Media Bias Monitor: Quantifying Biases of Social Media News Outlets at Large-Scale

Filipe N. Ribeiro[#o+], Lucas Henrique[o], Fabricio Benevenuto[o], Abhijnan Chakraborty[#*], Juhi Kulshrestha[#], Mahmoudreza Babaei[#], Krishna P. Gummadi[#]

[#]Max Planck Institute for Software Systems, Germany [+]Universidade Federal de Ouro Preto, Brazil
[*]Indian Institute of Technology Kharagpur, India [o]Universidade Federal de Minas Gerais, Brazil

Abstract

As Internet users increasingly rely on social media sites like Facebook and Twitter to receive news, they are faced with a bewildering number of news media choices. For example, thousands of Facebook pages today are registered and categorized as some form of news media outlets. Inferring the bias (or slant) of these media pages poses a difficult challenge for media watchdog organizations that traditionally rely on content analysis.

In this paper, we explore a novel scalable methodology to accurately infer the biases of thousands of news sources on social media sites like Facebook and Twitter. Our key idea is to utilize their advertiser interfaces, that offer detailed insights into the demographics of the news source's audience on the social media site. We show that the ideological (liberal or conservative) leaning of a news source can be accurately estimated by the extent to which liberals or conservatives are over-/under-represented among its audience. Additionally, we show how biases in a news source's audience demographics, along the lines of race, gender, age, national identity, and income, can be used to infer more fine-grained biases of the source, such as social vs. economic vs. nationalistic conservatism. Finally, we demonstrate the scalability of our approach by building and publicly deploying a system, called "Media Bias Monitor" [1], which makes the biases in audience demographics for over $20,000$ news outlets on Facebook transparent to any Internet user.

Introduction

Recent years have witnessed a radical change in the way news is being produced and consumed in our society. Online social media sites like Facebook and Twitter have emerged as popular destinations for users to receive, share, and discuss news about the world around them. A recent survey by Pew Research Center estimates that 62% of the U.S. adults consume news primarily from social media sites (Mitchell 2016), and this number is still growing. Similar to the traditional news media, the news stories disseminated over social media can also have a considerable impact on shaping people's opinions and influencing their choices, including having the potential to sway the outcomes of political elections (Allcott and Gentzkow 2017).

[1] twitter-app.mpi-sws.org/media-bias-monitor

A key characteristic of news on social media is that anyone can register as a news publisher without any upfront cost (*e.g.,* anyone can create a Facebook page claiming to be a newspaper or news media organization). Consequently, not only traditional news corporations are increasingly migrating to social media, but many social media only news outlets are also emerging (Lella 2016). With this recent transition, not surprisingly, there are growing concerns about 'fake' news publishers posting 'fake' news stories, and often disseminating them widely using 'fake' followers (Allcott and Gentzkow 2017; Vosoughi, Roy, and Aral 2018; Lazer et al. 2018).

Even when the accounts being used to publish or promote news stories are not 'fake bot' accounts (*i.e.,* they actually correspond to real persons or organizations), readers of news on social media are often not aware of the biases of these accounts. This situation is in sharp contrast to the news consumption over traditional news media channels, where because of the constant monitoring by media studies scholars and watchdog groups, at least well-informed consumers are aware of the biases of different news publishers.

For traditional media, two broad strategies have been used to quantify the biases of a given news outlet:

(i) The first strategy is to analyze the readership of the news outlets, which assumes that the content and attitudes of a news outlet end up driving the biases of its audience. Although this approach has been used by both researchers (Bakshy, Messing, and Adamic 2015; Gentzkow and Shapiro 2010; Zhou, Resnick, and Mei 2011) and think-tanks like Pew research (Mitchell et al. 2014), they often rely on readership surveys, and thus can not cover more than a few dozen mainstream news outlets.

(ii) The second class of approaches quantifies media bias directly by inspecting the published content (Covert and Wasburn 2007; Budak, Goel, and Rao 2016), specifically focusing on the coverage of important events by the media organizations. As there are significantly more news publishers on social media (with a constantly expanding list) than in the traditional media scenario, such strategies for measuring media bias do not scale for the current news ecosystem.
Thus, there is no mechanism available today for the users to know the biases of different publishers on social media.

In this work, we present a novel and scalable methodology to assess the biases of thousands of social media news

outlets. Facebook (as well as other social media sites) provides its advertisers access to its users through its targeted advertising platform. Before an ad is launched, and before any cost is incurred, Facebook exposes to the advertisers the size of the prospective audience that matches advertisers' targeting criteria composed of various dimensions like age, gender, political leaning, race, etc. Our *key idea* is to leverage the advertiser interfaces of social media sites that offer detailed insights into the demographics of the news source's audience on the social media site.

To that end, we design a crawler that exploits the Facebook marketing API to gather a large number of existing news outlets on Facebook. Then, we leverage the Facebook audience API to collect demographic information about the audience of these news outlets on Facebook. We hypothesize and empirically show that the ideological (liberal or conservative) leaning of a news source can be accurately estimated by the extent to which liberals or conservatives are over-/under-represented in the source's audience.

Finally, we demonstrate the scalability of our approach by building and publicly deploying a system called **Media Bias Monitor**[2], which quantifies the ideological biases of **20,448** news outlets in Facebook. Media Bias Monitor also provides demographic information along five other dimensions: gender, income level, racial affinity, national identity, and age. We hope that our system can bring more transparency to the biases of news publishers on social media, not only to the most popular ones, but also to small, niche news outlets.

The rest of the paper is organized as follows. The next section describes the related work. After which we present our strategies for gathering data from Facebook and discuss our method in detail. We then compare our approach for inferring ideological bias with four state-of-the-art methods, with the aim of validating our methodology. Next, we investigate other demographic aspects of news outlets with different ideological biases. We end with a brief discussion of our system design including examples of its application, before finally concluding the paper with the discussion about potential future research directions.

Related Work

Traditionally, news media organizations played an important role in societal evolution by acting as *gatekeepers of information*, and by deciding and regulating what news is consumed by the common people (Shoemaker, Vos, and Reese 2009). With this powerful role played by them, media studies researchers have long worried that an ideologically partisan and deregulated media can have a high impact on the political outcomes, and ultimately on our society (Groseclose and Milyo 2005; Chiang and Knight 2011). Therefore, a large number of research studies (as well as media watchdog groups like FAIR (`fair.org`) and AIM (`aim.org`)) have investigated *news media bias*, and evaluated the content produced by different news organizations for fairness, balance, and accuracy in news reporting.

Most of the efforts have focussed on studying political bias in traditional news media (Budak, Goel, and Rao 2016; Gentzkow and Shapiro 2010; Groseclose and Milyo 2005; Munson, Chhabra, and Resnick 2017). Particularly, Groseclose *et al.* (Groseclose and Milyo 2005) linked media sources to the members of the US Congress utilizing the co-citation of political thinktanks, and assigned them political bias scores based on the ADA scores of Congress members given by the political watchdog group 'Americans for Democratic Action' (`www.adaction.org`). Gentzkow *et al.* (Gentzkow and Shapiro 2010) inferred 'media slant' based on whether the language used by a media source is more similar to congressional Republicans or Democrats. Budak *et al.* (Budak, Goel, and Rao 2016) used a combination of crowdsourcing and machine-learning methods to study the selection and framing of political issues by different news organizations.

As online news sources are continuously gaining popularity, Munson *et al.* (Munson, Chhabra, and Resnick 2017) assigned political bias scores to popular news websites; whereas Babaei *et al.* (Babaei et al. 2018) proposed a system called "purple feed" to show users news which is likely to have high consensus between both republican and democrat leaning readers. In a recent work, Le *et al.* (Le, Shafiq, and Srinivasan 2017) presented a method to measure ideological slant of individual news articles by monitoring their consumption on Twitter. They analyzed the connectivity of the users tweeting an article to label them as republican or democrat leaning.

While political bias of news media has received a lot of attention, other forms of media biases have also been analyzed (*e.g.,* demographic bias (Chakraborty et al. 2017) such as gender (Shor et al. 2015) and racial biases (Ramasubramanian 2007)) to address concerns about these biases in news coverage, which can reinforce or even create certain forms of racial, gender, and ethnic stereotypes (Gilliam Jr et al. 1996). Similarly, efforts have been made to understand the topical coverage biases in news dissimination (Chakraborty et al. 2016) or recommendations (Bakshy, Messing, and Adamic 2015; Chakraborty et al. 2015), and whether they can lead to 'filter bubbles' (Pariser 2011). Being aware of such biases of different news media outlets is crucial for the society, since the awareness can play a critical role in shaping readers' assimilation of news published by these outlets (Dooling and Lachman 1971).

Overall, existing studies about the bias of news publishers have the following characteristics: (i) They infer bias based on either the content or active audience (*i.e.,* people sharing the news); and (ii) They are restricted to a small number of mainstream news publishers. In this work, we introduce a new approach to measure bias of news publishers based on their audiences (as inferred by Facebook), which allows us to study the bias of news outlets on a much larger scale. Additionally, our approach allows us to study the biases in the source's audience demographics along the lines of race, gender, age, national identity, and income, which can be used to infer more fine-grained leanings of news sources, such as social vs. economic vs. nationalistic conservatism.

Finally, a few efforts have explored the Facebook audience API, but with a focus on monitoring lifestyle diseases (Araujo et al. 2017), study worldwide gender inequal-

[2] `twitter-app.mpi-sws.org/media-bias-monitor`

ity (Garcia et al. 2017), and movement of migrants (Zagheni, Weber, and Gummadi 2017). Our work uses a similar strategy to gather demographic information from Facebook API, but to answer an orthogonal research question. We hope that our novel large-scale approach to measure ideological bias, as well as our system, will encourage a new research avenue of demographic studies related to news media.

Methodology

In this section, we describe how we identified news outlets on Facebook, and then gathered multiple demographic attributes of their audiences, which in turn enabled us to measure the biases of these news outlets.

Finding News outlets on Facebook

We start with a list of newspapers whose ideological biases we want to infer. To populate this list, we consider the news outlets used in the following prior efforts:

- 36 news outlets considered in the Pew Research survey on media habits (Mitchell et al. 2014).
- 15 news outlets from (Budak, Goel, and Rao 2016).
- 500 outlets used in (Bakshy, Messing, and Adamic 2015).
- 112 news outlets analyzed by the media bias monitoring website `AllSides.com`.[3]

To identify the Facebook Pages of these news outlets, we took the following three-pronged approach:
(i) First, we crawled the news media websites and searched for references to their corresponding Facebook pages. If we found such a reference, we fetched the name and URL of the referred page and compared with the name and URL of the newspaper to validate the mapping between the Facebook page and the media outlet.
(ii) If we did not get a match in the first step, we searched for the news domains (`nytimes.com`, `cnn.com`) using Facebook Graph API[4], and compared the name and URL in the returned Facebook page with the name and URL of the news media outlet.
(iii) If we did not succeed in establishing the mapping with the above steps, we searched for the news outlet's name using Facebook Graph API, and only included the pages where the names and URLs matched exactly.

After identifying the Facebook pages for these media outlets, we used the page names to search for their corresponding 'Interests' with Facebook Marketing API[5]. The API call returns a list of interests related to that name. If the interest name is identical to the Facebook page name, we link the Facebook page to the corresponding Interest ID. Such ID is key for our work as it allows us to gather the demographic information for the audience interested in the corresponding page. For example, the Interest ID of 'The New York Times' allows us to gather the demographics of the audience interested in 'The New York Times'.

In total, we were able to identify 32 news outlets (out of 36) from the Pew Research study (Mitchell et al. 2014), all 15 outlets from (Budak, Goel, and Rao 2016), 360 (out of 500) outlets from (Bakshy, Messing, and Adamic 2015), and 81 (out of 112) from `AllSides.com`.

As the above process of matching newspaper names using different APIs may result in errors, we conducted a manual validation for the mapping of news media sites to Facebook Pages, using a sample of 150 randomly selected outlets. We found the precision of the mapping to be 94.3% with 90% recall. Thus, we can conclude that using the steps described earlier, we could identify the Facebook pages belonging to different media outlets with high accuracy.

Gathering Audience Demographics

Facebook, similar to all large online social media, relies on advertisements for its revenue, and it provides advertisers with tools for highly targeted advertising. For example, before launching an advertisement, the advertiser can use the Audience API[6] to get the estimated audience (*i.e.,* number of monthly active users) likely to match the advertising criteria. In this work, we utilize this Audience API to gather the demographics of the audience of identified news outlets.

In brief, our approach consists of selecting an ad audience by specifying that the target population needs to have a certain attribute or a combination of attributes (this is the traditional way of targeting ads on Facebook (Speicher et al. 2018)), and then gathering the size of the targeted audience. Although targeting options are available for 197 countries worldwide, we focus only on the US-based Facebook users for this study. We plan to extend our effort to more countries in the future. For every identified Facebook page, we considered six demographic dimensions (*e.g.,* gender, race, *etc.*) and their corresponding attributes (*e.g.,* Male, Female, African-American, Hispanic, *etc.*), and computed the demographic composition of Facebook users interested in that page. Table 1 lists all six demographic dimensions we considered and their corresponding attributes.

Quantifying Ideological Biases

In this section, we first describe how we quantified the ideological bias of different news outlets. Then, to verify whether our inference strategy properly captures the news media bias, we compare our results with four very different approaches to measure media bias.

Measuring Bias using Facebook Audience Demographics

As detailed in the earlier section, utilizing the Facebook Audience API, we gathered the number of Facebook users, with different political leanings, interested in different media outlets. We use this strategy to measure the ideological bias score of an outlet. Specifically, we first find the fraction of users having different political leanings, and then multiply the fraction for each category with the following values: very liberal (-2), liberal (-1), moderate (0), conservative

[3] `allsides.com/media-bias/media-bias-ratings`
[4] `developers.facebook.com/docs/graph-api`
[5] `developers.facebook.com/docs/marketing-apis`
[6] developers.facebook.com/docs/marketing-api/audiences-api

Dimension	Attributes
Gender	Male, Female
Racial Affinity	African American, Asian American, Caucasian, Hispanic
Age	13-17, 18-24, 25-34, 35-44, 45-54, 55-64, above 65
National Identity	Australia, Africa, Canada, East Asia, Europe, Latin America, Mexico, Middle East, Russia, South Asia
Income Level	30-40K, 40-50K, 50-75K, 75-100K, 100-125K, 125-150K, 150-250K, 250-350K, 350-500K, >500K
Political Leaning	Very Conservative, Conservative, Moderate, Liberal, Very Liberal

Table 1: **Different demographic dimensions and attributes gathered from Facebook Audience API.**

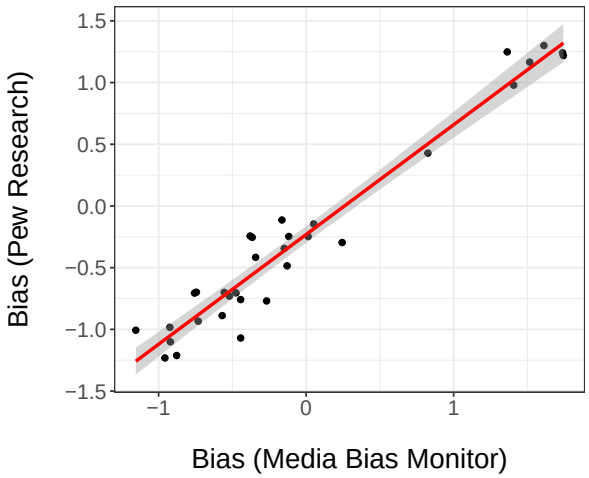


Figure 1: **Ideological leaning inferred by Media Bias Monitor in comparison with the bias inferred by the study from Pew Research (Mitchell et al. 2014).**

(1), and very conservative (2). The resultant sum gives the bias score which can vary from -2 to 2, where a high negative number indicates a highly liberal audience while a high positive number indicates a highly conservative audience for the media outlet. Accordingly, the media outlet is labeled as liberal leaning or conservative leaning. We utilized the above approach to quantify bias of different media outlets and built the system 'Media Bias Monitor' to make these biases more transparent to social media users (the details are presented in later sections). Next, we compare our approach with different approaches used to infer media bias.

Comparison with Survey Based Approach

We begin by comparing our approach with a study conducted by the Pew Research Center (Mitchell et al. 2014). Pew research classified the audience of popular news media outlets based on a ten question survey covering a range of issues like homosexuality, immigration, economic policy, and the role of government. In that study, the authors inferred the political leaning of the audience in a 5-point scale that are conceptually similar to those returned by Facebook Audience API – consistently liberal, mostly liberal, mixed, mostly conservative, and consistently conservative. In total, they evaluated 36 mainstream news media outlets, and we were able to gather the composition of their audience in Facebook for 32 of them.

To compare the bias inferred by Pew Research with ours, we compute the bias score from their data similar to how we compute the score for our method. For each category, we multiplied its fraction by its respective value in the scale ranging from -2 (consistently liberal) to 2 (consistently conservative). Figure 1 shows the scores obtained by our method for each news outlets along with the scores for the pew research study. Computing the Pearson's Correlation Coefficient (Lee Rodgers and Nicewander 1988) between the scores obtained by both methods, we found the Correlation Coefficient to be **0.97** (which is very high), with a 95% Confidence Interval of $[0.952, 0.986]$. This high correlation indicates that the results from both methods are almost perfectly matching.

Table 2 highlights the inferences from the two approaches for some popular news outlets. We can observe that both methods lead to the same conclusions about the political leaning of all these news outlets. Overall, the mean difference between the results of the two studies is 0.052 ± 0.016 for very liberal, 0.034 ± 0.012 for liberal, 0.070 ± 0.023 for moderate, 0.061 ± 0.022 for conservative, and 0.099 ± 0.034 for very conservative. We observe highest divergence for very conservative users, which can be explained by the possibility that number of conservative users may have grown in the US since 2014 (when the Pew Research study was conducted). We also note that in 26 out of the 32 media outlets, the number of moderate-leaning users decreased, which is also expected given the high polarization of the news discourse around the 2016 presidential election.

Comparison with News Sharing Approach

In (Bakshy, Messing, and Adamic 2015), the authors derived the alignment score of 500 media outlets by first identifying the political leaning of over 10 million Facebook users based on self-declarations, and then considering how users with different political leanings shared the stories published by these outlets. Similar to us, the authors measured the ideological leaning of the outlets on a scale ranging from -2 (Very Liberal) to $+2$ (Very Conservative). They identified the leaning of 500 news outlets, out of which we were able to find the Facebook pages (and thus identify the biases) for 342 outlets.

There are two reasons for not finding remaining outlets in Facebook: (i) we found that the domains of few outlets considered in their study (e.g., `dcbeacon.com`,

News Outlet	Source	V. Lib	Lib	Mod	Con	V. Con
NPR	Pew Res.	0.41	0.26	0.21	0.09	0.03
	Facebook	0.34	0.29	0.19	0.10	0.07
BBC	Pew R.	0.32	0.28	0.26	0.08	0.05
	Facebook	0.24	0.33	0.22	0.13	0.08
NYTimes	Pew R.	0.4	0.25	0.23	0.09	0.03
	Facebook	0.30	0.28	0.21	0.13	0.09
CNN	Pew Res.	0.19	0.25	0.4	0.12	0.04
	Facebook	0.23	0.27	0.22	0.15	0.12
Breitbart	Pew Res.	0.03	0.04	0.14	0.31	0.48
	Facebook	0.01	0.02	0.07	0.22	0.67
Fox News	Pew Res.	0.04	0.14	0.37	0.27	0.19
	Facebook	0.07	0.10	0.17	0.27	0.40

Table 2: **Pew Research results in comparison with our Facebook audience-based approach for measuring political leaning of different news media.**

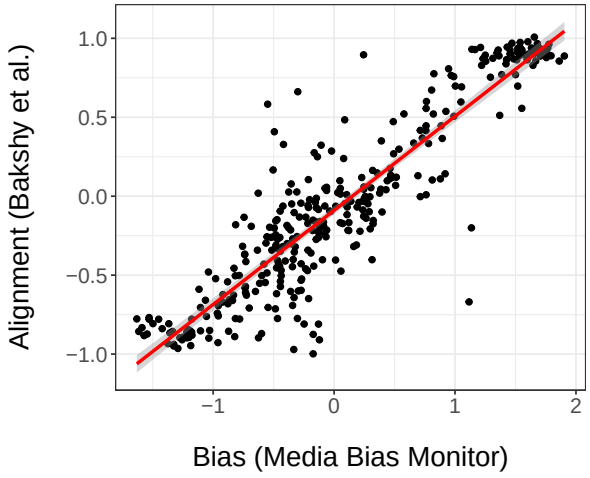


Figure 2: **Ideological leaning inferred by Media Bias Monitor in comparison with the bias inferred by (Bakshy, Messing, and Adamic 2015).**

scgnews.com etc.) are no longer active and hence could not be reached; (ii) we could not find the Facebook pages or Interest IDs for the remaining outlets, without which we can not gather the composition of Facebook users interested in those outlets.

Figure 2 shows the scatter plot of the scores obtained by two methods, where each dot in the figure is a news outlet and the scores of each method can be seen on the axes. Overall, the Pearson Correlation Coefficient for the scores obtained by our method and the method proposed by (Bakshy, Messing, and Adamic 2015) is **0.91**, with a 95% confidence interval of $[0.891, 0.927]$. Thus, we can note that despite the large number of news outlets considered, inferred ideological biases from both approaches are very close.

Comparison with Content Based Approach

Budak *et al.* (Budak, Goel, and Rao 2016) used a content-based approach to identify the slant of the top 13 U.S. news outlets and two popular political blogs. They sampled two

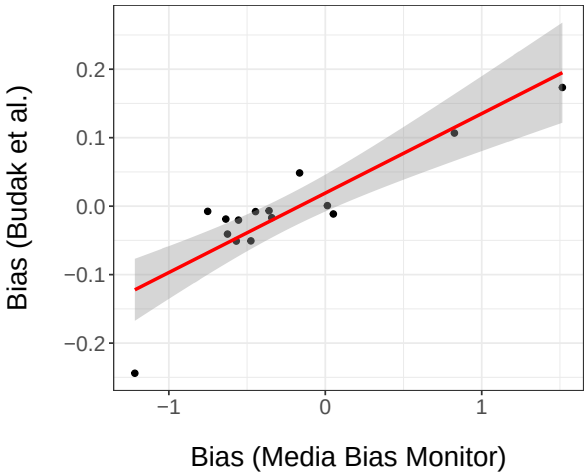


Figure 3: **Ideological leaning inferred by Media Bias Monitor in comparison with the bias inferred by (Budak, Goel, and Rao 2016).**

political stories per day for each outlet, and used Amazon Mechanical Turk platform[7] to ask human judges if the article was positive, negative or neutral towards Democrats or Republicans. The answer was encoded in separate 5-point scale with the values $\{-1, -0.5, 0, 0.5, 1\}$ for Democrats, and $\{1, 0.5, 0, -.0.5, -1\}$ for Republicans. Therefore, a negative average score implies the article is positive toward Democrats, while a positive average score indicates Republican leaning. Finally, the slant for each news outlet is calculated as an average of individual news' leaning scores.

Figure 3 shows the scatter plot between the bias scores obtained by us and by (Budak, Goel, and Rao 2016). Overall, the Pearson Correlation Coefficient between the scores obtained by these two methods is **0.87**, with a 95% confidence interval of $[0.650, 0.956]$. This implies that our approach inferred results similar to their content-based approach.

Comparison with Crowdsourcing Approach

Finally, we compare our approach with a crowdsourcing-based method to infer media bias deployed at the website AllSides.com. It encourages its users to rate different news outlets in one of the five categories: left, lean left, center, lean right, and right[8], using any of three different strategies: (i) blind surveys, in which users rate the bias of stories without knowing the news source; (ii) showing them the bias of the source as inferred by previous research efforts (e.g., the work by (Groseclose and Milyo 2005)), and (iii) showing them the past feedback from the other users. In (iii), a user can agree or disagree with the past ratings of the news outlets and can suggest new ones.

In total, AllSides.com presents bias of 112 media outlets, out of which we were able to identify the Facebook audiences for 81 outlets. Similar to the previous approaches, we defined a fixed bias score for each category assigned by

[7] mturk.amazon.com

[8] allsides.com/media-bias/media-bias-ratings

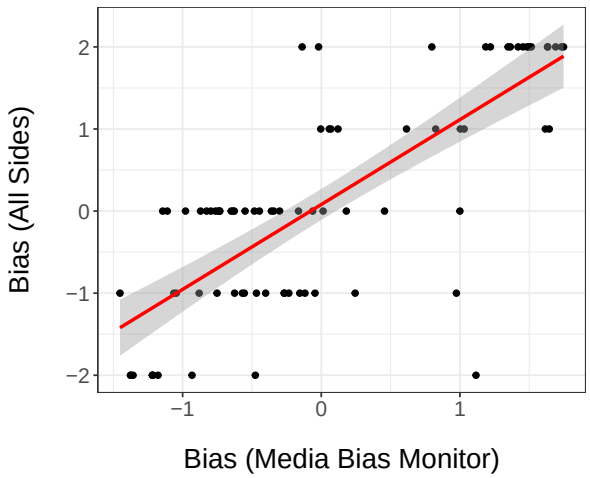


Figure 4: **Ideological leaning inferred by Media Bias Monitor in comparison with the bias inferred by** `Allsides.com`**.**

AllSides: left (-2), lean left(-1), center(0), lean right (1) and right (2). Figure 4 compares bias scores inferred by our approach vis-a-vis the scores from AllSides. The Pearson correlation coefficient obtained for the results of the two methods is **0.77** with a 95% confidence interval of $[0.658, 0.843]$.

Diving deeper into the mismatches, we observed that main difference occurs when AllSides marks a news outlet as center biased, whereas our approach assigns it a liberal bias score. Among those media outlets, we found BBC, CNN and Reuters, with the following liberal scores according to our approach: -0.521, -0.342, and -0.446 respectively. In order to verify the correctness of our score for these specific cases, we contrasted them with the results from (Bakshy, Messing, and Adamic 2015) and from (Budak, Goel, and Rao 2016). Both methods assigned liberal scores to these news outlets. Additionally, we verified that media outlets like Al Jazeera, FiveThirtyEight, and NPR have strong liberal bias according to the method by (Bakshy, Messing, and Adamic 2015) as well as our method, whereas AllSides marked them as moderate.

Summary

In summary, Table 3 presents the Pearson Correlation Coefficients (PCC) for the comparison between our approach with four existing state-of-the-art methods that use very different inference methods. Our method closely matches most of these studies, thus, validating our methodology to obtain the ideological bias. In the next section, we discuss two key benefits of using our approach over the existing ones.

Media Bias Monitor

In the previous section, we showed that our approach to quantify media bias can produce inferences similar to four very different state-of-the-art methods. However, the key advantage of our method over the existing approaches is that our approach is highly scalable, and can infer the ideological bias of several thousands of news media outlets that exist today. As a show case, we built a system, named *Media Bias Monitor* [9], which makes the biases in audience demographics for **20,448** news outlets in Facebook transparent to users. The number of news outlets we cover are at least two orders of magnitude more than any existing efforts.

Method	Total	Identified	PCC	CI (95%)
Pew Research	32	36	0.97	[0.946,0.987]
Bakshy *et al.*	500	342	0.91	[0.891,0.927]
Budak *et al.*	15	15	0.87	[0.650, 0.956]
AllSides.com	112	81	0.77	[0.658, 0.843]

Table 3: **Summary of the comparison between our approach to infer ideological bias and four previous efforts.**

Scaling Bias Inference

We begin by describing how we identified thousands of news outlets in Facebook, and then show the importance of identifying biases of these outlets.

Finding Large Number of News Outlets on Facebook
Our first step to identify a large set of news outlets on Facebook consisted of identifying large lists of news outlets on the Web. We used three different sources from the Web to identify names and web domains of known news outlets:
(i) list of news outlets included in Google News (also used in prior studies such as (Leskovec, Backstrom, and Kleinberg 2009)), (ii) $3,000$ most popular newspaper domains as determined by Alexa[10], and (iii) list of newspapers curated by a website[11]. We then combined the news outlets present in these lists and extracted an aggregated list of news outlet names and their corresponding website URLs.

To identify the Facebook Pages of these news outlets, we used the procedure described earlier to find news outlets on Facebook. After identifying the Facebook pages for these media outlets and subsequently their Interest IDs using Facebook Marketing API, we obtained a dataset of $2,466$ news outlets. Although this increased dataset size in one order of magnitude, to cover more outlets, we designed a new strategy using another Facebook API call.

Given the Interest ID of a Facebook page, the Facebook Marketing API also suggests a number of Facebook pages which are related to it[12]. We designed a Breadth-First Search (BFS) scheme which recursively collects the suggestions, starting from each Interest ID we had previously collected as a seed list. Our crawler exhausted the entire component of a graph in which nodes are Interest IDs and edges are suggestions. In total, we gathered near $240K$ related Interests from different categories (as defined by Facebook). We considered only those pages whose categories are related

[9] `twitter-app.mpi-sws.org/media-bias-monitor`
[10] `alexa.com/topsites/category/Top/News/Newspapers`
[11] `www.listofnewspapers.com`
[12] Technically, the Marketing API returns a list of related Interest IDs, and we run a reverse Interest Id – Facebook Page mapping. We omit the details for brevity.

Category	(#)	Example News outlets
Magazine	2,565	In Touch Weekly, Country Living, UNILAD
Newspaper	1,099	The Washington Post, The Daily Caller
Journalist	750	Bill O'Reilly, Lester Holt, Megyn Kelly
News Company	1,346	BuzzFeed Food, Conservative daily
Website	3,687	Topix, GroupMe, Delish
Other	900	BuzzFeed, Yahoo! News
Radio Station	992	2Day FM, Radio One Lebanon, Radio Disney
TV Show	4,447	NBC today show, The voice
Sports Team	2,615	Dallas Cowboys, Pittsburgh Steelers
TV Channel	2,047	ABC, CBS Sports

Table 4: **Number of news outlets in different categories covered by Media Bias Monitor.**

to News and Media (*e.g.,* 'Newspaper', 'Media/News Company', 'News & Media Website', 'Journalist', 'Magazine', 'Broadcasting & Media Production Company', 'Website', 'Publisher' *etc.*). After filtering out other categories, our final dataset of news outlets contains **20, 448** Facebook pages and their corresponding Interest IDs. Then, we gathered the demographics of the audiences of all these outlets by following the procedure detailed in the Methodology section.

Table 4 shows the number of news outlets in each category, as well as a few examples that help us to understand what kind of news outlets are grouped into each of these categories. The most popular category is TV show, which contains TV news programs such as NBC's Today Show. We note that there are also TV shows from outside US, but they have large following among the US users. The second most popular category corresponds to external websites, followed by sport teams' news and magazines. We also observe smaller yet considerable fractions of radio stations, newspapers, and individual journalists. Interestingly, although the number of individual journalists is small in comparison to other categories, some journalists have quite large audiences. For example, Bill O'Reilly, with an audience of 2.2 million users is on top of the list, followed by Lester Holt (1.8M) and Megyn Kelly (1.3M).

Importance of Measuring Bias at Scale

Figure 5 shows the audience size of all news outlets gathered using the above steps. We can observe from Figure 5 that although a small number of most popular news outlets on Facebook reach large number of news readers, still a large number of news outlets cater to small niche audiences, which account for a non-negligible fraction of the overall news audience.

Interestingly, we find that the news outlets with fewer audiences are also those that are most ideologically biased. For example, among the 10-percent most biased (i.e., either most conservative or most liberal) news outlets, 58% outlets have audience size less than $10,000$ users, whereas, among the 10-percent least biased outlets, only 31% outlets have less than $10,000$ audience. This suggest that the most biased news outlets are usually those that reach niche and smaller audience, thereby highlighting the importance of monitoring the news published by these outlets and not only those published by the mainstream news publishers.

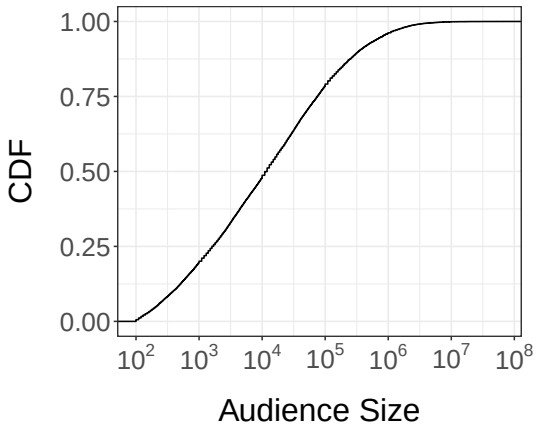


Figure 5: **Distribution of audience size of news outlets.**

Quantifying Biases at Finer Granularity

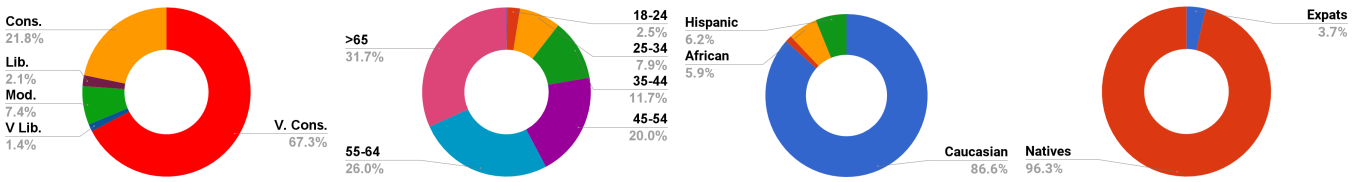
Most of the prior works on news media bias restricted themselves to political bias, leaving out other dimensions (e.g., racial bias, gender bias or age bias) that can be very helpful in providing a more fine-grained perspective of the complex news ecosystem. One might wonder, for instance, whether a highly conservative media outlet mostly has a young Black audience, or does their audience have a high prevalence of old Caucasian people. Next, we briefly discuss, through a series of examples, the benefits of incorporating the measurement of other demographic attributes as part of our system.

Breakdown of Demographic Dimensions

A key feature we have incorporated in Media Bias Monitor is a breakdown of the audience across different demographic attributes. For example, Figure 6 shows the breakdown of four demographic dimensions for Breitbart, a well-known conservative media outlet. As a reference for comparison, Table 5 shows the distribution for these demographic attributes for all Facebook users in the US.

As can be noted in Figure 6(a), number of conservative and very conservative users constitute more than 89% of the Breitbart audience. Figures 6(b), (c), and (d) present the breakdown of Breitbart audience along age, racial affinity and national identity. These figures show that the audience of Breitbart consists of 96.3% US natives, 86.6% Caucasians, and 57.7% of users older than 55 years. These values are much higher compared to the Facebook population in the US. Additionally, the proportion of men among Breitbart audience is 55% compared to 46% in the Facebook population in the U.S. However, in terms of Income Level, the distribution is quite similar to the overall Facebook population.

Observing the demographic dimensions for The Economist (see Figure 7), a liberal biased outlet (more than 65% interested users are Liberal and Very Liberal), we find that it has a higher fraction (21%) of well-paid audience, earning more than \$150K, against 14.1% in the US-based Facebook population. Men and expats are also higher compared to the overall population, while in terms of age and racial affinities, we don't see much difference.



(a) Political Leaning (b) Age Groups (c) Racial Affinities (d) National Identity

Figure 6: **Breitbart and its bias across four demographic dimensions.**

Gender	**Male**	46%
	Female	54%
Racial Affinities	**African American**	16.1%
	Asian American	3.5%
	Caucasian	64.3%
	Hispanic	16.1%
Income	**30k to 50k**	17.9%
	50k to 75k	35.6%
	75k to 150k	32.4%
	over 150k	14.1%
Age Groups	**under 18**	2.4%
	18-24	16.3%
	25-34	25.3%
	35-44	18.8%
	45-54	15.4%
	55-64	11.9%
	above 65	9.9%
National Identity	**Natives**	84.3%
	Expats	15.7%

Table 5: **The composition of the US-based Facebook users along different demographic dimensions and their corresponding attributes.**

Taking a closer look at other media outlets, we find other examples in which conservative news outlets have audiences that are over-represented by Men - Drudge Report (63%) and Rush Limbaugh Show (60%); by older people (aged above 55) - Rush Limbaugh Show (51%) and Sean Hannity Show (67%); and by Native Americans - The Blaze (97%). On the liberal side, we see an over-representations of women - ABC News (70%) and BuzzFeed (71%); African American - Daily Show (23%) and Al Jazeera America (35%); older people (aged above 55) - Politico (38%) and PBS (33%); younger people (aged between 25-34) - Daily Show (34%).

Apart from these well known media outlets, we can expand our analysis to a wider range of outlets. For example, we found a set of conservative media outlets that are more biased towards men. Such outlets include publishers of news related to Guns (e.g., Guns.com (92%), FourGuysGuns (94%)), or containing military news (e.g., The Fire Critic (83%), publishing stories from Fire Service, and SOFREP.com (92%), with news written and curated by former CIA and Veterans). Conservative women, in turn, have interest in media outlets publishing religious articles (e.g., PrayAmerica (82%), Breaking Christian News (71%)), or health related stories (e.g., Lifenews.com (76%), a website that post stories with topics against abortion and euthanasia). On the other hand, Liberal outlets with major predominance of men include gay magazines (like Gay Times (84%) and Instinct (88%)) and a left-wing magazine (Jacobin - (66%)). Liberal women, are over-represented in the audience of feminist and liberal magazines: The Man Repeller (96%), Feministing.com (91%) and Everyday Feminism (90%).

These examples show that our bias inference approach and the deployed system allows one to get a deeper understanding of bias in different media outlets. It not only presents the political bias of a large number of news outlets, but it also provides an interesting way of understanding other intrinsic biases (i.e., gender bias, age bias, etc.) of the audience interested in a certain news outlet.

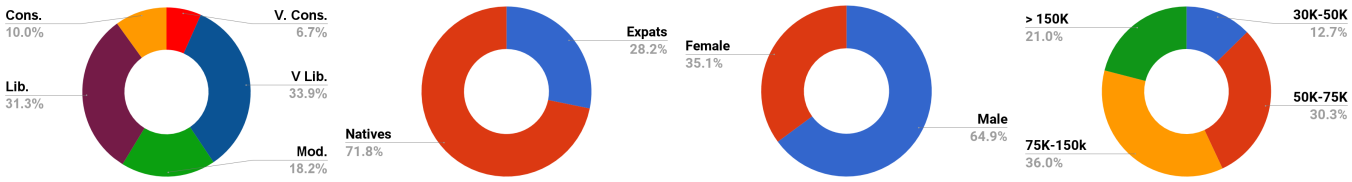
Search and Ranking Functions

As a final contribution, the system 'Media Bias Monitor' consists of a search function, which allows users to search for news outlets by name, as well as a ranking function that allows the users to find news outlets that have most over-representation of audience belonging to a particular demographic group. For instance, a user belonging to a particular demographic group can look up what news sources other fellow group members are subscribing to. First two rows in Table 6 show news outlets highly gender biased towards either men or women. We notice that many sports specific outlets are highly biased towards men; whereas, outlets related to fashion, makeup, and pregnancy tend to be the ones most biased towards women.

Similarly, the most racially biased outlets (third and fourth rows in Table 6) in terms of African-American and Asian-American users are clearly focused on these specific racial demographic groups. In terms of high age bias (as presented in Table 6), for under 18 years, we can note TV channels like Disney which target adolescents, as well as outlets related to games. For 18 to 24 years age group, we find many news outlets associated with dating, music, TV series, TV shows, and games. Interestingly, the outlets most biased towards the 25 to 34 years old users are associated with business, professions, and job seeking. For age groups higher than that, the most biased news outlets are related to parenting and family. Finally, the last two rows in Table 6 show very conservative and very liberal news outlets, which are all focused on politics, with their political leaning being expressed via their names themselves.

Demographic Dimension	Demographic Attributes	Sample of the Most Overrepresented News Outlets
Gender	**Male**	Velocity RC Magazine(99%), myGayTrip.com(99%), Best Motoring(99%), The Gentleman's Journal(99%)
	Female	Styletoday(100%), Makeuptalk.com(99%), Pregnancy and newborn(99%), Proud single Moms(98%)
Racial Affinities	**African American**	BlackamericaWeb(92%), BlackNews.com(89%), Black Men Magaz.(80%)
	Asian American	Hoa hoc Tro Magazine(97%), Kenh14.vn(97%), Sportsoho(100%)
Age Groups	**Under 18**	Fox Action Movies(25%), Disney Channel(23%), MuchGames.com (24%), BeingGirl(26%)
	18-24	Disaster Date(68%), Fairy Tail Fans(65%), Insert Gamer(76%), Speed and Sound Magazine(66%)
	25-34	JobTopGun(70%), Canadian Business(72%), Marketing na Cozinha(65%), WeddingSutra(76%)
	35-44	Fans of Being a Mom(50%), Scholastic Parents(47%), Growing Without Schooling(44%)
	45-54	Rush is a Band (59%), Ultimate Classic Rock(38%), Yahoo! Sports Radio(58%)
	55-64	SmartMoney(61%), The new avengers(42%), The Monkees(38%), I Love Being a Grandma(37%)
Political Leaning	**Very Cons.**	Legal Insurrection(90%), RedState(84%), Patriot Update(84%), Conservative Angle(82%), Fox Nation(75%)
	Very Liberal	Sister 2 Sister(76%), The Alaska Quarterly Review(66%),Democracy Now(59%)

Table 6: **Examples of highly biased news outlets in Facebook along different demographic dimensions. The percentage of audience belonging to the respective demographic groups are shown in parenthesis.**



(a) Political Leaning (b) National Identity (c) Gender (d) Income Level

Figure 7: **The Economist and its bias across four demographic dimensions.**

Conclusion

In this work, we proposed a novel methodology to quantify the ideological biases of thousands of news outlets on social media. To do so, we utilized the leaning of their audience which can be obtained from the social media site's advertising framework. Specifically, for this work, we collected $20,448$ pages categorized as news by Facebook, and then leveraged the Facebook audience API to obtain demographic information for their audiences. Such audience demographics allowed us to cover a large number of media outlets, which are at least two orders of magnitude more than what existing efforts have covered. Additionally, we also identified news outlets biased along five other axes: age, gender, income level, racial affinity, and national identity. Finally, we built and publicly deployed a system, called *Media Bias Monitor* [13], which makes the biases for these $20,448$ news outlets transparent to any Internet user.

We believe that systems such as ours are not only useful for the social media users, but also for journalists, social media researchers, developers of recommendation systems, as well as for governmental agencies wanting to understand the news generated by sources in the entire news media ecosystem. Our study forms the foundation for many research directions that can be pursued in the future for assessing and mitigating the impact of biases of news sources. As future work, we aim at expanding our system to other countries, particularly those with upcoming elections. Another research direction consists of assessing the advantages and pitfalls of audience-based and content-based methods for inferring news media bias.

Acknowledgments. This work was partially funded by a Data Transparency Lab grant. F. Ribeiro was supported by a grant from Capes. F. Benevenuto was supported by grants from Humboldt Foundation, Capes, and Fapemig. A. Chakraborty is a recipient of Google India PhD Fellowship and Prime Ministers Fellowship Scheme for Doctoral Research, a public-private partnership between Science & Engineering Research Board (SERB), Department of Science & Technology, Government of India and Confederation of Indian Industry (CII).

[13] `twitter-app.mpi-sws.org/media-bias-monitor`

References

Allcott, H., and Gentzkow, M. 2017. Social media and fake news in the 2016 election. Technical Report 2, National Bureau of Economic Research.

Araujo, M.; Mejova, Y.; Weber, I.; and Benevenuto, F. 2017. Using facebook ads audiences for global lifestyle disease surveillance: Promises and limitations. In *Proceedings of the 9th ACM Conference on Web Science (WebSci)*.

Babaei, M.; Kulshrestha, J.; Chakraborty, A.; Benevenuto, F.; Gummadi, K. P.; and Weller, A. 2018. Purple Feed: Identifying High Consensus News Posts on Social Media. In *Proceedings of the AAAI/ACM Conference on Artifical Intelligence, Ethics & Society (AIES)*.

Bakshy, E.; Messing, S.; and Adamic, L. A. 2015. Exposure to ideologically diverse news and opinion on facebook. *Science* 348(6239).

Budak, C.; Goel, S.; and Rao, J. M. 2016. Fair and Balanced? Quantifying Media Bias Through Crowdsourced Content Analysis. *Public Opinion Quarterly* 80(S1).

Chakraborty, A.; Ghosh, S.; Ganguly, N.; and Gummadi, K. P. 2015. Can trending news stories create coverage bias? on the impact of high content churn in online news media. In *Computation and Journalism Symposium*.

Chakraborty, A.; Ghosh, S.; Ganguly, N.; and Gummadi, K. P. 2016. Dissemination biases of social media channels: On the topical coverage of socially shared news. In *Proceedings of the Intl. AAAI Conference on Web and Social Media (ICWSM)*.

Chakraborty, A.; Messias, J.; Benevenuto, F.; Ghosh, S.; Ganguly, N.; and Gummadi, K. P. 2017. Who makes trends? understanding demographic biases in crowdsourced recommendations. In *Proceedings of the Intl. AAAI Conference on Web and Social Media (ICWSM)*.

Chiang, C.-F., and Knight, B. 2011. Media bias and influence: Evidence from newspaper endorsements. *The Review of Economic Studies* 78(3).

Covert, T. J. A., and Wasburn, P. C. 2007. Measuring media bias: A content analysis of time and newsweek coverage of domestic social issues, 1975–2000. *Social science quarterly* 88(3).

Dooling, D. J., and Lachman, R. 1971. Effects of comprehension on retention of prose. *Journal of experimental psychology* 88(2):216.

Garcia, D.; Kassa, Y. M.; Cuevas, A.; Cebrian, M.; Moro, E.; Rahwan, I.; and Cuevas, R. 2017. Facebook's gender divide. *arXiv preprint arXiv:1710.03705*.

Gentzkow, M., and Shapiro, J. M. 2010. What Drives Media Slant? Evidence From U.S. Daily Newspapers. *Econometrica* 78.

Gilliam Jr, F. D.; Iyengar, S.; Simon, A.; and Wright, O. 1996. Crime in black and white: The violent, scary world of local news. *Harvard Intl. Journal of Press/Politics* 1(3).

Groseclose, T., and Milyo, J. 2005. A measure of media bias. *The Quarterly Journal of Economics* 120.

Lazer, D. M. J.; Baum, M. A.; Benkler, Y.; Berinsky, A. J.; Greenhill, K. M.; Menczer, F.; Metzger, M. J.; Nyhan, B.; Pennycook, G.; Rothschild, D.; Schudson, M.; Sloman, S. A.; Sunstein, C. R.; Thorson, E. A.; Watts, D. J.; and Zittrain, J. L. 2018. The science of fake news. *Science* 359(6380).

Le, H. T.; Shafiq, Z.; and Srinivasan, P. 2017. Scalable news slant measurement using twitter. In *Proceedings of the Intl. AAAI Conference on Web and Social Media (ICWSM)*.

Lee Rodgers, J., and Nicewander, W. A. 1988. Thirteen ways to look at the correlation coefficient. *The American Statistician* 42(1).

Lella, A. 2016. Traditional news publishers take non-traditional path to digital growth. *ComScore*.

Leskovec, J.; Backstrom, L.; and Kleinberg, J. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining (KDD)*.

Mitchell, A.; Gottfried, J.; Kiley, J.; and Matsa, K. 2014. Political polarization and media habits. *Pew Research*.

Mitchell, A. 2016. Key findings on the traits and habits of the modern news consumer. *Pew Research Center*.

Munson, S.; Chhabra, S.; and Resnick, P. 2017. BALANCE - Tools for improving your news reading experience. *http://balancestudy.org/*.

Pariser, E. 2011. *The filter bubble: What the Internet is hiding from you*. Penguin UK.

Ramasubramanian, S. 2007. Media-based strategies to reduce racial stereotypes activated by news stories. *Journalism and Mass Communication Quarterly* 84(2).

Shoemaker, P. J.; Vos, T. P.; and Reese, S. D. 2009. Journalists as gatekeepers. *The Handbook of Journalism Studies*.

Shor, E.; van de Rijt, A.; Miltsov, A.; Kulkarni, V.; and Skiena, S. 2015. A paper ceiling: Explaining the persistent underrepresentation of women in printed news. *American Sociological Review* 80(5).

Speicher, T.; Ali, M.; Venkatadri, G.; Ribeiro, F. N.; Arvanitakis, G.; Benevenuto, F.; Gummadi, K. P.; Loiseau, P.; and Mislove, A. 2018. On the Potential for Discrimination in Online Targeted Advertising. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*'18)*.

Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *Science* 359(6380).

Zagheni, E.; Weber, I.; and Gummadi, K. 2017. Leveraging facebook's advertising platform to monitor stocks of migrants. *Population and Development Review*.

Zhou, D. X.; Resnick, P.; and Mei, Q. 2011. Classifying the political leaning of news articles and users from user votes. In *Proceedings of the Intl. AAAI Conference on Web and Social Media (ICWSM)*.