inferring the political bias of senators and self-identified common users.

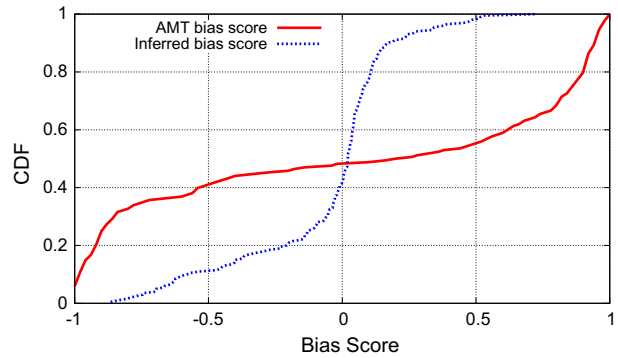
##### Evaluation for politically interested common users

*Identifying politically interested common users:* Following the methodology developed by Liu and Weber (2014), we collected up to 100 retweeters of each of the latest 3,200 tweets posted by the accounts of the two political parties—@TheDemocrats and @GOP. We removed the retweeters which retweeted the accounts of both the parties, obtaining 98,955 distinct retweeters of @TheDemocrats, and 71,270 distinct retweeters of @GOP. From each of these two sets of retweeters, we randomly selected 100 retweeters, giving us a total of 200 politically interested common users.

*Ground truth bias of test users:* We collected the ground truth bias annotations for these 200 politically interested users by conducting an AMT survey where human workers were shown a link to user's Twitter profile. We only used Master workers from the US who have had at least 500 HITS approved, with an approval rating of 95%. We paid the workers 4\$ for judging the political leaning of 45 Twitter users. The workers were asked to infer the user's political leaning as either pro-Democratic, pro-Republican or neutral based on the user's profile and tweets. For each user, we aggregated the judgements of 50 workers, adding  $+1$  for each pro-Democratic,  $-1$  for pro-Republican and  $0$  for each neutral judgement and normalizing by the total number of judgements to get an AMT bias score in the range  $[-1.0, 1.0]$ , where a more positive score indicates a stronger Democratic bias, while a more negative score indicates a stronger Republican bias.

*Evaluating our inferred score:* With our methodology, we were able to infer the bias of all 200 users (i.e., coverage is 100%). To quantify the accuracy of the methodology, we checked whether our inferred bias scores correlate well with the AMT bias scores. To

**Fig. 2** CDF of AMT bias scores and Inferred bias scores for politically interested common users



**Table 2** Confusion matrix of the match between AMT bias scores and Inferred bias scores

AMT bias score	Inferred rep (%)	Inferred neutral (%)	Inferred dem (%)
AMT Bin 1	84.05	13.04	2.89
AMT Bin 2	18.18	45.45	36.36
AMT Bin 3	3.89	12.98	83.11

**Table 3** Coverage and accuracy of the political bias inference methodology for (i) current US senators, and (ii) common users who have declared their political ideology in their Twitter account profiles

Political bias	Coverage (%)	Accuracy (%)
<i>Current US senators</i>		
Democratic ( $n = 45$ )	97.78	86.36
Republican ( $n = 54$ )	98.15	98.11
Average	97.96	92.23
<i>Self-identified common users</i>		
Democratic ( $n = 426$ )	92.01	88.52
Republican ( $n = 675$ )	90.22	82.95
Average	91.12	85.73

verify this, we binned our inferred bias score into three bins: Bin 1  $[-1.0, -0.5]$ , Bin 2  $(-0.5, 0.5)$ , and Bin 3  $[0.5, 1.0]$  and computed the average AMT bias scores for each bin. We observe a strongly Republican leaning score  $(-0.86)$  for Bin 1, while a strongly Democratic leaning score  $(0.93)$  for Bin 3. We observe a similar trend if we bin according to the AMT bias scores and compute the average inferred score for each bin  $(-0.32)$  for Bin 1 and  $0.14$  for Bin 3), demonstrating a good correlation between the two bias scores.

*Discretizing the bias score into categories:* While the inferred bias scores are highly correlated with the AMT bias scores, we observe that the distribution (CDF) of the two scores in the interval  $[-1.0, 1.0]$  are different, as shown in Fig. 2. Due to this difference in the distributions of the two scores, we decided to discretize our inferred bias score, and categorize users as—neutral, Democratic or Republican leaning.

In order to do the discretization, we needed to identify a suitable threshold  $x$  on our inferred score, such that users with scores in the range  $(-x, x)$  are categorized as neutral, while the ones with scores  $x$  and above are identified as Democratic leaning, while  $-x$  and

below are identified as Republican leaning. We experimented with  $x = 0.01, 0.03, 0.05, 0.08$  and  $0.1$ , and for each of these values computed a confusion matrix of the match between the AMT bias score and our inferred bias score. We selected  $x = 0.03$  to be the threshold as it maximizes the sum of the diagonal of the confusion matrix, as shown in Table 2. In the rest of this section, we will only label the users as Republican or Democratic leaning when their bias scores lie outside of the neutral zone ( $-0.03, 0.03$ ). We make this conservative choice to not overestimate the bias in the search results.

### Evaluation for US senators

Table 3 outlines the performance of our methodology for the 100 current US senators (45 Democrats, 54 Republicans, 1 Independent), showing that our methodology has very high coverage. Closer inspection of the two senators, for whom we could not infer the bias, disclosed that one of them does not follow any other users on Twitter while the other follows only one, making it impossible for us to infer their interests and consequently their bias. Our methodology also performs well in terms of accuracy by correctly identifying the bias for 86.4% of Democratic senators and 98.1% of Republican senators, out of the ones for whom we could infer the bias.

### Evaluation for self-identified common users

We collected our final set of self-identified common users using the service Followerwonk and gathering users located in the US, with less than 1000 followers, and whose Twitter account biographies contained keywords matching Democrats ('democrat', 'liberal', 'progressive') or Republicans ('republican', 'conservative', 'libertarian', 'tea party'). We manually inspected each user, and pruned out any users whose bios did not reflect their political ideology. For instance, users with erroneous bios like "*I am a #conservative #Christian who is neither a #Democrat nor a #Republican, but an #Independent voter*" and "*We hate Politicians - Democrats, Republicans, all of them.*" were removed. Following this procedure, we collected a total of 426 self-identified Democratic users, and 675 self-identified Republicans.

Table 3 also depicts the performance of our methodology for these self-identified users. The average coverage is again high (91.1%), with the users for whom we could not infer the bias either having protected accounts or following too few users such that it was impossible for us to infer their interests and therefore their political bias. Our proposed method also has a high accuracy of 85.7% on average across all these self-identified common users for whom we could infer the bias.

Further inspection of interest vectors of the users for whom we correctly inferred the political leaning reveals that the interest vectors of Democratic users not only contain political terms like 'liberal', 'progressive', and 'dem', but also other terms including 'gay', 'lgbt', 'science', and 'tech', while the interest vectors of Republican users contain terms like 'tea', 'gop', and 'palin' along with other related terms like 'patriots', 'military', and 'vets'.

### 4.3 Match between source bias and tweet bias

In this section, we focus on answering the question, "*how closely do source bias and bias of a tweet reflect each other?*".



**Table 4** Confusion matrix for source bias classification—gold standard tweet bias (based on AMT workers' judgement) versus source bias

Gold standard	Source bias		
	Republican (%)	Neutral (%)	Democratic (%)
Tweet bias			
Republican [− 1.0, 0.5]	70.44	9.36	20.2
Neutral (− 0.5, 0.5)	27.61	16.96	55.43
Democratic [0.5, 1.0]	11.71	10.24	78.05

**Table 5** Confusion matrix for content bias classification (support vector machine (SVM) classifier)—gold standard tweet bias (based on AMT workers' judgement) versus content bias

Gold standard	Content bias		
	Republican (%)	Neutral (%)	Democratic (%)
Tweet bias			
Republican [− 1.0, 0.5]	39.11	40.22	20.67
Neutral (− 0.5, 0.5)	16.67	76.19	7.14
Democratic [0.5, 1.0]	24.35	50.00	25.65

**Table 6** Confusion matrix for content bias classification (gradient boosted decision tree (GBDT) classifier)—gold standard tweet bias (based on AMT workers' judgement) versus content bias

Gold standard	Content bias		
	Republican (%)	Neutral (%)	Democratic (%)
Tweet bias			
Republican [− 1.0, 0.5]	79.88	11.17	8.94
Neutral (− 0.5, 0.5)	57.14	35.72	7.14
Democratic [0.5, 1.0]	64.10	8.97	26.93

*Measuring tweet bias:* For each of our selected queries, we gathered two search snapshots from our chosen period in December 2015, one during the Republican debate and one during the Democratic debate. Across all these snapshots, we gathered a total of 881 distinct tweets, and we use these to evaluate the extent to which the tweet bias matches the inferred source bias. We use AMT workers to measure the tweet bias by showing each tweet (but not the user who posted it) to 10 AMT workers and asking them to label the tweet as either pro-Democratic, pro-Republican or neutral. Then following the methodology in Sect. 4.2.2, we computed a tweet bias score for each tweet by aggregating the judgements of the 10 AMT workers. Using these scores, we generated the gold standard labels for the bias of the tweets, by dividing the range of AMT tweet bias scores into 3 intervals and labelling tweets in interval [− 1.0, 0.5] as Republican, in interval (− 0.5, 0.5) as neutral, and in interval [0.5, 1.0] as Democratic-leaning.

*How closely do source bias and tweet bias match each other?:* To investigate the match between source bias and tweet bias, Table 4 presents the confusion matrix for our source bias inference methodology. We observe that when the content is biased on either side, the match between source and AMT gold standard tweet bias is high (70% or more) indicating that strongly biased content is produced mostly by users with the same bias.

*How does our source based scheme compare to content based scheme for inferring the bias of a tweet?:* To evaluate how well does a content-based scheme work for inferring bias of social media posts, especially in comparison with our source based methodology, we represented the tweets by a bag-of-words model (i.e., using every distinct unigram as a feature) and applied two well-known classifiers—Support Vector Machine (SVM) and Gradient Boosted Decision Tree (GBDT).<sup>5</sup> The unigram features were generated from the tweet text by applying preprocessing steps of case-folding, stemming, stop word removal and removal of URLs. We used 5-fold cross validation for all the classification experiments.

Tables 5 and 6 depict the confusion matrices for the SVM and GBDT content based classifiers respectively. Comparing with Table 4, we observe that our source based method performs better than the content based scheme. While the accuracy for SVM classifier is quite low, the GBDT classifier seems to classify most tweets as Republican. However, we want a classifier where errors for the different classes are balanced, so that one class is not grossly over-estimated, and from this perspective also our source-based classification performs better.

#### 4.4 Characterizing the bias for political searches on Twitter social media

Having described our bias inference methodology, as well as the search data that we collected for political searches on Twitter social media, we next focus on analyzing the collected data to characterize the bias for political searches on Twitter. We begin by investigating the contributions of the two sources of bias—input data and ranking system—to the final output bias seen by the users. Then we examine the interplay between the input data and the ranking system that produces the output bias seen by the users. We end with an analysis of the variation of bias over time.

##### 4.4.1 Where does the bias come from: input data or ranking system?

*It is not always the ranking system, input data matters:* We show the three biases (output, input and ranking bias) for all our selected queries in Table 7. When we compute the average biases for the four sets of queries—Democratic and Republican candidates and debates—we find that the average input biases for all four sets are Democratic-leaning (i.e., larger than 0). Although the average input bias for Republican candidates and debates is less Democratic-leaning than Democratic ones, the full tweet stream containing all these query terms (without any interference from the ranking system) on an average contains a more Democratic slant. We observe that the input bias proves to be a prominent contributor to the final output bias seen by the users. For instance, the output bias for *Bernie Sanders* is very Democratic (0.71), with only a small amount of the bias being contributed by Twitter’s ranking system (0.16); the majority of bias originates from the input data (0.55),

<sup>5</sup> Currently, we have used unigrams as features, however in the future other features including n-grams as well as other classification methods can be explored to improve the content-based baselines.

**Table 7** Time-averaged bias in Twitter search “top” results, for selected queries (related to political candidates and debates)—output bias *TOB*, input bias *TIB*, and ranking bias *TRB*

Query	Output bias (TOB)	Input bias (TIB)	Ranking bias (TRB)
<i>Queries related to democratic candidates</i>			
<i>Hillary Clinton</i>	0.21	0.03	0.18
<i>Bernie Sanders</i>	0.71	0.55	0.16
<i>Martin O’Malley</i>	0.64	0.57	0.07
Average	0.52	0.38	0.14
<i>Queries related to republican candidates</i>			
<i>Donald Trump</i>	0.29	0.19	0.10
<i>Ted Cruz</i>	− 0.48	− 0.11	− 0.37
<i>Marco Rubio</i>	− 0.41	− 0.12	− 0.29
<i>Ben Carson</i>	0.46	0.20	0.26
<i>Chris Christie</i>	− 0.14	0.27	− 0.41
<i>Jeb Bush</i>	− 0.31	0.09	− 0.40
<i>Rand Paul</i>	− 0.37	− 0.18	− 0.19
<i>Carly Fiorina</i>	0.16	0.38	− 0.22
<i>John Kasich</i>	− 0.09	− 0.13	0.04
<i>Mike Huckabee</i>	0.30	0.12	0.18
<i>Rick Santorum</i>	− 0.04	0.18	− 0.22
<i>Lindsey Graham</i>	− 0.45	0.07	− 0.52
<i>George Pataki</i>	− 0.17	0.09	− 0.26
<i>Jim Gilmore</i>	− 0.35	− 0.11	− 0.24
Average	− 0.11	0.07	− 0.18
<i>Queries related to democratic debate</i>			
<i>democratic debate</i>	0.43	0.38	0.05
<i>dem debate</i>	0.52	0.29	0.23
<i>#democraticdebate</i>	0.28	0.19	0.07
<i>#demdebate</i>	0.57	0.56	0.01
Average	0.45	0.35	0.10
<i>Queries related to republican debate</i>			
<i>republican debate</i>	0.53	0.27	0.26
<i>rep debate</i>	0.31	0.40	− 0.09
<i>#republicandebate</i>	0.39	0.34	0.05
<i>#gopdebate</i>	0.04	0.10	− 0.06
Average	0.32	0.28	0.04

Here a bias value closer to + 1.0 indicates Democratic bias and a value closer to − 1.0 indicates Republican bias

indicating that most of the users that discuss *Bernie Sanders* on Twitter have a Democratic leaning. The effect of input data on the output bias highlights the importance of also taking into account the input data while auditing algorithms, to discern how much of the bias is due to the data and how much is contributed by the algorithmic system. This insight is particularly crucial in this digital era where many algorithms are trained using vast amounts of data (Barocas and Selbst 2014).

We also measured the *bias of overall Twitter corpus* in two ways: (i) *User population bias*: measured as the average bias of 1000 Twitter users selected randomly from the Twitter user-id space (i.e., the user-ids were randomly selected from the range of 1 through the id assigned to a newly created account in December 2015), and (ii) *Full tweet stream bias*: measured as the average source bias of 1000 tweets selected randomly from Twitter's 1% random sample for December 2015. We found the user population bias to be 0.25 and a full tweet stream bias to be 0.3 indicating that not only is the population of Twitter Democratic-leaning, but the active users (whose tweets have been included in Twitter's 1% random sample) are even more Democratic-leaning. These findings are in-line with prior studies (<http://www.pewresearch.org/2013/03/04/twitter-reaction-to-events-often-at-odds-with-overall-public-opinion/>) which have shown that Twitter has a high fraction of Democratic-leaning users.

Although Twitter has a Democratic-leaning corpus bias, the input bias (TIB) of the different queries varies across the spectrum (as shown in Table 7). This variation in bias likely occurs because each query acts as a filter to extract a subset of Twitter users whose tweets are relevant to that query, and the sets of users filtered out by different queries have differing biases. Therefore, even with the corpus bias of Twitter being the same, each query determines the input data set and hence the input bias, which in turn affects the final output bias observed by the user for that query.

*The power of the ranking system*: Although input data does contribute to the final output bias, the ranking system also exerts power over the final bias by shifting the bias or even changing its polarity, as demonstrated by the ranking biases shown in Table 7. Even though we observed that the input biases for both the Democratic and Republican candidates on an average were Democratic-leaning, we notice that on an average the ranking system adds a Democratic-leaning ranking bias for the Democratic candidates making the output more Democratic-leaning ( $TOB = 0.52$ ), while it adds a Republican-leaning ranking bias for Republican candidates making the output more Republican-leaning ( $TOB = -0.11$ ). This change of polarity from a Democratic-leaning input bias to a Republican-leaning output bias is particularly noticeable for some Republican candidates like *Chris Christie*, *Jeb Bush* and *Lindsey Graham*. These shifts in the bias caused by the ranking system (that can also result in a polarity change), exhibit the ranking system's power in altering the inherent bias of the input data.

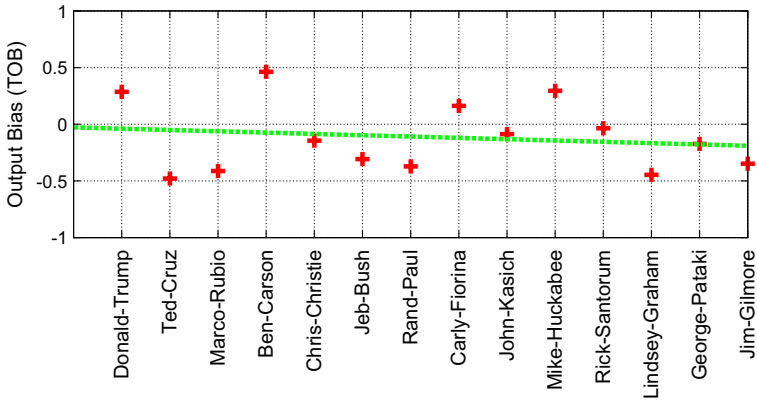
The ranking of posts in social media search systems is a complex process with the platform providers trying to provide the most relevant posts within the highest ranked items. They use a number of factors to measure the relevance of posts for ranking search results, including the keywords it contains, the popularity of the post in terms of users' engagements with it (e.g., number of retweets, favorites or replies) (<https://help.twitter.com/en/using-twitter/top-search-results-faqs>, [https://blog.twitter.com/engineering/en\\_us/a/2014/building-a-complete-tweet-index.html](https://blog.twitter.com/engineering/en_us/a/2014/building-a-complete-tweet-index.html)), as well as the recency of the post ([https://blog.twitter.com/engineering/en\\_us/topics/infrastructure/2016/search-relevance-infrastructure-at-twitter.html](https://blog.twitter.com/engineering/en_us/topics/infrastructure/2016/search-relevance-infrastructure-at-twitter.html)). Our goal in this work is not to reverse engineer Twitter's ranking system. However, we take a step towards gaining insight into the ranking system of Twitter by examining the impact of the popularity of the posts on the search rankings. For doing so, we take the posts included in Twitter's top search results and rerank them based on the popularity of the post (i.e., the number of retweets and the number of favorites). We then compared the bias of these simulated rankings with the ranking bias of Twitter's ranking system (shown in Table 8). For most of our queries, the ranking biases of the three strategies are quite similar to each other, indicating that popularity of the post can explain much of the observed bias in Twitter's ranking.

**Table 8** Time-averaged ranking bias for different ranking strategies: (i) Twitter’s ranking (Twitter search “top” results), (ii) Most retweeted tweet first ranking, and (iii) Most favorited tweet first ranking

Query	TRB of ranking strategies		
	Twitter’s ranking	Most retweeted first	Most favorited first
<i>Queries related to democratic candidates</i>			
<i>Hillary Clinton</i>	0.18	0.33	0.25
<i>Bernie Sanders</i>	0.16	0.22	0.16
<i>Martin O’Malley</i>	0.07	0.001	0.1
<i>Queries related to republican candidates</i>			
<i>Donald Trump</i>	0.10	0.06	0.09
<i>Ted Cruz</i>	− 0.37	− 0.49	− 0.35
<i>Marco Rubio</i>	− 0.29	− 0.36	− 0.27
<i>Ben Carson</i>	0.26	0.23	0.25
<i>Chris Christie</i>	− 0.41	− 0.40	− 0.34
<i>Jeb Bush</i>	− 0.40	− 0.46	− 0.34
<i>Rand Paul</i>	− 0.19	− 0.25	− 0.17
<i>Carly Fiorina</i>	− 0.22	− 0.17	− 0.18
<i>John Kasich</i>	0.04	0.04	0.11
<i>Mike Huckabee</i>	0.18	0.11	0.19
<i>Rick Santorum</i>	− 0.22	− 0.34	− 0.16
<i>Lindsey Graham</i>	− 0.52	− 0.45	− 0.56
<i>George Pataki</i>	− 0.26	− 0.22	− 0.23
<i>Jim Gilmore</i>	− 0.24	− 0.22	− 0.21
<i>Queries related to democratic debate</i>			
<i>democratic debate</i>	0.05	0.21	0.12
<i>dem debate</i>	0.23	0.22	0.22
<i>#democraticdebate</i>	0.07	0.08	0.14
<i>#demdebate</i>	0.01	− 0.01	0.01
<i>Queries related to republican debate</i>			
<i>republican debate</i>	0.26	0.274	0.268
<i>rep debate</i>	− 0.09	− 0.09	− 0.09
<i>#republicandebate</i>	0.05	0.08	0.17
<i>#gopdebate</i>	− 0.06	− 0.06	− 0.02

Here a bias value closer to + 1.0 indicates Democratic bias and a value closer to − 1.0 indicates Republican bias

However, in case of some queries (e.g., *Martin O’Malley*, *John Kasich*, *democratic debate* and *#republicandebate*), the difference in the ranking bias values between Twitter’s ranking and the popularity based rankings indicates that there are probably other factors also that contribute to the overall bias of the search results. Note that this analysis is just a first step towards understanding the influence of the different factors on the overall bias of search results and we defer a more in-depth analysis for the future.



**Fig. 3** The time-averaged output bias *TOB* in Twitter “top” search results for the Republican candidates—candidates are listed left to right from highest to lowest popularity

**Table 9** Randomly selected tweets from the search results for the queries *Hillary Clinton* and *Donald Trump*, which are posted by a user with an opposite bias as compared to the candidate

Randomly selected tweets from <i>Hillary Clinton</i> search results, which are posted by a republican leaning user	Randomly selected tweets from <i>Donald Trump</i> search results, which are posted by a democratic leaning user
WT: Watchdog wants federal ethics probe of Clinton, possible improprieties <a href="http://bit.ly/1NvIrpA">http://bit.ly/1NvIrpA</a>	Williamsburg, #Brooklyn Dec 15 #trump2016 #MussoliniGrumpyCat #MakeAmericaHateAgain #DonaldTrump @realDonaldTrump pic.twitter.com/Hj6DC7M7V1
The Clintons both Bill and Hillary have a very long history of framing others while they commit the Crimes. History has destroyed the proof	Scotland defeats Trump on clean energy. Hopefully hell have a lot of time for golfing soon [url]
@CarlyFiorina: @realDonaldTrump is a big Christmas gift wrapped up under the tree for @HillaryClinton. [url]	Dirty little secret: Donald Trump is not a good debater.
s@CNN @HillaryClinton @BernieSanders hell no shes a murderer pic.twitter.com/zGQwR7dLZj	<a href="http://MLive.com">http://MLive.com</a> - Where Donald Trumps Michigan campaign donations come from <a href="http://ow.ly/39hCWt">http://ow.ly/39hCWt</a>
I dont care if youre a Democrat or Republican, how can you trust a word Hillary Clinton says and how can you consider voting for her??	Enjoy the sweet music of Donald Trump in Carol of the Trumps [url]

#### 4.4.2 Collective contribution of the input data and the ranking system

Having observed that both the input data and the ranking system contribute prominently to shape the final output bias seen by the users, we next explore the dynamics between these two sources of bias. Here, we discuss two cases in which the interplay between the input and ranking biases lead to an output bias which can noticeably affect a user’s search experience.

*The case of popular candidates:* Comparing the output biases for the candidate queries in Table 7, we found that the search results for the more popular candidates have a higher bias towards the opposing perspective.<sup>6</sup> For example, the top search results for the most popular Democratic candidate—*Hillary Clinton*—contained lesser Democratic-leaning results than other Democratic candidates, while the results for the most popular Republican candidate—*Donald Trump*—contained fewer Republican-leaning results as compared to other Republican candidates. In Fig. 3, we plot the output bias for the Republican candidates ranked by their popularity. The negative slope of the line of best fit the figure seems to suggest that the more popular a candidate is, the more is the opposing perspective in their top search results (however we are limited in the number of data points to be able to make any statistical inferences).

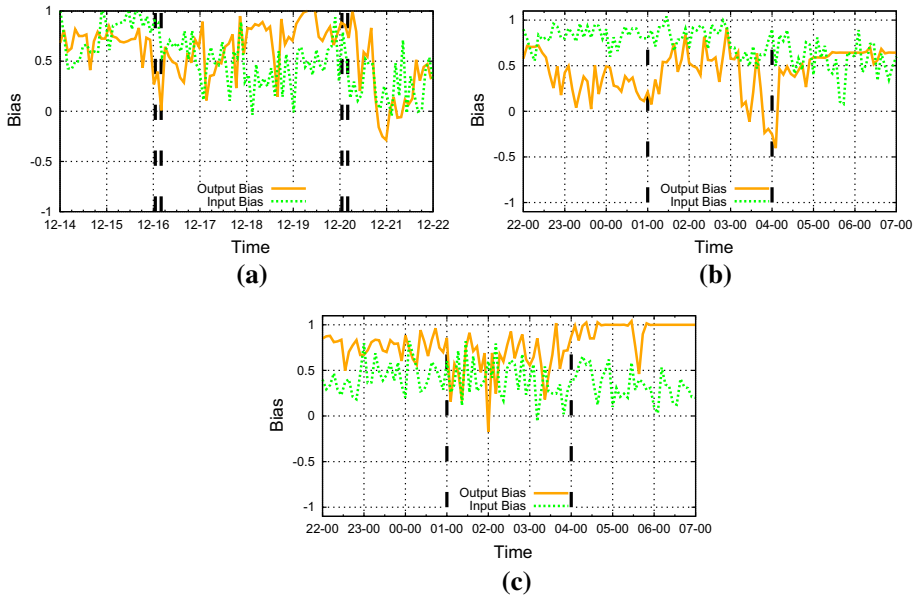
This situation may be undesirable for popular candidates, especially if users from the opposite perspective are more likely to speak negatively about the candidate and indeed this is what we find. Table 9 shows tweets randomly sampled from the set of tweets included in the top search results for a candidate, which were posted by users with an opposing polarity as compared to the candidate and they all either criticize or ridicule the candidates. Such negative tweets could alter the opinions of undecided voters (Epstein and Robertson 2015) and thus the situation is less than ideal for the popular candidates.

When we examine the input biases for *Hillary Clinton* and *Donald Trump*, we observe that they too lean towards opposite leaning indicating that opposite leaning users are more likely to talk about the popular candidates as opposed to less popular candidates. However, we observe that the ranking system altered input bias for the two most popular candidates in different manners—while the ranking system improved the situation for *Hillary Clinton* by adding a Democratic-leaning ranking bias and directing the search results towards her own party’s perspective, it does the opposite for *Donald Trump* by adding Democratic-leaning ranking bias and thus increasing the opposite leaning bias for him. These opposing interplay between the input data and ranking system (though possibly inadvertent) can have serious implications for the candidates, especially the one for whom the ranking system made the tweets of opposite leaning users more visible in the final output search results.

*Different phrasings of similar queries:* While looking for information about the same topic, different users may use different phrasings of the query. For instance, for searching for the event Republican debate, users can use different queries like *republican debate*, *rep debate*, *#republicandebate* or *#gopdebate*. If users from different leanings preferentially use different keywords, phrases or hashtags to refer to the same event in their tweets, then this might lead to differing biases for these differently phrased queries about the same event. To investigate whether different phrasings of the query about the same event lead to different biases, we compare the bias values for the queries related to Democratic and Republican debates, shown in Table 7. The first thing we observe is that the output biases for similar queries are noticeably different. For instance, the output bias of *republican debate* ( $TOB = 0.53$ ) has a lot more Democratic-leaning bias than the query *rep debate* ( $TOB = 0.31$ ), while the bias in search results for *#demdebate* ( $TOB = 0.57$ ) are much more Democratic-leaning than bias for the query *#democraticdebate* ( $TOB = 0.28$ ).

When we examine the input and ranking biases for our similarly phrased queries from Table 7, we observe that for most of them, the input bias is the more prominent contributor to the final output bias. However, in some cases even when the input biases are similar, as in the case of queries *rep debate* ( $TIB = 0.40$ ) and *republican debate* ( $TIB = 0.27$ ), the

<sup>6</sup> The popularity of a candidate is estimated from the polling data obtained from RealClearPolitics (2015) for December 2015.



**Fig. 4** Temporal variation of output and input bias for the query *dem debate*—(a) variation across the full duration over which we collected data (December 14–22, 2015), (b) variation during a 9-h window around the republican debate on December 15, 2015, (c) variation during a 9-h window around the democratic debate on December 19, 2015

**Table 10** Statistical analysis of temporal variation of output bias for the query *dem debate*—(i) variation in output bias across 3 h before and 3 h after the Republican debate on December 15, 2015, (ii) variation in output bias across 3 h before and 3 h after the Democratic debate on December 19, 2015, (iii) variation in output bias across the 3 h time periods during the Republican debate (on December 15, 2015) and the Democratic debate (on December 19, 2015)

Output bias across 3-h time periods	T1	T2	Paired <i>t</i> test	Effect size <i>r</i>
	Mean	Mean		
	(Std_dev)	(Std_dev)	<i>df</i>	<i>p</i> val
Before Rep debate (T1) versus after Rep debate (T2)	0.3783 (0.2025)	0.5189 (0.2415)	35	0.0380 – 0.3008
Before Dem debate (T1) versus after Dem debate (T2)	0.7675 (0.1133)	0.9576 (0.1164)	35	0.0000 – 0.6375
During Rep debate (T1) versus during Dem debate (T2)	0.4200 (0.2974)	0.6089 (0.2565)	35	0.0034 – 0.3219

ranking system shifts their biases in opposite directions, by adding a Democratic-leaning ranking bias for *republican debate* ( $TRB = 0.26$ ), while a Republican-leaning ranking bias for *rep debate* ( $TRB = -0.09$ ). This example illustrates the power the ranking system exerts on the input data, which can lead to search results for similar queries with similar input biases having different output biases. These observations about different biases for



similar queries raise questions about the impact that features like autocomplete queries and suggested queries can have on the bias that the users see, and what mechanisms can be designed to make the users aware of these effects. These are open research questions that can be pursued in the future and in Sect. 8.4 we briefly discuss some solutions for signaling the bias in the search results to the users.

#### 4.4.3 Variation of bias over time

Finally, we explore whether the bias in the search results for a particular query varies with the time at which the query is issued. As described earlier, we collected the Twitter top search results for our selected queries at 10-min intervals during the period December 14–21, 2015, which included both a Republican debate (December 15) and a Democratic debate (December 19).

To illustrate how the bias in the search results for a query varies with time, Fig. 4 shows the variation in the output and input biases for the query *dem debate* during the entire one week period (Fig. 4a), during a 9-h interval around the Republican debate (Fig. 4b), and during a 9-h interval around the Democratic debate (Fig. 4c). We observe noticeable variation in the bias over time. The variation is lower for the input bias because we compute input bias over cumulative sets of tweets and hence it is less affected by instantaneous events. However, the variation in output bias is much higher, especially during and immediately after the debate events. (Fig. 4b, c). In Table 10, we present the statistical analysis for the temporal variations in the output bias for the query *dem debate* (corresponding to Fig. 4). The first row of Table 10 shows the comparison between the output bias values for the search snapshots for the query *dem debate* in the 3 h period before the start of the Republican debate and the 3 h period after the end of the Republican debate. For comparison, we computed the significance of difference by performing paired *t* test and determining the *p* value for 95% confidence interval, and also computed the value of effect size *r*. Similarly, the second row in the table shows these values for the 3 h period before and after the Democratic debate. For both the debates, we find that the differences in output bias before and after the debate are statistically significant with medium to large effect sizes. We observed similar statistically significant temporal differences for other debate-related queries too.

We also observe another common trend in the variation of bias across different queries. The output bias for most debate related queries shifted down (towards the Republican perspective) during the Republican debate when possibly a larger number of influential or popular Republican users were actively posting on Twitter. Correspondingly, the output bias for most debate related queries shifted up (towards the Democratic perspective) during the Democratic debate. This trend is visible in Fig. 4b, c for the query *dem debate*, and we observed similar trends for most other debate-related queries. The third row in Table 10, shows the comparison between the output bias values for the search snapshots for the query *dem debate* during the 3 h period during the Republican debate and the 3 h period during the Democratic debate. Again, we observe that difference between the two is statistically significant (with medium effect size) with the output bias being lower (more Republican-leaning) during Republican debate than during the Democratic debate. Therefore, we find that which perspective is reflected more in the top Twitter search results varies with the time at which the query is issued.

## 5 Comparing relative bias in political searches on the web and social media

Next, we apply our bias quantification framework to compare the relative biases of political searches on two different search systems—Twitter social media search and Google Web search. This second study highlights another useful application scenario for our bias quantification framework where we can observe the output search results, but we do not have access to the input data to the ranking system (as is the case with most commercial search systems). This unavailability of input data makes it infeasible to disentangle the effect of input data and ranking system by measuring input bias and ranking bias separately, however, we can still compare the relative biases of different search systems.

Our choice of the two search systems to compare (Google and Twitter search) was driven by the fact that these are two popular channels by which internet users are finding news and information on the Web. Traditional media channels like Fox News or CNN have often been scrutinized by academics (Ribeiro et al. 2015; Babaei et al. 2018; Budak et al. 2016; Gentzkow and Shapiro 2010; Groseclose and Milyo 2005; Baron 2006; Munson et al. 2013b), as well as media watchdog groups (like FAIR ([fair.org](http://fair.org)) and AIM ([aim.org](http://aim.org))) for fairness, accuracy and balance in the news they report. Additionally, tools have also been developed to mitigate or expose the media bias (Purple Feed 2018; Park et al. 2009; Munson et al. 2013b; <https://twitter-app.mpi-sws.org/media-bias-monitor/>; <https://mediabiassfactcheck.com>) to users. However, the relative biases of newer digital channels like search systems are not as well studied and documented as yet, and thus users may not be taking their relative biases into account while selecting where to get their information from. With this study, we aim to highlight the differences in the bias of these two popular search systems—Twitter social media search and Google Web search. To have a fair comparison, we compare the Google search results with Twitter ‘news’ search results (<https://twitter.com/search-home>), both of which frequently contain results from news media sources.

### 5.1 Query selection and data collection

#### 5.1.1 Collecting Google web search data

We collected the top 20 Google search results for the queries stated in Sect. 4.1.1.<sup>7</sup> The results were collected at 10-min intervals during the period December 14–21, 2015, gathering a total of 714 distinct web-links across all the queries. As was done while collecting the Twitter search results, to minimize any personalization effects, all the Google search results were collected without logging in to Google, and from the same IP subnet in Germany.

Note that in the case of Web search, it is infeasible to gather the set of all relevant web-links for a query. Therefore we did not attempt to measure the input and ranking bias separately. Instead, we used the collected search snapshots to measure the bias in the output.

<sup>7</sup> We did not consider the hashtags as queries in this case, since hashtags are usually popular only on social media.

## 5.1.2 Collecting Twitter news search data

Following the methodology described in Sect. 4.1.2, we collected the first page of top 20 “news” search results for each query at 10-min intervals for the whole period. In total, across all the selected queries, Twitter news search results contained tweets posted by 7512 distinct accounts, an order of magnitude more than the number of distinct web-links in the dataset. We used these output search results to measure the output bias for Twitter news search.

## 5.2 Measuring political bias of a search result

For applying the bias quantification framework, we need a methodology for inferring the political bias score of each data item. Next, we describe how we measured the political bias of Google search results and Twitter news search results.

### 5.2.1 Measuring bias of Google search results

We observed that the top Google search results for our chosen set of queries (US presidential debates/candidates) contained a significant fraction of links from *news media websites* for which the political biases have been documented (Baron 2006; Gentzkow and Shapiro 2010; Groseclose and Milyo 2005; Munson et al. 2013b). We use the results from Balance study (Munson et al. 2013b) which identified the political bias of a large number of popular news media sources, to infer the political bias of the news media links in the web search results. We mapped the URLs in the search results to media sources in the Balance list (Munson et al. 2013a), by considering the longest matching substring.

Apart from links from news media sources, Google search results also frequently contain Wikipedia articles, and personal websites and social media accounts of the political candidates (as also observed in Trielli et al. 2015). We considered all Wikipedia URLs to have a zero or neutral bias,<sup>8</sup> all personal websites of the candidates to have their own leanings (e.g., *trump.com*, the website of Donald Trump, gets labelled as Republican), and all the social media profile links of the candidates to have their own leanings (e.g., the links to the Facebook, Twitter, Instagram accounts of Bernie Sanders are labelled as Democratic). Following this procedure, we were able to infer bias for 86% of the top Google search results on an average across all the queries. The rest of the domains, for which we did not attempt to infer bias, are mostly political facts websites (e.g., *ontheissues.org*, *bal-lotopedia.org*), informative websites (e.g., *biography.com*), or government websites (e.g., \*.gov pages).

### 5.2.2 Measuring bias of Twitter news search results

To have a fair comparison between Google and Twitter news search results, we switch our methodology to infer the political leaning of Twitter results in this section and utilize Balance scores (Munson et al. 2013a) for them too. We observed that the 7512 accounts which were included in the Twitter news search results include not only news media sources and

<sup>8</sup> Given Wikipedia’s policy of neutral point of view (<https://en.wikipedia.org/wiki/Wikipedia:Neutralpointofview>), we make this simplifying assumption. Though sometimes Wikipedia does contain misinformation, prior work Kumar et al. (2016) has shown that most hoaxes are quickly detected and have little impact on Wikipedia.

journalists, but also other users like politicians and even academicians; hence, there was no way to match all these accounts to Balance scores. Therefore, we ranked these accounts based on their frequency of occurrence in the Twitter news search results for all the queries and tried to manually map the top 200 accounts (which account for 63% of all the Twitter news search results) to Balance scores. Additionally, we attempted to match the 100 of the most influential media accounts on Twitter (Bremmen 2010) to Balance scores as well. Twitter news results also contained posts from journalists and political workers, and there was no way to map them to Balance score, so we manually labeled such accounts with their self-declared leaning from their profile bios (whenever available). Finally, as before, we marked the Twitter accounts of the presidential candidates with the candidate's own bias. By following this methodology, we were able to get the bias annotations for 155 media accounts on Twitter, which cover 45% of the Twitter news search results on average across the different queries.<sup>9</sup>

### 5.3 Comparing relative biases of Google search and Twitter news search

Our analysis shows three interesting ways in which the search bias for political queries on Google web search differs from that for Twitter social media search: (i) first, we investigated the temporal dynamics of the bias in the search results on the two systems and found the bias in social media search results to be significantly more dynamic across time, (ii) next, we compared their time-averaged output bias values to capture the overall trend and observed that for Google search the bias for most queries matches the leaning of the person or event being queried for, while the bias of Twitter news search for most queries is Democratic-leaning, and (iii) finally, we noticed that on Google search, a much higher fraction of search results are candidate-controlled sources (e.g., candidate's website or social media accounts), leading to more favorable results for the candidates on Web search than on social media search. Next, we elaborate on each of our findings about the differences in bias of Google Web search and Twitter social media search.

#### 5.3.1 Temporal variation in search bias

We began by comparing the two search systems along the temporal aspect by computing the standard deviation in the output biases of search result snapshots across time, for the different queries. We observe that the Google web search results are much more stable over time with a mean standard deviation of 0.046 in the output bias across all snapshots of all queries, while the standard deviation for Twitter news search results is an order of magnitude higher at 0.452, highlighting their highly dynamic nature in comparison.

#### 5.3.2 Higher democratic bias in Twitter news search results

Next, to compare the overall trend in relative biases of the two search systems, we computed the time-averaged output bias (*TOB*) for all the queries on Google and Twitter news

---

<sup>9</sup> The political leaning inferred by the source bias method and the Balance score based method match for 76% of these 155 media accounts. Here, we ignored the 9% of cases where our source bias methodology inferred the political leaning as neutral, which lead to a mismatch since the Balance Score does not output a neutral leaning.

**Table 11** Comparing time-averaged output bias *TOB* in (i) Google search results, (ii) Twitter news search results

Query	Google TOB	Twitter news TOB
<i>Queries related to events</i>		
<i>democratic debate</i>	− 0.039	0.271
<i>dem debate</i>	0.016	0.881
<i>republican debate</i>	− 0.224	0.216
<i>rep debate</i>	0.073	0.07
<i>Queries related to democratic candidates</i>		
<i>Hillary Clinton</i>	0.766	0.3
<i>Bernie Sanders</i>	0.577	0.42
<i>Martin O'Malley</i>	0.552	0.701
Average	0.631	0.473
<i>Queries related to republican candidates</i>		
<i>Donald Trump</i>	− 0.524	0.542
<i>Ted Cruz</i>	− 0.543	0.288
<i>Marco Rubio</i>	− 0.055	0.253
<i>Ben Carson</i>	− 0.259	0.191
<i>Chris Christie</i>	− 0.105	− 0.286
<i>Jeb Bush</i>	− 0.201	0.236
<i>Rand Paul</i>	− 0.642	− 0.006
<i>Carly Fiorina</i>	− 0.487	0.09
<i>John Kasich</i>	− 0.364	0.442
<i>Mike Huckabee</i>	0.006	0.058
<i>Rick Santorum</i>	− 0.229	− 0.041
<i>Lindsey Graham</i>	− 0.183	− 0.12
<i>George Pataki</i>	− 0.259	0.125
<i>Jim Gilmore</i>	0.138	− 0.608
Average	− 0.264	0.083

search, which are shown in Table 11. As can be observed from the table, there is a striking difference between the two—the *TOB* values for Twitter news search are positive (i.e., more Democratic-leaning) for most of the queries, including many of the Republican candidates, while the *TOB* values for the Google search results in most cases match the leaning of the candidate or event being searched for. So although the average *TOB* values for Democratic candidates are Democratic-leaning for both systems, the average output bias for Republican candidates is Republican-leaning ( $TOB = -0.264$ ) for Google, while it is on the positive side ( $TOB = 0.083$ ) for Twitter news search results.

This difference between Google and Twitter news search results may be due to the larger fraction of Democratic-leaning users on Twitter as indicated by the Democratic-leaning corpus bias we computed in Sect. 4, as well as the Democratic-leaning input bias *TIB* values for most queries reported in Table 7. These bias values mean that not only are there more Democratic-leaning users on Twitter, but the users tweeting about many of our queries are also Democratic-leaning. These results hint at the tremendous influence that corpus and input data have on determining the final output bias.

### 5.3.3 Favorable bias on Google search via candidate controlled sources

When we dug deeper, we found that another potential reason for the differences in the relative bias in Google search and Twitter news search results for a particular candidate is the difference in the *fraction of search results that come from sources controlled by the candidate themselves*. For the Google search results, a significant fraction—24.48% on average across all queries—of the results for the presidential candidates are from sources they control, i.e., either their personal websites or their social media profile links (e.g., for Donald Trump, we consider the webpage *trump.com* and his Twitter profile link <https://twitter.com/realDonaldTrump> to be sources controlled by him). A similar result is also reported in Trielli et al. (2015). This fraction is much smaller for most candidates on Twitter—across all the presidential candidates, only 7.14% of the Twitter news search results are from their own Twitter account. However, there are a few exceptions like *Martin O'Malley*, *Chris Christie* and *Jim Gilmore*, for whom 16.46%, 14.62% and 19.65% respectively of their Twitter news search results come from their own Twitter accounts. And correspondingly, the search results for these candidates show a strong bias towards their own perspective (as shown in Table 11). But, for most other candidates, the fractions of such tweets is much lower, and the bias in the Twitter news search results towards their own perspective is also lower.

Since sources other than the candidate's websites and social media profile links can also be controlled by the candidates, our measure likely underestimates the proportion of candidate controlled sources and thus provides a lower bound estimate of the observed favorable bias. In future, a more extensive analysis could be pursued using a larger set of possible candidate controlled sources.

The above observations about web search, including lower dynamicity over time and the candidates having favorable biases due to controlling a significant fraction of the links which come up in their top search results, make it easier for candidates to manipulate the Web search results in their own favor. While, the results on Twitter are much more dynamic and affected more by popular users on Twitter, rather than the candidates themselves, making them much harder to manipulate.

## 6 Comparing relative bias of Twitter's different ranking systems

In this paper, we measure the output bias of two different ranking systems of Twitter search—'top' and 'news' search filters—for the same set of queries. Since the input biases for the two are the same, we can compare the relative ranking biases for these two different ranking systems of Twitter. When we consider the average biases for the Republican candidates, we find that the input bias is slightly Republican-leaning (average  $TIB = 0.07$ , shown in Table 7), the Twitter 'top' search ranking system adds a Republican-leaning bias making the output bias Republican-leaning (average  $TOB = -0.11$ , shown in Table 7). While the Twitter 'news' search ranking system adds a little Democratic-leaning ranking bias making the output bias even more Democratic-leaning (average  $TOB = 0.083$ , shown in Table 11). This comparison of their relative ranking biases indicates that the 'news' filter of Twitter search highlights much more Democratic-leaning posts than the 'top' search filter.

## 7 Limitations

In this study, we focused on a limited set of queries that were either related to a political event or a political candidate. The main obstacles for expanding this set of queries included finding a set of queries that are not biased towards any party or candidate (as described in Sect. 4.1.1) and Twitter data collection limitations (API keys and infrastructure). Extending our query set to include more general political queries on polarizing topics like gun control or immigration could be done in the future to understand how the search systems are biasing the discourse about these popular debates in the society. Additionally, in the future, less popular queries which are less likely to be manually intervened could be analyzed to understand the influence of the ranking system.

Another limiting factor in our study was using the simplifying assumption of considering a user as either neutral, pro-Democrat or pro-Republican. Under this assumption we can not have a user who is partially both pro-Republican and pro-Democrat. However, we should clarify that for doing this classification, we still considered two scores for each user, one which captures the similarity to Republicans and the other to Democrats. Currently, to give the user a final leaning, we consider the difference between these similarities. However, in the future, we can use these two similarities to determine the extent to which a user is pro-Democrat as well as pro-Republican to have a more nuanced view of political leanings of users. Another interesting direction of future work would be to measure the users' opinions towards the different candidates in a more fine grained manner.

Also, the bipolar nature of US politics makes for a conducive environment for our bias measurement methodology. Extending our methodology for a multidimensional (political) space is likely to be quite challenging. Since our bias quantification framework can as easily work with a different methodology for inferring the bias of an individual item, future advances in measuring multidimensional bias could be plugged into our framework to quantify search bias for more nuanced and complex multidimensional bias search scenarios.

And lastly, while we discuss some potential solutions for signaling political bias in search results, and we have implemented our proposed split search as a Twitter application, however we have not done a user study to investigate the effect of this signaling on the users' search experience. This exploration is an important follow up of our current work.

## 8 Discussion

### 8.1 Generalizability of our search bias quantification framework

Having presented our results from applying our search bias quantification framework to measure the bias in political searches on Twitter social media search and Google Web search in the context of US politics, we now present a brief discussion of how our bias quantification framework can be generalized to scenarios of multiple perspectives, limited search data, and other search systems.

*Extending to multiple perspectives scenario:* In this paper, we have focused on US politics, and we have applied our bias quantification framework to this two-perspective scenario. However, it is possible to extend our framework to multiple perspective scenarios, for instance with  $p$  different perspectives. These  $p$  different perspectives could correspond

to the bias towards  $p$  different socio-political issues, or they could correspond to  $p$  different political parties. Our framework can be extended to  $p$  different perspectives, by associating a  $p$ -dimensional bias vector with each item, rather than a scalar bias score, as we did currently. More formally, the bias vector for the  $i$ -th data item would be given by  $V_i = [v_i^1, v_i^2, \dots, v_i^p]$ , where  $v_i^j$  gives a measure of how biased the  $i$ -th data item is along the  $j$ -th perspective, with values in the range of  $[-1, 1]$ . Here a value of  $v_i^j = 1$  could indicate support for the  $j$ -th perspective,  $v_i^j = -1$  could indicate opposition, whereas  $v_i^j = 0$  could indicate that the item is neutral with respect to that perspective. By converting Eqs. 1 to 4, to their vector addition formulations, we can measure the input, output and ranking biases for this  $p$ -dimensional scenario. The primary challenge for pursuing this direction in the future is the development of a methodology to capture these bias vectors.

*Extending to limited data availability scenario:* In many (if not most) cases, it may not be possible or feasible to either access or collect the input dataset of all items containing the selected queries. In such scenarios, we can adopt one of the following two approaches for applying our quantification framework for estimating the search bias:

1. Compare relative biases of two different search systems that function on similar input data: For many modern IR systems, the items in the corpus are directly ranked according to their relevance for a query, without explicitly extracting an intermediate relevant item set. For such systems, we can compute the relative ranking biases of the two systems assuming them to operate upon similar input sets. For instance, we could compare the relative ranking of different web search engines (e.g., Google vs. Bing vs. Yahoo), by observing the output bias for the same set of queries.
2. Approximate the input bias from the output search result snapshots: A simple approximation of the input bias based on the output search snapshot could be computed by taking an unweighted average of the bias scores of the items in the output set. This naive approximation can be improved by averaging over items in multiple search snapshots (e.g.,  $n$  search snapshots), or averaging over items in a larger snapshot with more search results (e.g., top-10k instead of top- $k$  results).

*Extending to other search systems:* Our bias quantification framework follows a black box approach and does not require the knowledge of the internal details of retrieval and ranking systems to quantify the search bias. As a result, it can be easily applied to study the bias of a wide range of search systems, as long as a methodology for computing the bias of an individual item (e.g., web-pages, tweets, posts) is available. Measuring the bias of an individual item in a search system is a context-dependent task, and since each platform is different, this in itself requires a significant effort. In this paper, we have delineated bias measurement techniques for tweets (Sect. 4) and web-links (Sect. 5). Also, in Sect. 2, we have briefly described prior work which has developed techniques for measuring the bias of users (Purver and Karolina 2015; Makazhanov and Rafiei 2013; Fang et al. 2015; Golbeck and Hansen 2011; Conover et al. 2011a, b; Pennacchiotti and Popescu 2011; Bond and Messing 2015; Wong et al. 2016) or content (Zafar et al. 2016; Weber et al. 2013) on social media as well as blogs and news stories (Adamic and Glance 2005; Yano et al. 2010; Zhou et al. 2011; Budak et al. 2016; Munson et al. 2013b) on the Web. In the future, when bias quantification schemes are developed for other search systems, for instance for videos (e.g., Youtube search) or music (e.g., Spotify), these methodologies can be plugged into our bias quantification framework and be used to analyze the bias of these other search systems.



## 8.2 Search bias and personalization

In this work, we have focused our attention on non-personalized search results, by adopting measures to mitigate the personalization effects as described earlier in the paper. We do acknowledge that in reality, most searches made by users are personalized. Therefore our results may not be representative of the searches mostly done in the wild. However, we believe that the personalization is most likely to exacerbate the biases we observe and report in this paper.

In the future, our bias quantification framework can be applied to study bias in personalized search scenarios as well. By performing carefully controlled experiments (Hannak et al. 2013; Kliman-Silver et al. 2015), along with our framework, the different sources of bias in personalized search scenarios can potentially be discerned. We leave the detailed design and implementation of such a study for the future.

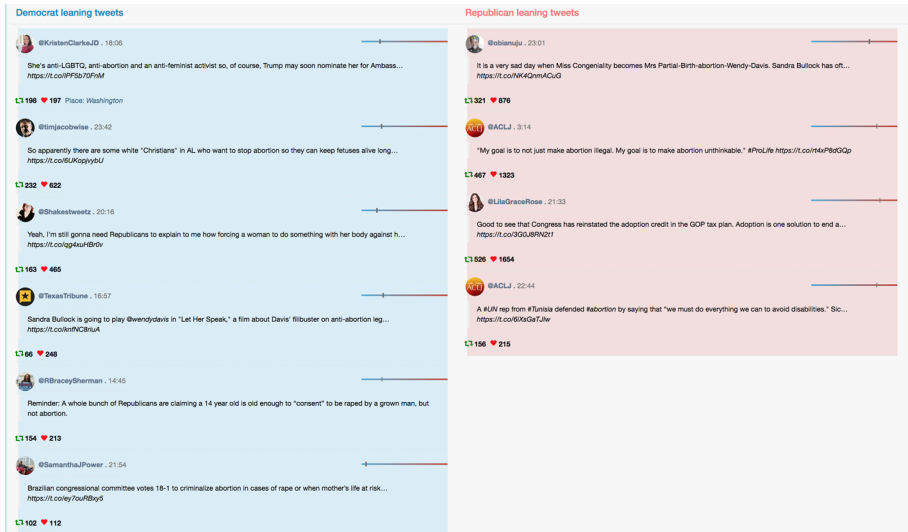
## 8.3 The black box approach

Recently, the rise of algorithmic platforms' influence on users' online experience has motivated many studies (Datta et al. 2015; Sweeney 2013; Hannak et al. 2014; Sandvig et al. 2014; Executive Office of the President 2016) to audit these platforms and understand their biases. While some of these algorithmic systems' functionalities are open to the public, making the auditing process easier, most of them are not. The walls of intellectual proprietary, high complexity of these algorithms and the perils of gaming a system via malicious users put these algorithms in a black box, making it almost infeasible to have access to an algorithm's specifications from outside, like in our study.

While we know about a few general factors that a search engine takes into account in curating the search results (such as relevancy, popularity, and recency), there are hundreds of other features that are hidden in a black-box, preventing us as researchers from being able to pinpoint the exact feature(s) of the algorithm which might be leading to the bias being introduced in the search results. However, it was possible for us (as outsiders) to observe the unranked set of items that contained a query (input data) and the ranked list of items output to the end user.

Therefore, building on previous studies that have adopted the "black-box" view for an algorithmic system while auditing it (Eslami et al. 2015; Liao et al. 2016; Hannak et al. 2014, 2013; Chen et al. 2015), we characterized the bias of the ranking algorithm in Twitter's search platform and Google Web search platform, without knowing their internal functioning. We assume a simplistic view of the search engine where in the first step we determine the set of items containing the query term, and in the next step, the black-box ranking system ranks these retrieved set of items into a ranked search output list, while taking into account all the relevance factors. Therefore we measure the input bias as the average bias of this set of input items that contain the query term, while we measure the bias of the output ranked list using a MAP-style score that weighs higher ranked items higher (thereby capturing the impact of relevance and other factors used for ranking).

In this paper, we report the bias observed in the search outputs. However, we do not claim that the search engines are intentionally adding bias to the search results because it is possible that the bias is introduced due to the numerous factors that the search system is using for ranking.



**Fig. 5** Screenshot of our Twitter-based split search service showing the results for the search term ‘abortion’. A widget adjacent to each result shows its bias measure

### 8.4 Signaling bias in search results

In this work, we have shown that both social media search, as well as Web search results, display varying degrees of bias. Next, we briefly discuss some solutions for tackling the bias, though their in-depth evaluation is left for the future.

*Designing bias-aware ranking systems:* A potential solution to address search bias is to design bias-aware ranking systems, which trade-off other metrics like relevance, popularity or recency with the bias of the search results. For instance, this could be achieved by minimizing the overall bias of search results by interleaving results with different biases using methods similar to the ones used for injecting diversity in results (Welch et al. 2011; Yom-Tov et al. 2013). However, this may lead to a degradation of the quality of search results along these relevance metrics, and finding an optimal trade-off point might be domain and user specific.

*Making bias transparent in search interface design:* An alternative method for addressing search bias could be to make the bias of each result transparent to the user by incorporating it into the search engine’s front-end design. Such a nudging practice has been used widely in the literature for purposes like delivering multiple aspects of news in social media (Park et al. 2009) and encouraging reading of diverse political opinions (Munson and Resnick 2010; Munson et al. 2013c). In a recent field study, it has been shown that by showing users alerts about the ranking bias in the search results can suppress the impact of the ranking bias on undecided voters’ voting preferences and also encourage them to read lower ranked results (Epstein et al. 2017b).

*Hybrid approach—split search:* A hybrid approach of the above two methods could also be proposed, which not only shows the bias of each search result, but also separates the results from the two political perspectives (Republican and Democratic) and shows them as *distinct ranked lists*, with each distinct list retaining the ranking of the results in the original ranked list. This solution can be particularly effective in the cases where re-designing

the algorithm and reaching a trade-off point for considering both bias and other relevance factors in an algorithm's design is infeasible. This method is similar to how several product companies like Amazon separate product reviews into positive and negative reviews, such that a user searching for that product can read the perspectives of others who either liked or disliked the product. By preserving the original search engine's ranking within each list, this methodology ensures that the quality of the top search results does not degrade across other metrics such as relevance, popularity, and recency.

We have deployed the proposed split search methodology as a live Twitter-based search service (<http://twitter-app.mpi-sws.org/search-bias-split-view/>) which allows users to log in with their Twitter credentials and do real-time searches for political queries on Twitter. The top search results are presented to the user as two distinct ranked lists containing Democratic- and Republican-leaning tweets, with each list maintaining the relevance rankings of the original search results returned by Twitter. Figure 5 shows a snapshot of the tool for the search term 'abortion'. Besides showing the bias of each search result, this split search design helps users to understand what fraction of the top results are related to each political leaning. For example, Fig. 5 shows that there are more Democratic-leaning search results for the query 'abortion' than Republican-leaning ones amongst the first page of top search results. Such differences can nudge users to notice which is the dominant political leaning for the top search results for a search query and encourage them to read more results from the other political side to gain more balanced information about a topic. A similar system has been developed by Wall Street Journal (Keegan 2017) which presents posts from the most biased news publishers on Facebook as chronological lists, with the aim of showing both sides of the stories. However, how users interact with such alternative search interface designs remains to be investigated and is left for future work.

## 9 Conclusion

To our knowledge, this work presents the first search bias quantification framework which not only quantifies the bias in the output search results but also discerns the contributions of two sources of bias—input data and ranking system. We have applied our framework to investigate the sources of bias for political searches on Twitter social media and found both input data and the ranking system to be prominent contributors of the final bias seen by the users in the output ranked list of search results. We found that factors such as the topic of the query, the phrasing of query and the time at which a query is issued also impact the bias seen by the users. We also applied our framework to compare the relative biases of Google Web search and Twitter social media search and found that Web search results are typically more favorable for the candidates from the two parties because many of the top results include links to candidate-controlled sources like their own or their party's websites and social media accounts.

While we do measure and report the bias introduced by the ranking systems of Twitter and Google search engines, we do not claim that these biases are intentionally added by the platform. In fact, we did not find evidence of any systemic bias, i.e., the platforms consistently ranking the items from one political leaning higher than the other, or consistently making the search results more polarizing by adding a Democratic-leaning bias to Democratic party related queries and Republican-leaning bias to Republican party related queries.

As an increasing number of people are relying on search systems to follow on-going events and news and public opinion on well known personalities (Teevan et al. 2011), the biases in search results can shape the users' opinions about these events and personalities (Pan et al. 2007; Epstein and Robertson 2015). Our work lays the groundwork for the design of new mechanisms for making the users more aware of search bias, for instance by making the potential biases in the search results transparent to the users. For users, this awareness can lead to more intelligent use of the system to mitigate the effects of search bias. For system designers, the search bias framework can be used to audit their systems, especially in cases when the bias is introduced by the ranking system and not the input data. And lastly, researchers and watchdog organizations can utilize our framework to audit and compare the biases of different search platforms, especially to unearth cases where the search bias may be ending up misleading the users.

**Acknowledgements** Open access funding provided by Max Planck Society.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## References

- Adamic, L. A., & Glance, N. (2005). The political blogosphere and the 2004 U.S. election: Divided they blog. In *Proceedings of the 3rd international workshop on link discovery* (pp. 36–43). ACM.
- Babaei, M., Kulshrestha, J., Chakraborty, A., Benevenuto, F., Gummadi, K. P., & Weller, A. (2018). Purple feed: Identifying high consensus news posts on social media. In *Proceedings of the AAAI/ACM conference on artificial intelligence, ethics & society*, AIES 2018, New Orleans, USA.
- Barocas, S., & Selbst, A. D. (2014). Big data's disparate impact. Available at SSRN 2477899.
- Baron, D. P. (2006). Persistent media bias. *Journal of Public Economics*, 90(1–2), 1–36.
- Bhattacharya, P., Ghosh, S., Kulshrestha, J., Mondal, M., Zafar, M. B., Ganguly, N., & Gummadi, K. P. (2014a). Deep twitter diving: Exploring topical groups in microblogs at scale. In *Proceedings of the 17th ACM conference on computer supported cooperative work and social computing, CSCW '14*, pp. 197–210. ACM, New York, NY, USA. <http://doi.acm.org/10.1145/2531602.2531636>
- Bhattacharya, P., Zafar, M. B., Ganguly, N., Ghosh, S., & Gummadi, K. P. (2014b). Inferring user interests in the twitter social network. In *Proceedings of the 8th ACM conference on recommender systems, RecSys '14*, pp. 357–360. ACM, New York, NY, USA. <https://doi.org/10.1145/2645710.2645765>.
- Bond, R., & Messing, S. (2015). Quantifying social media's political space: Estimating ideology from publicly revealed preferences on facebook. *American Political Science Review*, 109(01), 62–78.
- Bremmen, N. (2010). The 100 most influential news media Twitter accounts. <https://memeburn.com/2010/09/the-100-most-influential-news-media-twitter-accounts/>.
- Budac, C., Goel, S., & Rao, J. M. (2016). Fair and balanced? Quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, 80(S1), 250–271. <https://doi.org/10.1093/poq/nfw007>.
- Carney, D. R., Jost, J. T., Gosling, S. D., & Potter, J. (2008). The secret lives of liberals and conservatives: Personality profiles, interaction styles, and the things they leave behind. *Political Psychology*, 29(6), 807–840. <https://doi.org/10.1111/j.1467-9221.2008.00668.x>.
- Chen, L., Mislove, A., & Wilson, C. (2015). Peeking beneath the hood of UBER. In *Proceedings of the 2015 ACM conference on internet measurement conference* (pp. 495–508). ACM.
- Cohen, R., & Ruths, D. (2013). Classifying political orientation on twitter: It's not easy!. In *Proceedings of AAAI international conference on web & social media*, ICWSM 2013, Boston, USA.
- Coletto, M., Garimella, K., Gionis, A., & Lucchese, C. (2017). A motif-based approach for identifying controversy. [arXiv:1703.05053](https://arxiv.org/abs/1703.05053).
- Conover, M., Gonçalves, B., Ratkiewicz, J., Flammini, A., & Menczer, F. (2011a). Predicting the political alignment of twitter users. In *Proceedings of IEEE third international conference on social computing, SocialCom '11, IEEE*.

- Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., Flammini, A. (2011b). Political polarization on twitter. In *Proceedings of AAAI international conference on web & social media*, ICWSM 2011.
- Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. In *Proceedings on privacy enhancing technologies*, <http://arxiv.org/abs/1408.6491>.
- Edelman, B. (2010). Hard-coding bias in google “algorithmic” search results. <http://www.benedelman.org/hardcoding/>.
- 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 (PNAS)*, 112(33), E4512–E4521.
- Epstein, R., Robertson, R. E., Lazer, D., & Wilson, C. (2017a). Suppressing the search engine manipulation effect (SEME). *Proceedings ACM Human–Computer Interaction*, 1, 42:1–42:22. <https://doi.org/10.1145/3134677>.
- Epstein, R., Robertson, R. E., Lazer, D., & Wilson, C. (2017b). Suppressing the search engine manipulation effect (SEME). *Proceedings of the ACM: Human–Computer Interaction*, 1(2), 452.
- Eslami, M., Karahalios, K., Sandvig, C., Vaccaro, K., Rickman, A., Hamilton, K., & Kirlik, A. (2016). First i “like” it, then i hide it: Folk theories of social feeds. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 2371–2382). ACM.
- Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., Hamilton, K., & Sandvig, C. (2015). I always assumed that i wasn’t really that close to [her]: Reasoning about invisible algorithms in news feeds. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (pp. 153–162). ACM.
- Eslami, M., Vaccaro, K., Karahalios, K., & Hamilton, K. (2017). “Be careful; things can be worse than they appear”: Understanding biased algorithms and users’ behavior around them in rating platforms. In *Proceedings of AAAI international conference on web & social media*, ICWSM 2017 (pp. 62–71).
- Executive Office of the President, U. (2016). Big data: A report on algorithmic systems, opportunity, and civil rights. <http://tinyurl.com/Big-Data-White-House>.
- Fang, A., Ounis, I., Habel, P., Macdonald, C., & Limsopatham, N. (2015). Topic-centric classification of twitter user’s political orientation. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, SIGIR ’15 (pp. 791–794).
- Fortunato, S., Flammini, A., Menczer, F., & Vespignani, A. (2006). Topical interests and the mitigation of search engine bias. *Proceedings of the National Academy of Sciences (PNAS)*, 103(34), 12684–12689.
- Garimella, K., De Francisci Morales, G., Gionis, A., & Mathioudakis, M. (2016). Quantifying controversy in social media. In *Proceedings of the 9th ACM international conference on web search and data mining*, WSDM ’16.
- Gentzkow, M., & Shapiro, J. (2010). What drives media slant? Evidence from U.S. daily newspapers. *Econometrica*, 78(1), 35–71.
- Ghosh, S., Sharma, N., Benevenuto, F., Ganguly, N., & Gummadi, K. (2012). Cognos: Crowdsourcing search for topic experts in microblogs. In *Proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval*, SIGIR ’12 (pp. 575–590). ACM, New York, NY, USA. <https://doi.org/10.1145/2348283.2348361>.
- Golbeck, J., & Hansen, D. (2011). Computing political preference among Twitter followers. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1105–1108). ACM.
- Groseclose, T., & Milyo, J. (2005). A measure of media bias. *The Quarterly Journal of Economics*, 120(4), 1191–1237.
- Hannak, A., Sapiezynski, P., Molavi Kakhki, A., Krishnamurthy, B., Lazer, D., Mislove, A., & Wilson, C. (2013). Measuring personalization of web search. In *Proceedings of the 22nd international conference on world wide web*, WWW ’13 (pp. 527–538). ACM, New York, NY, USA. <https://doi.org/10.1145/2488388.2488435>.
- Hannak, A., Soeller, G., Lazer, D., Mislove, A., & Wilson, C. (2014). Measuring price discrimination and steering on e-commerce web sites. In *Proceedings of the 2014 conference on internet measurement conference*, IMC ’14, pp. 305–318. ACM, New York, NY, USA. <https://doi.org/10.1145/2663744>.
- Hern, A. (2015). Flickr faces complaints over ‘offensive’ auto-tagging for photos. <http://tinyurl.com/Flickr-AutoTagging>.

- Himelboim, I., McCreery, S., & Smith, M. (2013). Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on twitter. *Journal of Computer-Mediated Communication*, 18(2), 40–60.
- Keegan, J. (2017). Blue feed, red feed—see liberal facebook and conservative facebook, side by side. <http://graphics.wsj.com/blue-feed-red-feed/>.
- 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, IMC '15* (pp. 121–127). ACM, New York, NY, USA. <https://doi.org/10.1145/2815675.2815714>.
- Koutra, D., Bennett, P. N., & Horvitz, E. (2015). Events and controversies: Influences of a shocking news event on information seeking. In *Proceedings of the 24th international conference on world wide web, WWW '15, International world wide web conferences steering committee, Republic and Canton of Geneva, Switzerland* (pp. 614–624). <https://doi.org/10.1145/2736277.2741099>.
- Kulshrestha, J., Eslami, M., Messias, J., Zafar, M. B., Ghosh, S., Gummadi, K. P., & Karahalios, K. (2017). Quantifying search bias: Investigating sources of bias for political searches in social media. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing, CSCW '17* (pp. 417–432). ACM, New York, NY, USA. <https://doi.org/10.1145/2998181.2998321>.
- Kumar, S., West, R., Leskovec, J. (2016). Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th international conference on world wide web, WWW '16, International world wide web conferences steering committee, Republic and Canton of Geneva, Switzerland* (pp. 591–602).
- Liao, Q. V., Fu, W. T., & Strohmaier, M. (2016). # Snowden: Understanding biases introduced by behavioral differences of opinion groups on social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 3352–3363). ACM.
- Lichterman, J. (2010). New pew data: More Americans are getting news on Facebook and Twitter. <http://www.niemanlab.org/2015/07/new-pew-data-more-americans-are-getting-news-on-facebook-and-twitter/>.
- Liu, Z., & Weber, I. (2014). Is twitter a public sphere for online conflicts? A cross-ideological and cross-hierarchical look. In *International Conference on Social Informatics* (pp. 336–347). Springer.
- Lu, H., Caverlee, J., Niu, W., & Biaswatch, A. (2015). A lightweight system for discovering and tracking topic-sensitive opinion bias in social media. In *Proceedings of the 24th ACM international conference on information and knowledge management, CIKM '15*.
- Makazhanov, A., & Rafiei, D. (2013). Predicting political preference of twitter users. In *Proceedings of advances in social networks analysis and mining, 289–305, ASONAM '13*.
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge: Cambridge University Press.
- Media Bias / Fact Check - The most comprehensive media bias resource. <https://mediabiasfactcheck.com>.
- Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. (2013). Is the sample good enough? Comparing data from twitter's streaming API with twitter's firehose. In *International AAAI conference on web and social media, ICWSM '13, AAAI*.
- Mowshowitz, A., & Kawaguchi, A. (2005). Measuring search engine bias. *Information Processing and Management*, 41(5), 1193–1205.
- Munson, S., Chhabra, S., & Resnick, P. (2013a). BALANCE—Tools for improving your news reading experience—List of Classified sources. <http://balancestudy.org/whitelist-classifiable.html>.
- Munson, S., Chhabra, S., & Resnick, P. (2013b). BALANCE—Tools for improving your news reading experience. <http://balancestudy.org/>.
- Munson, S. A., Lee, S. Y., & Resnick, P. (2013c). Encouraging reading of diverse political viewpoints with a browser widget. In *International AAAI conference on web and social media, ICWSM '13, AAAI*.
- Munson, S. A., & Resnick, P. (2010). Presenting diverse political opinions: How and how much. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1457–1466). ACM.
- News Media Bias Monitor Discover the Demographics of News and Media Outlets. (2018). <https://twitter-app.mpi-sws.org/media-bias-monitor/>.
- Official Google blog: New ways to stay informed about presidential politics. <http://tinyurl.com/OfficialGoogleBlog>.
- 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, 801–823.

- Park, S., Kang, S., Chung, S., & Song, J. (2009). NewsCube: Delivering multiple aspects of news to mitigate media bias. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 443–452). ACM.
- Pennacchiotti, M., & Popescu, A. M. (2011). Democrats, republicans and starbucks aficionados: User classification in twitter. In *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '11* (pp. 430–438). ACM, New York, NY, USA. <https://doi.org/10.1145/2020408.2020477>.
- ProCon.org. (2015). Differences in conservative and liberal brains. <http://2012election.procon.org/view.resource.php?resourceID=004818>.
- Purple Feed. (2018). Identifying high consensus news posts on social media. <https://twitter-app.mpi-sws.org/purple-feed/>.
- Purver, M., & Karolina, S. (2015). Twitter language use reflects psychological differences between democrats and republicans. *PLoS ONE*, 10(9), e0137–e0422.
- RealClearPolitics—Election 2016—2016 Republican Presidential Nomination. (2015). <http://tinyurl.com/us-republican-polling-data>.
- Ribeiro, F. N., Lucas Henrique, F. B., Chakraborty, A., Kulshrestha, J., Babei, M., & Gummadi, K. P. (2015). Media bias monitor: Quantifying biases of social media news outlets at large-scale. In *Proceedings of the 12th international AAAI conference of web and social media, ICWSM '18*.
- 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.
- Semaan, B. C., Robertson, S. P., Douglas, S., & Maruyama, M. (2014). Social media supporting political deliberation across multiple public spheres: Towards depolarization. In *Proceedings of the 17th ACM conference on computer supported cooperative work and social computing* (pp. 1409–1421). ACM.
- Sharma, N. K., Ghosh, S., Benevenuto, F., Ganguly, N., & Gummadi, K. (2012). Inferring who-is-who in the twitter social network. *ACM SIGCOMM Computer Communication Review*, 42(4), 533–538. <https://doi.org/10.1145/2377677.2377782>.
- Shi, Y., Mast, K., Weber, I., Kellum, A., & Macy, M. (2017). Cultural fault lines and political polarization. In *Proceedings of the ACM conference on web science, WebSci '17* (pp. 213–217). ACM, New York, NY, USA (2017). <http://doi.acm.org/10.1145/3091478.3091520>.
- Smith, L. M., Zhu, L., Lerman, K., & Kozareva, Z. (2013). The role of social media in the discussion of controversial topics. In *2013 International conference on social computing (SocialCom)*, (pp. 236–243). IEEE.
- Springer, A., Hollis, V., & Steve, W. (2017). Dice in the black box: User experiences with an inscrutable algorithm. In *The AAAI 2017 Spring symposium on designing the user experience of machine learning systems*. AAAI.
- Sweeney, L. (2013). Discrimination in online ad delivery. *Queue*, 11(3), 10.
- Tavani, H. (2014). Search engines and ethics. In Zalta, E. N. (Ed.), *The Stanford encyclopedia of philosophy*, spring 2014 Edn.
- Teevan, J., Ramage, D., Morris, & M. R. (2011). #twittersearch: A comparison of microblog search and web search. In *Proceedings of the 4th ACM international conference on web search and data mining, WSDM '11* (pp. 35–44). ACM, New York, NY, USA. <https://doi.org/10.1145/1935826.1935842>.
- Trielli, D., Mussenden, S., & Diakopoulos, N. (2015). Why Google Search results favor democrats. <http://tinyurl.com/google-search-favor-dems>.
- Twitter Application: Make Users Aware of the Biases in Search. <http://twitter-app.mpi-sws.org/search-bias-split-view/>.
- Twitter Blog: Building a complete Tweet index. [https://blog.twitter.com/engineering/en\\_us/a/2014/building-a-complete-tweet-index.html](https://blog.twitter.com/engineering/en_us/a/2014/building-a-complete-tweet-index.html).
- Twitter Blog: Search Relevance Infrastructure at Twitter. [https://blog.twitter.com/engineering/en\\_us/topics/infrastructure/2016/search-relevance-infrastructure-at-twitter.html](https://blog.twitter.com/engineering/en_us/topics/infrastructure/2016/search-relevance-infrastructure-at-twitter.html).
- Twitter Help Center: Search result FAQs. <https://help.twitter.com/en/using-twitter/top-search-results-faqs>.
- Twitter Reaction to Events Often at Odds with Overall Public Opinion. (2013). <http://www.pewresearch.org/2013/03/04/twitter-reaction-to-events-often-at-odds-with-overall-public-opinion/>.
- Twitter Search Home. <https://twitter.com/search-home>.
- Van Couvering, E.(2010). Search engine bias: The structuration of traffic on the World-Wide Web. Ph.D. thesis, The London School of Economics and Political Science.
- Vaughan, L., & Thelwall, M. (2004). Search engine coverage bias: Evidence and possible causes. *Information Processing and Management*, 40(4), 693–707.

- Weber, I., Garimella, V. R. K., & Borra, E. (2012). Mining web query logs to analyze political issues. In *Proceedings of the 4th annual ACM web science conference, WebSci '12*, pp. 330–334. ACM, New York, NY, USA. <https://doi.org/10.1145/2380718.2380761>.
- Weber, I., Garimella, V. R. K., & Teka, A. (2013). Political hashtag trends. In *European conference on information retrieval, ECIR '13* (pp. 857–860). Berlin: Springer.
- Welch, M. J., Cho, J., & Olston, C. (2011). Search result diversity for informational queries. In *Proceedings of the 20th international conference on world wide web* (pp. 237–246). ACM.
- Wikipedia: Neutral point of view. [https://en.wikipedia.org/wiki/Wikipedia:Neutral\\_point\\_of\\_view](https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view).
- Wong, F. M. F., Tan, C. W., Sen, S., & Chiang, M. (2016). Quantifying political leaning from tweets, retweets, and retweeters. *IEEE Transactions on Knowledge and Data Engineering*, 28, 2158.
- Yano, T., Resnik, P., & Smith, N. A. (2010). Shedding (a thousand points of) light on biased language. In *Proceedings of NAACL HLT workshop on creating speech and language data with Amazon's mechanical turk (CSLDAMT)*.
- Yardi, S., & Boyd, D. (2010). Dynamic debates: An analysis of group polarization over time on twitter. *Bulletin of Science, Technology and Society*, 30(5), 316–327.
- Yilmaz, E., & Aslam, J. A. (2006). Estimating average precision with incomplete and imperfect judgments. In *Proceedings of the 15th ACM international conference on information and knowledge management* (pp. 102–111). ACM.
- Yom-Tov, E., Dumais, S., & Guo, Q. (2013). Promoting civil discourse through search engine diversity. *Social Science Computer Review*, 32, 145–154.
- Zafar, M. B., Gummadi, K. P., & Danescu-Niculescu-Mizil, C. (2016). Message impartiality in social media discussions. In *Proceedings in international AAAI conference on web and social media. ICWSM '16. AAAI*.
- Zhou, D. X., Resnick, P., & Mei, Q. (2011). Classifying the political leaning of news articles and users from user votes. In *Proceedings in international AAAI conference on web and social media. ICWSM '11. AAAI*.