The Interaction between Political Typology and Filter Bubbles in News Recommendation Algorithms

Ping Liu
Illinois Institute of Technology
Chicago, IL, USA
pliu19@hawk.iit.edu

Karthik Shivaram
Tulane University
New Orleans, LA, USA
kshvaram@tulane.edu

Aron Culotta
Tulane University
New Orleans, LA, USA
aculotta@tulane.edu

Matthew A. Shapiro
Illinois Institute of Technology
Chicago, IL, USA
shapiro@iit.edu

Mustafa Bilgic
Illinois Institute of Technology
Chicago, IL, USA
mbilgic@iit.edu

ABSTRACT

Algorithmic personalization of news and social media content aims to improve user experience; however, there is evidence that this filtering can have the unintended side effect of creating homogeneous "filter bubbles," in which users are over-exposed to ideas that conform with their preexisting perceptions and beliefs. In this paper, we investigate this phenomenon in the context of political news recommendation algorithms, which have important implications for civil discourse.

We first collect and curate a collection of over 900K news articles from 41 sources annotated by topic and partisan lean. We then conduct simulation studies to investigate how different algorithmic strategies affect filter bubble formation. Drawing on Pew studies of political typologies, we identify heterogeneous effects based on the user's pre-existing preferences. For example, we find that i) users with more extreme preferences are shown less diverse content but have higher click-through rates than users with less extreme preferences, ii) content-based and collaborative-filtering recommenders result in markedly different filter bubbles, and iii) when users have divergent views on different topics, recommenders tend to have a homogenization effect.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; **Personalization**.

KEYWORDS

filter bubbles, news recommendation, political polarization, policy issues, simulation

ACM Reference Format:
Ping Liu, Karthik Shivaram, Aron Culotta, Matthew A. Shapiro, and Mustafa Bilgic. 2020. The Interaction between Political Typology and Filter Bubbles in News Recommendation Algorithms. In *Proceedings of the Web Conference 2021 (WWW '21), April 19–23, 2021 ,Ljubljana, Slovenia.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3442381.3450113

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '21, April 19–23, 2021, Ljubljana, Slovenia
© 2020 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-8312-7/21/04.
https://doi.org/10.1145/3442381.3450113

1 INTRODUCTION

Machine learning algorithms provide personalized curation of news, blogs, and social media posts to improve user experience. However, there is mounting evidence that this automated filtering leads to "filter bubbles," in which users are over-exposed to ideas that conform with their preexisting perceptions and beliefs, prompting intellectual isolation [35]. In this paper, we investigate this phenomenon in the context of political news recommendation algorithms, which can have significant and often confounding effects with regard to how people perceive consensus and mobilize around partisan and policy issues [3, 15, 18, 32, 39].

Prior work typically simplifies the problem space by reducing user preferences to a single partisan score (e.g., strong liberal to strong conservative) [37]. However, this ignores the nuanced and varied preferences users have by topic. For example, a user may have conservative views on abortion but liberal views on health care. In this work, we are interested in understanding how a user's preferences influence the behavior of recommendation algorithms, and in turn the diversity of news content to which they are exposed.

To investigate this, we first collect over 900K news articles from 41 sources annotated by topic and partisan lean. Then, drawing on recent Pew surveys of political typology [13], we simulate nine classes of users (e.g., solid liberals, disaffected Democrats, country first conservatives, etc.) with differing partisan preferences across 14 news topics. We conduct simulation studies to compare the articles recommended by *content* and *collaborative filtering* algorithms with those articles recommended by an "*oracle*" approach that observes the user's true preferences. This allows us to measure the change in diversity of recommendations introduced by the recommendation system versus what would be expected based solely on the user's true preferences. Specifically, we compare recommendation diversity and user utility measures to address the following research questions:

- **How do user preferences influence the diversity of recommendations?** We find that users with more extreme preferences are shown less diverse content but have higher click-through rates than users with less extreme preferences.
- **How do filter bubbles vary by the type of recommendation system?** We find that the filter bubbles created by content-based recommenders and collaborative filtering are

markedly different. Content-based recommendations are susceptible to biases based on how distinctive the partisan language used on a topic is, leading to over-recommendation of the most linguistically polarized topics. Collaborative filtering recommenders, on the other hand, are susceptible to the majority opinion of users, leading to the most popular topics being recommended regardless of user preferences.

- **How does recommendation diversity vary for users with heterogeneous preferences?** We find that when users have divergent views on different topics, recommenders tend to have a homogenization effect. For example, if a user is conservative on most issues, but liberal on health care, they are shown more conservative articles on health care than desired. The reasons again differ based on the type of recommender: for content-based, lexical overlap between topics can mislead the recommender; whereas for collaborative filtering, a small group of users with heterogeneous preferences are "subsumed" by a majority group that has less diverse views.

These results provide insight into the trade-off between diversity and utility in recommendation algorithms, which can help guide attempts to reduce filter bubble effects in online systems.

The rest of paper is organized as follows. We discuss the related work in Section 2. We describe our data collection, data processing, and data annotation in Section 3. We illustrate our simulation framework in Section 4, then discuss and analyze the experimental findings in Section 5. We conclude in Section 7.

2 RELATED WORK

Two primary, intersecting factors – technological and psychological – contribute to the formation of filter bubbles. The technological component refers to filtering algorithms that are designed to increase user engagement by presenting users with content that they are more likely to click on [11, 16, 27]; the psychological component refers to the tendency for users to seek out or be more accepting of information that is consistent with their preexisting attitudes and beliefs [1, 7, 21, 29, 30]. For our purposes, news source and content are central to both of these factors [34].

Much attention has been given to filter bubbles in the context of social media. For instance, research on filter bubbles has shown that, with regard to Twitter, segregation is neither uniform across ideological orientations nor across the range of topics available for consumption [4]. On Facebook, Bakshy et al. [2] examined 10 million users to quantify individual exposure to diversified news, finding that liberals are less likely to encounter ideologically cross-cutting news content than conservatives, a finding consistent with parallel research of Twitter [17]. Yet, online and offline political engagement can increase with exposure to this cross-cutting news, particularly when it originates from individuals not necessarily in one's own filter bubble, i.e. individuals with whom one has weak connections [33]. Beyond news articles themselves, and highlighting the role of influential elites in filter bubble formation [24], comments about content on Facebook and YouTube can also be predictors of echo-chamber formation [5, 41, 42].

Beyond social media-based experiments, and given that, in the U.S., nearly one-fifth of Democrats and Republicans obtain news in a filter bubble-like dynamic [28], efforts have been made to simulate recommender systems to more closely observe filter bubble dynamics. These simulations are able to control select parameters, altering specific characteristics of the online environment. Epstein et al. [19], for example, evaluated "Search Engine Manipulation Effects" and confirmed that ranking bias shifts the behavior of the voting population, thus increasing the vote share for targeted candidates. This finding has since been confirmed via experiments using representative samples of the American public [43]. Elsewhere, Geschke et al. [23] constructed an agent-based model to test the emergence of the filter bubble effect, while Chaney et al. [9] and Jiang et al. [26] attempted to build a simulation environment defining and measuring the filter bubble effect across a variety of recommender algorithms.

Ultimately, filter bubbles have significant and often confounding effects with regard to how people perceive consensus and mobilize around partisan and policy issues [3, 8, 15, 18, 32, 39]. Without some form of intervention, there are significant implications for how one is able to properly receive and process information, accurate or otherwise. Information distortions may not consistently have lasting effects [38], but filter bubbles can affect voters' election-related decisions nonetheless [18].

A number of strategies that aim to alleviate filter bubbles are proposed. Masrour et al. [31] study filter bubbles created by network link prediction algorithms and propose a framework that utilizes adversarial learning to create more heterogeneous links in the network. Bhargava et al. [6] propose providing transparency and content control mechanisms to the users to combat filter bubbles on social media. In the news consumption domain, "bias alerts" sent to users can be considered partially effective in mitigating the voting-related implications described above [19]. Providing accuracy reminders before news is consumed may minimize the likelihood that people will trust and share potentially inaccurate information [14, 36]. Yet, one's understanding of what is truly inaccurate is confounded by news source. Specifically, Dias et al. [12] find that source identification by users may help identify implausible news content from trusted new sources while simultaneously making it more difficult to identify plausible news content from untrusted news sources. This only reinforces the need to use bias alerts and accuracy reminders before news is consumed and perhaps periodically afterwards, too.

Having identified the need to account for both technological and psychological factors, the present study examines precisely how machine learning algorithms create a filter bubble effect for individuals with varying political views and vary levels of exposure to the gamut of news content. In this way, we will be able to comment directly on the causes and conditions of the filter bubble beyond the single dimension of political ideology or a single policy issue area.

3 DATA COLLECTION AND ANNOTATION

For our study, we require a large set of news articles annotated by both political stance and topic. In this section, we summarize our data collection and annotation process. Our overall approach is to use the news source as a proxy for political stance, and to use text classifiers to assign one or more topics to each article.

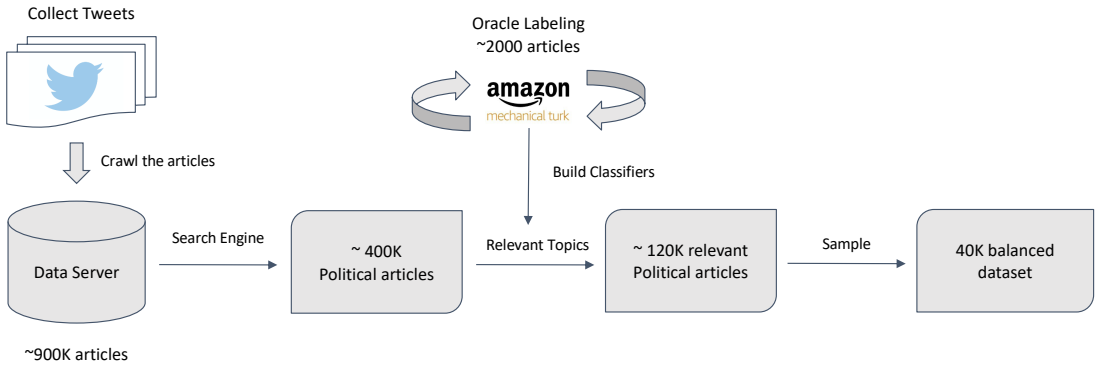

Figure 1: The pipeline to collect and label the data

stance	interpretation	# sources	# articles	% articles
-2	extreme left	10	93,700	10.1
-1	moderate left	11	282,432	30.3
0	neutral	8	286,639	30.8
+1	moderate right	4	93,279	10.0
+2	extreme right	8	175,998	18.9

Table 1: Statistics of collected news articles.

3.1 News article collection

To collect a range of political news articles, we first identified 41 featured news sources from *www.allsides.com*, which annotates each source with a *political stance* in $\{-2, -1, 0, +1, +2\}$, ranging from very liberal (-2) to very conservative (+2). The ratings are based in part on user surveys of the perceived slant of the news source.

To collect articles, we next query the Twitter API with the URL of each source to identify tweets that contain links to news articles. We then crawl each URL and collect the title, source, and content of each article. We submitted these queries continuously from September 2019 to August 2020, resulting in over 900K articles. These articles are summarized in Table 1. Popular sources from each stance include DailyBeast (-2, 17k articles), New York Times (-1, 47k), Forbes (0, 74k), Fox News (1, 36k), and Brietbart (2, 28k). Each article is annotated with the partisan score of its source.[1]

While this process gives us a broad range of articles from across the political spectrum, it is of course not without some sampling bias. E.g., articles shared on Twitter differ from a uniform random sample of all articles from all news sources. However, given that our focus is on articles likely to be read and shared by users, this sampling methodology seems appropriate for our purposes. To account for the unequal distribution of articles by partisan stance, in our experiments below we sample to have a balanced distribution of articles.

[1]While this may introduce some label noise at the article level [22], we expect this to have limited impact in aggregate.

Topic	Negative Labels	Positive Labels
LGBTQIA	1,972	114
abortion	1,909	177
environment	1,963	123
guns	2,014	72
health care	1,947	139
immigration	1,978	108
racism	1,986	100
taxes	1,963	123
technology	2,032	54
trade	2,006	80
trump impeachment	1,803	283
us 2020 election	1,725	361
us military	2,001	85
welfare	2,002	84

Table 2: Label Distributions of Training Data for Topic Classification

3.2 Topic classification

From the 900K articles we collected, our next goal is to build a classifier to annotate each article with the topics it discusses. To do so, we trained a two-stage classifier: one to determine if the article is relevant to U.S. politics, and a second to assign one or more topics to the article.

To collect training data, the five co-authors first independently annotated a sample of documents with political relevance and topics. Through several discussions and iterative refinement, we arrived at the following list of 14 topics: *abortion, environment, guns, health care, immigration, LGBTQIA, taxes, technology, trade, Trump impeachment, US military, welfare, US 2020 election*, and *racism*.

To increase the training sample, we next sampled additional documents to be annotated using Amazon Mechanical Turk. Using our expert annotations as a guide, we identified 12 high-quality AMT annotators, and had them annotate 3,250 total documents, of which 2,086 were annotated as politically relevant. The label distribution of this annotated dataset can be seen in Table 2.

Accuracy	F1	Recall	Precision
0.7865	0.8307	0.7909	0.8773

Table 3: Performance of Relevance Classifier

Topic	F1	Topics	F1
abortion	0.942	environment	0.898
guns	0.906	healthcare	0.785
immigration	0.853	LGBTQIA	0.894
racism	0.776	taxes	0.848
technology	0.538	trade	0.839
impeachment	0.888	US military	0.773
US election 2020	0.847	welfare	0.598

Table 4: The F1 scores of the Topic Classifiers

	before sampling		after sampling	
topics	# articles	% articles	# articles	% articles
abortion	3,421	1.7	1,382	2.6
environment	4,329	2.2	1,656	3.2
guns	4,647	2.4	1,787	3.4
healthcare	14,823	7.6	5,444	10.6
immigration	10,736	5.5	4,308	8.3
LGBTQIA	2,848	1.5	1,126	2.1
racism	10,051	5.1	4,069	7.9
taxes	8,187	4.2	3,055	5.9
technology	3,722	1.9	1,379	2.6
trade	6,739	3.4	2,323	4.5
impeachment	45,989	23.4	6,811	13.2
US military	17,205	8.8	9,409	18.3
US election 2020	57,996	29.6	6,501	12.6
welfare	5,413	2.7	2,054	4.0
# labels	196,106		51,304	
# articles	167,431		40,000	

Table 5: News article topics distribution.

From these labeled data, we next trained a binary classifier to determine if the article is relevant to U.S. politics or not. For this we used a standard logistic regression model using tf-idf features. Table 3 summarizes the accuracy of this classifier.

For topic classification, as it is a multi-label classification task, we trained 14 independent binary classifiers (one per topic). As the label distributions is highly imbalanced, we used SMOTE (Synthetic Minority Oversampling Technique) [10] to over-sample the positive class. Each of these topic classifiers uses logistic regression and tf-idf based features. The settings for the tf-idf vectorizer are as follows: the maximum number of features is 5,000, the maximum document frequency is 0.95, and the minimum document frequency is 30. These classifiers were separately optimized using a 5-fold cross validation loop with grid-search using the F1-score as the optimization metric. Table 4 shows the final cross-validation results for each topic. While F1 is generally high, we note that the classifier has smaller F1 score for the technology and welfare topics. For technology, this is likely do to ambiguity of whether an article is related to U.S. politics – e.g., an article about Facebook's earnings is not relevant, but one that discusses new regulations is. For welfare, this topic is much broader than the rest, covering everything from cash assistance programs to homelessness issues. More training data would likely help here.

3.3 Article Sampling

With the two classifiers described above, we then annotated all collected articles with relevance and topic. Table 5 shows the predicted topic distribution of those articles determined to be relevant and to have at least one topic assigned. To ensure that the final sample has a uniform distribution of political stance, we randomly sample 8K articles from each stance, resulting in the final topic distribution in the final two columns in the table. (Note that many articles have more than one topic assigned.) Given the high fraction of articles about the 2020 election and Trump's impeachment, we additionally down-sampled these topics to ensure a broader diversity of articles.

4 SIMULATION FRAMEWORK

In order to study the relationship between user preferences and recommendation systems, we would ideally conduct large-scale user studies to observe real-world interactions. However, given the challenges of conducting such studies, we instead build on the growing line of research conducting simulation studies of recommendation systems [9, 25, 26, 40].

To conduct such a simulation, we must make some assumptions about the interaction model. Our approach largely follows that of prior work [9, 25], though here we use real news articles annotated by stance and topic. We assume that each user has a predefined, fixed set of preferences over articles they would like to read. These preferences are parameterized by the topic and stance of the article; e.g., a user may prefer to read a liberal article about healthcare more than a conservative article about immigration. As we are interested in short-term effects of recommenders, for this study we assume that user preferences do not change over time, though this is of course an important consideration for future studies.

The simulation proceeds by first showing the user an article. We then simulate the user's response: either "like" or "dislike," sampled proportional to the user's preferences. With this feedback, the recommender updates its model to re-sort the remaining articles, then shows the next article to the user.

In the following sections, we describe this process in more detail, including the user profile model, a user-choice model, and specific recommendation engines we implement.

4.1 User utility model

We represent each user's preferences with a two-dimensional matrix of utility values $U = \{u_{ij}\}$, where $u_{ij} \in [0, 1]$ indicates the user's utility for reading an article on topic i with political stance j. (Thus, U is a 14×5 matrix.) Large values indicate greater utility and therefore a larger probability of clicking on an article with topic i and stance j.

We wish to investigate how recommender behavior varies with heterogeneous utility matrices. Rather than randomly generate these matrices, in order to make them more reflective of reality, we sampled them based on Pew surveys of U.S. political typologies [13]. This comprehensive survey attempts to identify more nuanced political ideologies than a simple left/right spectrum. The survey contains many questions relevant to our identified topics above. E.g., for abortion, there is a survey question asking whether abortion should be legal in all/most cases. For immigration, there is a question asking whether immigrants strengthen or weaken the country. Pew clustered the responses to identify nine political types: *solid liberals, opportunity Democrats, disaffected Democrats, bystanders, devout and diverse, new era enterprisers, market skeptic Republicans, country first conservatives*, and *core conservatives*. These types capture a number of common heterogeneous ideologies – for example, the devout and diverse type leans conservative on issues of abortion and LGBTQIA, but leans liberal on race and health care. Similarly, the market skeptic Republicans lean liberal on issues of trade and taxation.

For each political type, then, we have a list of survey responses indicating the fraction of respondents who agree with the statement (e.g., 92% of solid liberals think that abortion should be legal in all/most cases). In our simulations, to generate a new user, we first pick a political type, then sample a utility matrix based on these survey responses. We convert these responses into a utility matrix as follows: for each survey question, we separate the responses into quantiles (0-20%, 21-40%, etc.), and assign the response to one of the five political stance categories $\{-2, 1, 0, +1, +2\}$. Thus, the fact that 92% of solid liberals think abortion should be legal means that their primary stance is -2 on abortion. To generate the utility value for each topic/stance pair, we first sample a utility value for the primary stance using a Beta distribution centered on their survey response (e.g., $Beta(.92, 1)$ for the running example). We then decay this value for the other stances for this topic as a function of standard deviation of responses on this topic (i.e., a measure of how divisive this topic is). We then repeat this process for each topic. Table 6 shows an example utility matrix for the devout and diverse profile.

As with any simulation, one can question how reflective the simulated users are of the real world. The key aspect that these utilities do capture, however, is a broad spectrum of ideologies with which we can investigate variation in recommender behavior.

4.2 User interaction model

Given a user's utility matrix, we next must simulate their behavior when presented with a recommended article. To do so, we follow the approach of prior work [9]. To represent each article, we create a binary matrix of the same shape as the user utility matrix, containing 1 in cell (i, j) if the article has been assigned topic i and stance j. (Recall that the topic is derived from the text classifier, and the stance from the news source.) To sample whether a user will "like" or "dislike" an article, we first flatten both the utility matrix and the item matrix into 1d arrays, then compute the dot product between them. We then sample a value from a *Beta* distribution centered on this dot product value. Finally, a random number is generated and compared to the sampled value to determine the action of the user. Algorithm 1 formalizes this process.

topics	-2	-1	0	+1	+2
abortion	0.276	0.411	0.546	**0.682**	0.546
environment	0.298	0.505	**0.711**	0.505	0.298
guns	0.332	0.490	**0.648**	0.490	0.332
healthcare	0.515	**0.711**	0.515	0.319	0.122
immigration	0.045	0.285	0.525	**0.766**	0.525
LGBTQIA	0.250	0.423	0.596	**0.769**	0.596
racism	**0.815**	0.575	0.335	0.095	0.010
taxes	0.080	0.283	0.486	**0.689**	0.486
technology	0.228	0.397	0.567	**0.737**	0.567
trade	0.400	0.511	0.622	**0.733**	0.622
Trump impeachment	0.313	0.452	**0.592**	0.452	0.313
US military	0.171	0.362	0.553	**0.744**	0.553
US election 2020	0.180	0.395	**0.610**	0.395	0.180
welfare	**0.860**	0.582	0.304	0.025	0.010

Table 6: An example of the utility matrix for a "devout and diverse" user.

Algorithm 1 The user interaction model

Input: u – the user vector; v – the article vector
Output: B – a Boolean variable to indicate whether the user likes this article or not.

$v_{ui} = Beta^1(dot(\text{u}, normalized(\text{v}))$
$p_{ui} = v_{ui} \times Beta^1(0.98)$
if $Random < p_{ui}$ **then**
  return Like
else
  return Dislike
end if

In the algorithm, the function takes the user vector u and the item vector v. We calculate the dot product with u and normalized v to constrain the output as a probability from 0 to 1. Following previous work [9], we choose a modified $Beta^1$ distribution (for which the mean and standard deviation are given) to calculate the probability p_{ui} the user will click the given article. A random number is generated and used to determine whether the user will click this article, given p_{ui}.

4.3 Recommender models

We implemented five recommender systems, including a random recommender (as a baseline), a content-based recommender, a collaborative filtering recommender, an oracle recommender, and a hybrid recommender.

4.3.1 Random recommender. A random news recommender randomly selects the articles from the pool without replacement.

4.3.2 Content-based recommender. A content-based recommender (CBR) is a user-personalized model that learns the user's preference, given the user's previous interactions. We treat this as a binary classification problem – given an article, will the user like or dislike it? As training data, we seed the model with 700 simulated examples per user, sampled uniformly for each topic. We train a standard logistic regression classifier separately for each user, using tf-idf word features from each article. During the simulation, the

training data is updated after each user interaction, and the model is retrained. Note that the classifier does not observe the stance and topic assignments for each document – this simulates the situation where neither the structure nor values of the user's utility matrix are known to the recommender.

4.3.3 Collaborative Filtering recommender. A collaborative filtering recommender (CFR) uses the concept of similarities between users and items and recommend similar users the 'liked' items from each other's 'like' history. We use nonnegative matrix factorization [20] on the user-item matrix to construct the collaborative filtering recommender.

4.3.4 Oracle recommender. We also implement an oracle recommender, which observes the user's utility matrix and news' topic and stance matrix. This algorithm samples documents proportional to the user's probability of liking these documents. This baseline enables us to observe what biases are introduced by the recommender algorithms versus those that are inherent in the user's pre-existing preferences.

4.3.5 Hybrid recommender. A simple way to try to reduce filter bubbles is to inject random recommendations into the user's article list. We are interested in how the systems behave as the amount of randomness is injected. How quickly does the diversity increase as we introduce randomness? To investigate this, we consider three settings for each recommender above: randomness as 0% (totally personalized), 50% (hybrid), and 100% (totally random).

5 EXPERIMENTAL METRICS AND RESULTS

In order to answer the three questions we proposed in Section 1, we designed simulations to study recommender behavior for users of different political types. In this section, we formulate our filter bubble metrics and the details of experimental setup, then discuss the experimental results. [2]

5.1 Problem formulation and metrics

Let V be a collection of news articles. Each article $v \in V$ is associated with one or more of 14 topics introduced in Section 3.2. Let U be a group of users. Each user $u \in U$ belongs to one of the nine political types introduced in Section 4.1. In each simulation run, every user u is recommended N articles, one at a time. For each recommended article i, we simulate a binary random variable r_i, where $r_i = 1$ mean the user clicks on /likes the article and $r_i = 0$ means they do not. We propose the following metrics to study the filter bubble effect of different algorithms on different political types.

5.1.1 Click-through rate. The click-through rate (CTR) is the fraction of recommended articles that the user clicks on. A high CTR indicates that the algorithm can deliver accurate recommendations to the users, and thus has high utility. The CTR is defined as follows.

$$CTR = \frac{\sum(r_i)}{N}, 1 \le i \le N \tag{1}$$

[2]The code and data that we used to derive the experimental results in this paper are available https://github.com/IIT-ML/WWW21-FilterBubble

5.1.2 Average document stance. Average document stance is the average partisan score of the articles that are *shown* to the users. Letting $s(v_i) \in \{-2, -1, 0, 1, 2\}$ be the partisan score for article v_i, then the average document stance for a sequence of recommended articles is:

$$\bar{s} = \frac{\sum s(v_i)}{N}, 1 \le i \le N \tag{2}$$

5.1.3 Normalized stance entropy. Let p_i represent the fraction of articles that are shown to the users that have stance i. Normalized stance entropy is the entropy of this distribution, normalized by $\log m$ so that its maximum is 1, where $m = 5$ in our case, representing the five stances:

$$entropy = \frac{-\sum_{i=1}^{m} p_i \log p_i}{\log m} \tag{3}$$

A high value of normalized stance entropy would indicate a smaller filter bubble effect since the stances of the shown articles are more diverse.

5.1.4 Normalized topic entropy. Similar to normalized stance entropy, we also measure the diversity of topics. This provides a measure of topical diversity, in addition to stance diversity above. The metric is the same as Equation 3, where p_i is instead the probability of articles having topic i in a sequence of recommendations, and $m = 14$ since there are 14 topics. A low value of normalized topic entropy indicates that the recommender is recommending documents in a small set of topics.

5.2 Experimental setup

We generate 100 synthetic users for each political type following the user utility model described in Section 4.1. To initialize the recommendation models, we initially bootstrap 50 articles per topic for each user, resulting in 700 articles in total. Then the recommender recommends 1,000 articles, one by one, in a sequence and updates the algorithm after each recommendation. The CBR and CFR have three different randomness settings as we mentioned in the previous section.

We simulate the oracle recommender explicitly as follows. For a given political type, for every article v, we calculate the probability p_v that the given political type would click that article if they are shown that article, based on their user profile. To study varying degrees of randomness in the oracle recommender, we compute a sampling weight for each article as $\exp(w \times p_v)$ where w is a hyper-parameter. We sample K articles from our dataset, using weighted sampling without replacement. We repeat this process M times. The probability q_v that the article will be shown by the oracle is the fraction of samples that contain v. When $w = 0$, each article has $\exp(0 \times p_v) = 1$ weight, resulting in uniform sampling, and hence results in the random algorithm. As $w > 0$, articles that have a higher chance of being clicked gets a higher weight.

Once we have the shown (q_v) and click (p_v) probabilities, we can calculate the expectations for the CTR and all other metrics for all the political types using the whole dataset. We choose to use K as 1000, and M as 5000 in our case. For the hyper-parameter w, we vary the value from 0 (totally random) to 9 (optimal personalized solution). For comparing CBR and CFR to the oracle recommender,

Figure 2: Simulation results by political typology, showing click-through rate vs average document stance for three levels of randomness.

we use w that achieves a similar CTR for that prototype, and analyze where the CBR and CFR differ from the oracle. This analysis allows us to measure the bias introduced by the recommender beyond that inherent in the user preferences.

5.3 Experimental results

5.3.1 How do user preferences influence the diversity of recommendations? We first investigate how the user's political type influences the diversity of the recommended documents. Because there is a strong relationship between diversity and utility (i.e., CTR), we are particularly interested in their trade-off. We consider content-based recommender, collaborative filtering recommender, and the oracle recommender. For each, we have varying levels of randomness through the hybrid recommendation approach. In this way, we can plot how the CTR varies with filter bubble measures such as average document stance, stance entropy, and topic entropy. We would like to determine how this trade-off varies by political type.

Figure 2 shows the main results of CTR versus average document stance. Each panel summarizes the results of multiple simulation runs. Each dot represents the result for one user. For content-based recommender and collaborative filter recommender, each political type has three settings, which are 0% randomness, 50% randomness (hybrid recommender), and 100% randomness (random recommender). The larger symbols (e.g., circle, triangle, and cross) represent the centroids of each setting. For the oracle recommender, the randomness is controlled by the w parameter, where w ranges from $w = 0$ (fully random) to $w = 9$ (user preferences are given high priority). We also fit a LOWESS curve for each political type to visualize the tradeoff between CTR and document stance.

The first observation is that more extreme political types have both higher CTR and higher magnitude document stances. E.g., when no randomness is used, country-first conservatives have over a 60% CTR, and an average partisan score of nearly 1.0 for both content-based and collaborative filtering recommendations. On the other hand, more moderate political types, such as bystanders and devout & diverse, do not attain such high CTRs. These results make clear the intuitive finding that the more extreme a user's preferences are, the more extreme their recommendations will be, and that it is easier to find articles that they are likely to click.

We can also see from the third panel that the oracle is able to achieve even higher CTRs, though to do so it must recommend even more extreme and homogeneous documents.

Figure 3 shows a similar result instead using stance entropy as a measure of diversity. For more extreme users, stance entropy decreases more quickly as CTR increases.

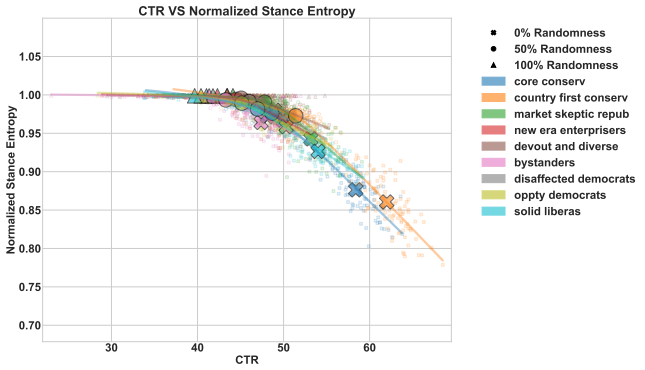

Figure 3: Click-through rate vs normalized stance entropy for the content-based recommender.

Examining these figures, there is a notable difference in the recommendation behavior for left-leaning versus right-leaning users. In the first panel of Figure 2, we see that right-leaning users ultimately exhibit higher CTRs, and more extreme partisan scores, than left-leaning users. Furthermore, we only see this difference in the content recommender, not for collaborative filtering or oracle recommenders. Upon further inspection, we conjecture that this is in part due to the asymmetry in the textual similarities between documents of different partisan scores. In particular, it appears that articles with score 0 are more similar to left-leaning articles (scores -2, -1) than they are to right-leaning articles (scores +1, +2). The result is that the content-based recommender has a more difficult

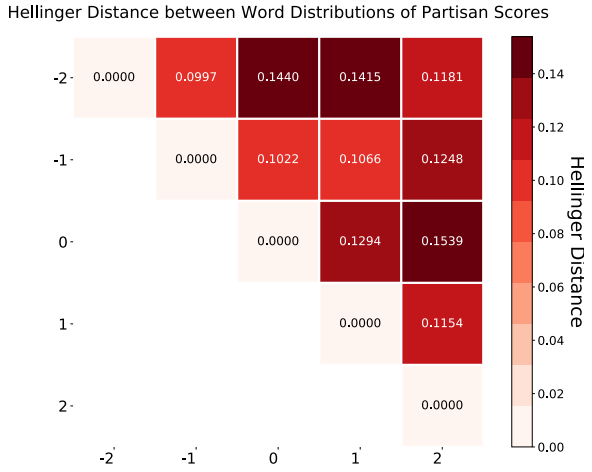

Figure 4: Hellinger Distance between different Partisan Scores

time distinguishing between -2 and 0 articles than it does distinguishing between +2 and 0 articles. To further investigate this, we fit five different multinomial bag-of-words models, one per partisan score, by grouping together all articles with the same partisan score. We then compute the Hellinger distance between each pair of multinomials to determine how similar the word distributions are. We find that the differences between -2 and 0 (.1415) and -1 and 0 (.1022) are substantially smaller than that between +2 and 0 (.1539) and +1 and 0 (.1294), further supporting this interpretation (Figure 4).

5.3.2 How do filter bubbles vary by type of recommendation system? As we have just seen, different recommendation systems can have different impact on filter bubble formation. In this section, we further compare CBR and CFR to their comparable oracle recommender counterpart to investigate possible biases introduced by CBR and CFR into the recommendation processes. To do so, we first compute the average number of articles recommended from each topic/partisan score pair for each political type, using the versions of CBR and CFR with the highest overall click-through rate. We then compare these values with the corresponding recommendations provided by the oracle recommender.[3]

Figure 5 shows the results for three political types: country-first conservatives (CFC), devout and diverse (D&D), and Opportunity Democrats (OPD). Each cell in the heat map displays the difference between the average number of articles recommended by either CBR/CFR and those recommended by the oracle. For example, in the top left panel, we see that the content-based recommender shows on average 113 more immigration/+2 documents than the oracle does to country-first conservatives. By examining these results, we can identify a few trends that characterize the different sorts of bias introduced by either content-based or collaborative filtering recommenders.

[3]We select the randomness hyper-parameter w to result in an oracle with the same click-through rates as the CBR or CFR method it is being compared with.

For CBR, a key source of bias is **linguistic polarization**. For some topics, there is a clear distinction between the language used in right-leaning articles versus left-leaning articles. For example, in the immigration topic, terms like "illegal" and "alien" are much more likely to appear in right-leaning articles, while terms like "undocumented" are more common in left-leaning articles. In such cases, it will take few training examples for the recommender to develop an accurate model of user preferences, resulting in an over-recommendation of such topics. Furthermore, this can often result in a feedback loop, wherein immigration/+2 articles are recommended and clicked on, further reinforcing the over-recommendation of such articles.

This behavior is most noticeable in the immigration/+2 cell of the first panel of Figure 5. We can further see this behavior in Figure 6, which shows that content-based recommenders tend to have lower entropy over topics shown than the other two recommendation models for all of the political types at the extreme ends.

For collaborative filtering, we identify two sources of bias. The first is that the distribution of preferences across all users will influence the popularity of some topics over others. For example, across all political types, abortion and trade have high utilities, so they tend to be over-recommended across all user types. We also observe that minority groups tend to be 'subsumed' by larger groups. For example, the devout and diverse group appears to be grouped with more right-leaning groups and hence recommended more right articles across almost all topics, whereas the opportunity Democrats are grouped with left-leaning groups and hence are recommended more left articles across almost all topics, as the bottom row of Figure 5 shows.

A final source of bias that affects both recommendation systems is the overall makeup of the pool of articles to be recommended. As Table 5 indicates, topics such as US military, US election, and impeachment are the most common. The initial bootstrap for CBR and CFR had equal articles from each topic (50 articles from each topic), hence these topics were underrepresented compared to their representation in the overall pool. Thus, articles from these topics tend to be under-recommended by CBR and CFR systems compared to the oracle recommender, which does not have a bootstrap and hence is unaffected by it.

5.3.3 How does recommendation diversity vary for users with heterogeneous preferences? The biases described above can also have effects on users with heterogeneous preferences. For example, Devout and Diverse users lean right on most issues, but lean left on issues of race, welfare, and health care. Both content-based and collaborative filtering systems under-recommend left leaning articles on these topics, but for different reasons. For collaborative filtering, the devout and diverse users are clustered together with other right-leaning users (e.g., core conservatives). Because those other users have right-leaning preferences for race and welfare, the devout and diverse users are recommended similar articles. Similarly, while the content-based recommender over predicts immigration/+2 for country-first conservatives, the collaborative filtering algorithm instead *under* predicts this category. The CFC type is most distinct because it is more conservative on immigration than "typical" right-leaning users, and so they are grouped together with these more typical users and shown less extreme views on immigration.

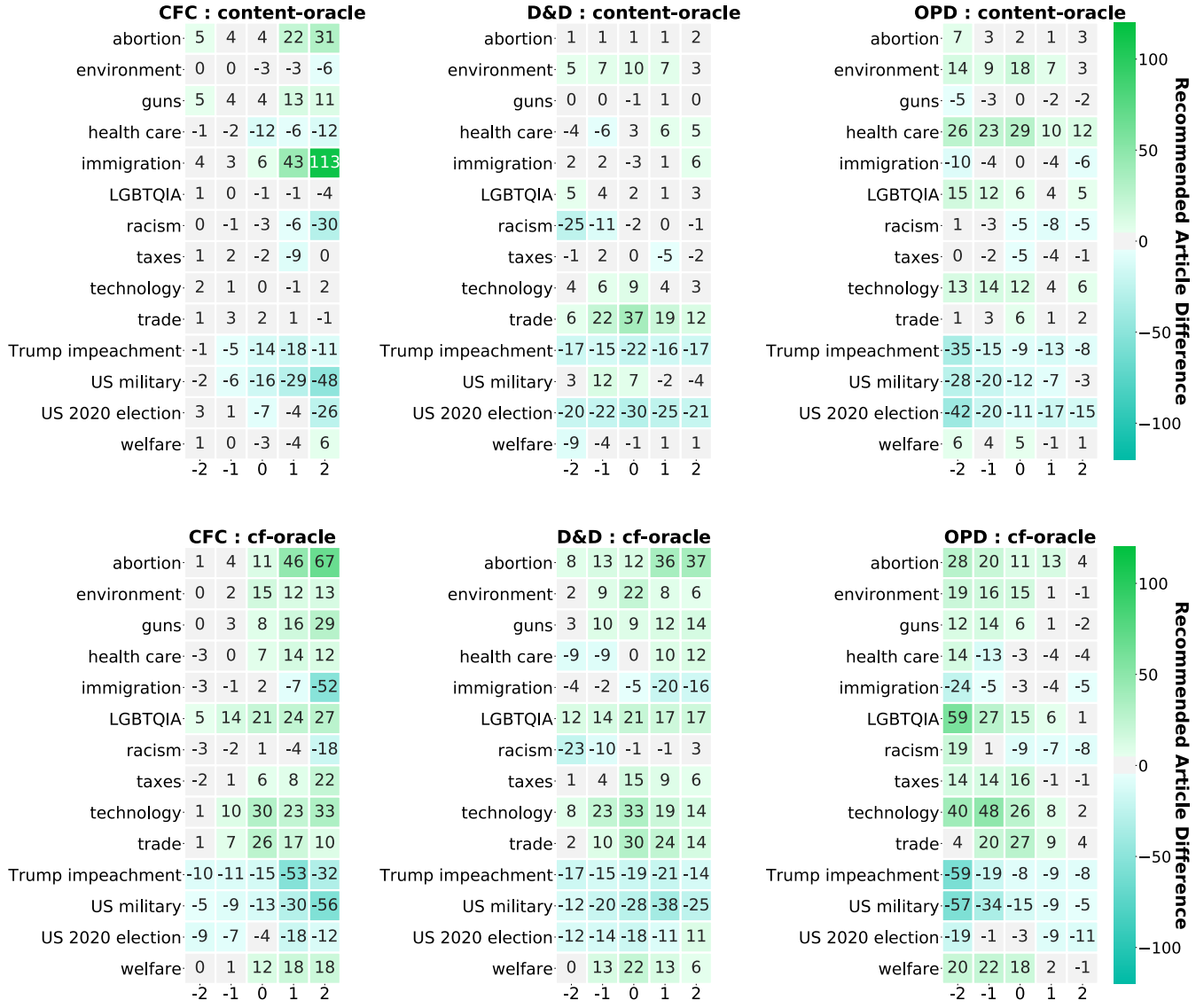

Figure 5: Difference in the number of articles recommended by the content-based and collaborative filtering recommenders as compared to the oracle recommender. Results are the average of 1,000 recommendations for 100 users from three user types: country first conservatives (CFC), devote and diverse (D&D), and opportunity Democrats (OPD).

The explanation for the content-based recommender is more nuanced. A central issue is that there is keyword overlap across topics that can mislead the recommender. For example, the keyword "baby" correlates with right-leaning articles both for the abortion topic and the health care topic. Because D&D users lean right on abortion issues, after clicking on several right-leaning abortion articles, the recommender may also start to recommend right-leaning health care articles, contrary to their preferences. Similar behavior occurs between the welfare and taxes topic, where the term "socialist" correlates with right-leaning articles for both topics. As D&D users lean right on taxes but left on welfare, left-leaning articles on welfare are under-recommended.

Together, these examples suggest that recommender systems can have a homogenization effect on such users, for example by pushing D&D users to more typical right-leaning articles, and by pushing opportunity democrats to more typical left-leaning articles, even though their true preferences are more mixed. Importantly, we do not see such behavior for the oracle recommender, but rather these are artifacts of the biases of recommendation systems that learn imperfect models of user preferences.

6 LIMITATIONS AND FUTURE WORK

We assumed that the news source's partisan score was reflective of its articles. While this appears to be a reasonable assumption

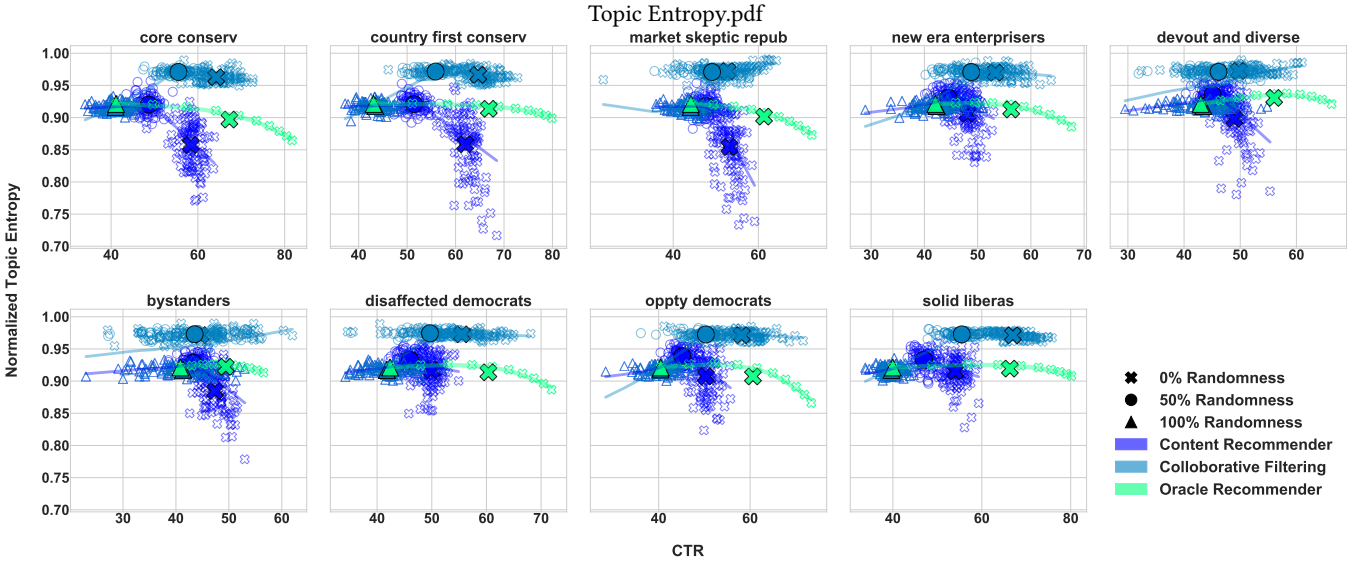

Figure 6: Click-through rate vs normalized topic entropy for all recommenders. The content-based recommender exhibits much lower topic diversity than others.

in aggregation, there are undoubtedly some individual errors introduced here. We plan to build partisan score classifiers for each topic to relax this assumption. In the meantime, we need to take into account that the classifiers might introduce their own bias. Further, the user utility model is constant during the recommendation process. Modeling long-term effects requires further assumptions about the causal effect of news consumption on reader beliefs. Typical recommender systems suffer from self-reinforcement because their training data is tainted by skewed recommendations. One might expect that the filter bubbles could cause the user views to become less heterogeneous, further reinforcing and exacerbating filter bubbles. Our paper focuses on short-term effects on news consumption, leaving effects on reader beliefs for future work.

7 CONCLUSION

In this paper, we have presented several simulations to understand the relationship between political typology and news recommendation algorithms. We find that users with more extreme views tend to be easier for recommendation systems to model, and thus tend to enjoy higher click-through rates, though this is only possible with less diverse recommendations both in terms of political views and topics. Furthermore, we find that two common classes of recommendation algorithms, content-based and collaborative filtering, can each result in filter bubbles, though of different types and for different reasons. Finally, we find that users with heterogeneous preferences tend to be recommended articles that reflect more homogeneous viewpoints. These results suggest that future work in news article recommendation should consider a wider range of metrics when measuring diversity and also consider a wider range of user preferences.

As with any simulation, this work must be further supported by studies with real users. While large scale studies remain challenging, in future work, we plan to conduct user studies using a custom-built recommendation engine to test the external validity of the conclusions drawn here.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under grants #1350337 and #1927407.

REFERENCES

[1] Francis Bacon. 2000. Three steps toward a theory of motivated political reasoning. *Elements of reason: Cognition, choice, and the bounds of rationality* 183 (2000).
[2] Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132.
[3] Delia Baldassarri and Andrew Gelman. 2008. Partisans without constraint: Political polarization and trends in American public opinion. *Amer. J. Sociology* 114, 2 (2008), 408–446.
[4] Pablo Barberá, John T Jost, Jonathan Nagler, Joshua A Tucker, and Richard Bonneau. 2015. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science* 26, 10 (2015), 1531–1542.
[5] Alessandro Bessi, Fabiana Zollo, Michela Del Vicario, Michelangelo Puliga, Antonio Scala, Guido Caldarelli, Brian Uzzi, and Walter Quattrociocchi. 2016. Users polarization on Facebook and Youtube. *PloS one* 11, 8 (2016), e0159641.
[6] Rahul Bhargava, Anna Chung, Neil S Gaikwad, Alexis Hope, Dennis Jen, Jasmin Rubinovitz, Belén Saldías-Fuentes, and Ethan Zuckerman. 2019. Gobo: A System for Exploring User Control of Invisible Algorithms in Social Media. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. 151–155.
[7] Toby Bolsen, James N Druckman, and Fay Lomax Cook. 2014. The influence of partisan motivated reasoning on public opinion. *Political Behavior* 36, 2 (2014), 235–262.
[8] Pablo Briñol, Derek D. Rucker, Zakary L. Tormala, and Richard E. Petty. 2003. *Individual differences in resistance to persuasion: The role of beliefs and meta-beliefs*. Routledge Taylor & Francis Group, 83–104.
[9] Allison JB Chaney, Brandon M Stewart, and Barbara E Engelhardt. 2018. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 224–232.
[10] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.
[11] Wei Chu and Seung-Taek Park. 2009. Personalized recommendation on dynamic content using predictive bilinear models. In *Proceedings of the 18th international*

conference on World wide web. 691–700.
[12] Nicholas Dias, Gordon Pennycook, and David G Rand. 2020. Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media. *Harvard Kennedy School Misinformation Review* 1, 1 (2020).
[13] C Doherty, J Kiley, and B Johnson. 2017. Political typology reveals deep fissures on the right and left: Conservative Republican groups divided on immigration,'openness'. *Pew Research Center* (2017).
[14] James N. Druckman. 2015. Communicating Policy-Relevant Science. *PS: Political Science & Politics* 48, S1 (2015), 58–69.
[15] James N Druckman and Arthur Lupia. 2017. Using frames to make scientific communication more effective. *The Oxford handbook of the science of science communication* (2017), 243–252.
[16] Susan Dumais, Thorsten Joachims, Krishna Bharat, and Andreas Weigend. 2003. SIGIR 2003 workshop report: implicit measures of user interests and preferences. In *ACM SIGIR Forum*, Vol. 37. ACM New York, NY, USA, 50–54.
[17] Gregory Eady, Jonathan Nagler, Andy Guess, Jan Zilinsky, and Joshua A. Tucker. 2019. How Many People Live in Political Bubbles on Social Media? Evidence From Linked Survey and Twitter Data. *SAGE Open* 9, 1 (2019), 2158244019832705.
[18] Robert Epstein and Ronald E Robertson. 2015. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences* 112, 33 (2015), E4512–E4521.
[19] Robert Epstein, Ronald E Robertson, David Lazer, and Christo Wilson. 2017. Suppressing the search engine manipulation effect (SEME). *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–22.
[20] Cédric Févotte and Jérôme Idier. 2011. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural computation* 23, 9 (2011), 2421–2456.
[21] Jonathan L Freedman and David O Sears. 1965. Selective exposure. In *Advances in experimental social psychology*. Vol. 2. Elsevier, 57–97.
[22] Soumen Ganguly, Juhi Kulshrestha, Jisun An, and Haewoon Kwak. 2020. Empirical Evaluation of Three Common Assumptions in Building Political Media Bias Datasets. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 939–943.
[23] Daniel Geschke, Jan Lorenz, and Peter Holtz. 2019. The triple-filter bubble: Using agent-based modelling to test a meta-theoretical framework for the emergence of filter bubbles and echo chambers. *British Journal of Social Psychology* 58, 1 (2019), 129–149.
[24] A Guess. 2018. (Almost) everything in moderation: New evidence on Americans' online media diets.
[25] Eugene Ie, Chih wei Hsu, Martin Mladenov, Vihan Jain, Sanmit Narvekar, Jing Wang, Rui Wu, and Craig Boutilier. 2019. RecSim: A Configurable Simulation Platform for Recommender Systems. (2019). arXiv:1909.04847 [cs.LG]
[26] Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli. 2019. Degenerate feedback loops in recommender systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 383–390.
[27] Christopher C Johnson. 2014. Logistic matrix factorization for implicit feedback data. *Advances in Neural Information Processing Systems* 27 (2014), 78.
[28] M Jurkowitz and A Mitchell. 2020. About one-fifth of Democrats and Republicans get political news in a kind of media buble. *Pew Research Center* (2020).
[29] Dan M Kahan. 2015. The politically motivated reasoning paradigm, part 1: What politically motivated reasoning is and how to measure it. *Emerging trends in the social and behavioral sciences: An interdisciplinary, searchable, and linkable resource* (2015), 1–16.
[30] Ziva Kunda. 1990. The case for motivated reasoning. *Psychological bulletin* 108, 3 (1990), 480.
[31] Farzan Masrour, Tyler Wilson, Heng Yan, Pang-Ning Tan, and Abdol Esfahanian. 2020. Bursting the Filter Bubble: Fairness-Aware Network Link Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 841–848.
[32] Aaron M McCright and Riley E Dunlap. 2011. The politicization of climate change and polarization in the American public's views of global warming, 2001–2010. *The Sociological Quarterly* 52, 2 (2011), 155–194.
[33] Seong Jae Min and Donghee Yvette Wohn. 2018. All the news that you don't like: Cross-cutting exposure and political participation in the age of social media. *Computers in Human Behavior* 83 (2018), 24 – 31.
[34] Subhayan Mukerjee and Tian Yang. 2020. Choosing to Avoid? A Conjoint Experimental Study to Understand Selective Exposure and Avoidance on Social Media. *Political Communication* 0, 0 (2020), 1–19.
[35] Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you*. Penguin UK.
[36] Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G Lu, and David G Rand. 2020. Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science* 31, 7 (2020), 770–780.
[37] Cristian G Rodriguez, Jake P Moskowitz, Rammy M Salem, and Peter H Ditto. 2017. Partisan selective exposure: The role of party, ideology and ideological extremity over time. *Translational Issues in Psychological Science* 3, 3 (2017), 254.
[38] Matthew J. Salganik and Duncan J. Watts. 2008. Leading the Herd Astray: An Experimental Study of Self-fulfilling Prophecies in an Artificial Cultural Market. *Social Psychology Quarterly* 71, 4 (2008), 338–355.
[39] Glenn S Sanders and Brian Mullen. 1983. Accuracy in perceptions of consensus: Differential tendencies of people with majority and minority positions. *European journal of social psychology* 13, 1 (1983), 57–70.
[40] Sven Schmit and Carlos Riquelme. 2018. Human interaction with recommendation systems. In *International Conference on Artificial Intelligence and Statistics*. 862–870.
[41] Matthew A. Shapiro and Han Woo Park. 2015. More than entertainment: YouTube and public responses to the science of global warming and climate change. *Social Science Information* 54, 1 (2015), 115–145.
[42] Matthew A. Shapiro and Han Woo Park. 2018. Climate Change and YouTube: Deliberation Potential in Post-video Discussions. *Environmental Communication* 12, 1 (2018), 115–131.
[43] Yotam Shmargad and Samara Klar. 2020. Sorting the News: How Ranking by Popularity Polarizes Our Politics. *Political Communication* 37, 3 (2020), 423–446.