More Accounts, Fewer Links: How Algorithmic Curation Impacts Media Exposure in Twitter Timelines

JACK BANDY and NICHOLAS DIAKOPOULOS, Northwestern University, USA

Algorithmic timeline curation is now an integral part of Twitter's platform, affecting information exposure for more than 150 million daily active users. Despite its large-scale and high-stakes impact, especially during a public health emergency such as the COVID-19 pandemic, the exact effects of Twitter's curation algorithm generally remain unknown. In this work, we present a sock-puppet audit that aims to characterize the effects of algorithmic curation on source diversity and topic diversity in Twitter timelines. We created eight sock puppet accounts to emulate representative real-world users, selected through a large-scale network analysis. Then, for one month during early 2020, we collected the puppets' timelines twice per day. Broadly, our results show that algorithmic curation increases source diversity in terms of both Twitter accounts and external domains, even though it drastically decreases the number of external links in the timeline. In terms of topic diversity, algorithmic curation had a mixed effect, slightly amplifying a cluster of politically-focused tweets while squelching clusters of tweets focused on COVID-19 fatalities and health information. Finally, we present some evidence that the timeline algorithm may exacerbate partisan differences in exposure to different sources and topics. The paper concludes by discussing broader implications in the context of algorithmic gatekeeping.

Additional Key Words and Phrases: algorithm auditing, twitter, content ranking, social media

ACM Reference Format:
Jack Bandy and Nicholas Diakopoulos. 2021. More Accounts, Fewer Links: How Algorithmic Curation Impacts Media Exposure in Twitter Timelines. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 78 (April 2021), 28 pages. https://doi.org/10.1145/3449152

1 INTRODUCTION

As an algorithmic intermediary, Twitter's timeline curation system sorts, filters, and supplements personalized content for more than 150 million daily active users [32]. Yet despite its influential role in the flow of information across society, Twitter's curation algorithm generally remains a black box outside of Twitter's own explanations [51]: "our ranking algorithm is powered by deep neural networks... [that] predict whether a particular Tweet would be engaging to you." This opacity can be problematic for the public, especially since the system operates as an "algorithmic gatekeeper" [69] which can help set the agenda of public discourse analogously to traditional gatekeepers [6, 28, 80]. For example, amplifying particular tweets may push topics from the fringe of public discourse into the mainstream agenda [56].

The implications of algorithmic gatekeeping are further heightened during a public health emergency such as the COVID-19 pandemic. Health communication research has shown that different news information can change how people adopt protective measures [3, 34, 78] (such as wearing a mask or social distancing). A recent Pew survey underscored this point, finding

Authors' address: Jack Bandy, jackbandy@u.northwestern.edu; Nicholas Diakopoulos, nad@northwestern.edu, Northwestern University, Evanston, Illinois, USA.

2573-0142/2021/4-ART78 $15.00
https://doi.org/10.1145/3449152

that different media consumption patterns corresponded to different perceptions of COVID-19 risks, as well as differences in partisan affiliation [65]. Partisan differences may be exacerbated by social media feeds and their driving algorithms, which have often been embroiled with ideas of "echo chambers" and "filter bubbles" that may serve to reinforce existing viewpoints [7, 16, 73, 85]. Researchers have often called for countermeasures that deliberately increase the diversity of exposure to different sources and topics [13, 41, 76].

To address the opacity of Twitter's curation system and its potential echo chamber effects, this work aims to glimpse how algorithmic timeline curation affects media exposure on Twitter. We analyze how algorithmic curation compares to chronological curation in terms of its impact on *source diversity*, "the extent to which the media system is populated by a diverse array of content providers" [68], and *topic diversity*, the extent to which the timelines contain a diverse array of topics. More specifically, we ask the following research questions: **(RQ1)** How does Twitter's timeline curation algorithm affect (a) *source diversity*, with respect to Twitter accounts and external sources, and (b) *topic diversity*, with respect to different topic clusters? **(RQ2)** How does Twitter's timeline curation algorithm affect *partisan differences* in (a) source diversity and (b) topic diversity?

We address these questions through a sock-puppet algorithm audit [55, 79], creating eight "archetype puppets" (identified via network analysis) which follow the same accounts as real-world users, then collecting Twitter's algorithmic and chronological timelines for each of these puppets over the course of one month in early 2020. Using a number of different metrics, we present evidence from all eight accounts that Twitter's timeline curation algorithm substantially increases source diversity compared to the chronological baseline. On average, the algorithmic timeline featured more than twice as many unique accounts as the chronological timeline. It also reined in more dominant accounts, with the top ten accounts in the algorithmic timeline claiming fewer than half as many tweets (21% on average) as they did in the chronological timeline (45% on average). In terms of partisan differences, we found evidence that algorithmic curation may slightly exacerbate echo chambers, especially in terms of exposure rates to partisan Twitter accounts. Partisan differences in topic exposure did not exhibit these effects, even after algorithmic curation. Finally, we found that algorithmic curation dramatically decreased exposure to tweets with external links, from an average of 51% in chronological timelines to less than 20% in algorithmic timelines.

The paper makes two main contributions: (1) it introduces a network-based method for identifying "archetype puppets," strategically selecting real-world users to emulate in sock puppet audits; and (2) it applies this method to audit Twitter's timeline curation algorithm in terms of sources and content. While limited in sample size, we present experimental evidence from all eight accounts that, in comparison to a chronological baseline, Twitter's algorithmically curated timeline increases the diversity of sources, slightly exacerbates partisan echo chambers, shifts exposure to different news topics, and limits access to content outside the platform. These effects highlight the growing implications of algorithms acting as information gatekeepers in society. Traditionally, gatekeeping power has been held by journalists, who make decisions with the goal of "providing citizens the information they need to make the best decisions" [2]. But as DeVito [24] observed with the Facebook News Feed, a shift to algorithmic decision-making could represent a divergence from traditional editorial values. The paper concludes by discussing the implications of this divergence, as well as suggesting areas for future research to help address its potential negative impact.

2 BACKGROUND AND RELATED WORK

This study intersects with two main bodies of work: (1) audit studies that examine how algorithms affect the public, especially audits that focus on information curation algorithms, and (2) studies that explore how Twitter affects the production and distribution of news information.

2.1 Auditing News Curation Algorithms

News distribution has undergone rapid changes in recent decades, spurring research that revisits foundational theories of journalism such as gatekeeping, agenda-setting, and framing [70]. As Nechushtai and Lewis [71] put it, "prioritizing of news historically has been associated with human deliberations, not machine determinations." But today, machines play a definitive role in news curation. Whether audiences are visiting the website of their favorite newspaper, encountering news content on social media timelines, or searching for news on Google, their information diet is now mediated by algorithmic systems [26]. This has motivated audit studies aiming to better understand how algorithms influence the media ecosystem.

2.1.1 Audits of Partisan Differences. Some audit studies aim to determine whether gatekeeping algorithms – such as search algorithms and recommendation algorithms – personalize content to the point of creating "echo chambers" that silo users and limit news exposure diversity. These echo chambers form when a group is "isolated from the introduction of outside views," while "the views of its members are able to circulate widely," according to one synthesized definition by Bruns [16]. The phenomenon stems from well-documented patterns in media consumption related to social homophilies and selective exposure [13, 68], however, it may be exacerbated by curation algorithms. For example, a left-leaning user is likely to have many left-leaning friends on social media, and these friends are likely to share mostly left-leaning news articles [8]. A curation algorithm for the social media feed might cater to the user's ideology and increase exposure to left-leaning news articles, recognizing the user is less likely to engage with right-leaning news articles.

However, this algorithmic exacerbation of the echo chamber phenomenon often fails to materialize [16]. Focusing on Google search, an early study by Hannák et al. [39] found that on average only 11.7% of results were personalized, although personalization occurred more often in some query categories ("gadgets," "places," and "politics"). A study of 350 Dutch users found no substantial differences due to personalization [20] across 27 socio-political queries. Moving beyond the "blue links," an audit by Robertson et al. [77] measured personalization on all different components of a search engine, such as the "people also ask for" box, "tweet cards," and "knowledge components." This full-page audit "found negligible or non-significant differences between the SERPs" of personalized (logged in to Google) and non-personalized (incognito) windows. A recent audit found that even with partisan differences in query terms, Google search exhibited a mainstreaming effect, showing largely similar results for queries associated with left-leaning and right-leaning users [88]. Studies also find limited evidence for partisan echo chambers in recommender systems like Google News [38, 71] and the Facebook News Feed [8, 10], though Bakshy et al. [8] found that Facebook's algorithm slightly decreased exposure to ideologically cross-cutting content.

2.1.2 Audits of Curation Systems. Other studies focus less on personalization and explore more general trends in algorithmic news curation, such as source concentration, ideological bias, and churn rate. Trielli and Diakopoulos [87] conducted a targeted audit of the "Top Stories" carousel on Google's results page, finding a high concentration of sources (the top three sources – CNN, The New York Times, and The Washington Post – accounted for 23% of impressions). The substantial dependence on mainstream media outlets is evidenced in several studies of Google News [38, 71], Google search [20, 52, 77, 92], and Apple News [9], suggesting that algorithmic curation may exacerbate existing disparities in the media industry and exclude content from local and regional publications [30]. Some of these studies have also found small but statistically significant amplification of left-leaning news sources [77, 87], although the evidence is open to different interpretations. Finally, in terms of churn rate, Trielli and Diakopoulos [87] found that Google's

"Top Stories" carousel tends to "concentrate on articles that are more recent in age," though other studies found that "blue links" exhibit more stability over time [52, 64].

One audit study with particular relevance to our work is the "FeedVis" system developed by Eslami et al. [29], which presents users with a side-by-side comparison of their "unadulterated" Facebook News Feed with the algorithmically curated News Feed. Although the study focuses on algorithm awareness and user experience, it presents some methods for evaluating algorithmically curated timelines. For example, we adopt a similar framework in our methods for comparing the "unadulterated" chronological Twitter timeline with the algorithmically curated timeline.

2.1.3 Sock Puppet Audits. Audit studies in the social sciences have long faced experimental challenges related to resource constraints [33], often forcing experiments to utilize some kind of sampling. For example, an influential audit of employment discrimination in the United States only collected samples in Boston and Chicago [11]. In technical audits, sock-puppet auditing is one way of handling resource constraints and sampling challenges, allowing researchers to "use computer programs to impersonate users" [79]. As social media platforms further restrict access to APIs [15], expansive scraping audits become more challenging and sometimes impossible, and sock puppets provide a critical method for auditing platforms under these conditions.

Sock puppets have already proven to be an effective method for several technical audit studies of algorithmic curation. In one study conducted by Haim et al. [38], the authors created four sock puppets based on different life standards: (1) an elderly female conservative widow, (2) a bourgeois father in his fifties, (3) a 40-year-old job-oriented male single, and (4) a wealthy 30-year-old female marketing manager and early adopter. These agents were based on a media-user typology (MUT), which "aims to classify diverse user behaviours into meaningful categories of user types" [14]. The corresponding sock-puppet users experienced different personalized recommendations in their Google News accounts, though the effects were small.

Another exemplary sock puppet audit was conducted by Le et al. [55]. Based on Twitter accounts with opposing political views, the authors trained browser profiles to represent pro-immigration, anti-immigration, and control users, then collected search results for those users over the course of one week. Notably, this study created sock puppets to represent *ideological* views, specifically on the topic of immigration, rather than the demographic attributes (ex. gender and age) often used in related work [38, 39].

Our work builds on these previous algorithm audits in a number of ways. Specifically, we audit Twitter's timeline curation algorithm along many dimensions used in related work, anticipating the algorithm's effects on source diversity and partisan echo chambers, while adding a dimension of topic diversity. We also employ similar methods to related work, creating sock puppets to represent users with different ideological views, while also introducing a network-based approach to identify and simulate these archetypal users. In many ways, prior research on other platforms (e.g. Facebook [8] and Google [87]) suggest that algorithmic curation would have a negative impact, for example by introducing echo chamber effects and exacerbating existing inequalities in source concentration. Some work has already hinted at these effects on Twitter.

2.2 Twitter and News Information

The second body of related work generally relates to news information on Twitter. This research area is vast, and was even the subject of a recent special issue edition of *Journalism*, which addressed a broad range of topics including credibility perception, user experience, as well as celebrity news and gossip [99]. Studies that explore news content on Twitter have particular relevance for our work because they tend to account for concepts such as gatekeeping and source diversity, which we explore in our study.

2.2.1 News on Twitter. A number of surveys, interviews, and other studies demonstrate that "Twitter has become an important tool for journalists around the world" [91] since its founding in 2006. In fact, some have suggested that Twitter has fundamentally changed how the institution works, shifting the media ecosystem toward "ambient journalism" [42] characterized by an always-on news cycle. As of 2016, Weaver and Willnat [94] found that 54.8% of journalists reported regularly using microblog sites, predominantly Twitter. This usage impacts how journalists carry out their work. Based on a recent interview study in the United States by Peterson-Salahuddin and Diakopoulos [75], the algorithms that drive social media platforms "have become a new element influencing gatekeeping practices, especially with regards to content framing and resource allocation to stories," even though they do not dominate the editorial decision-making process.

While Twitter is in some ways impacting the practice of journalism, more importantly, journalism has a significant impact on Twitter. A 2010 study [53] found that over 85% of trending topics were related to news media, with mainstream news accounts most often benefiting from retweets compared to other types of accounts. The preponderance of news content on Twitter has been corroborated in a number of other studies (see Orellana-Rodriguez and Keane [72] for a survey of the literature). For example, Wu et al. [98] showed that even though celebrity accounts had more followers than news media accounts, news media accounts produced the most information. Also, according to a 2018 survey by Pew Research Center, "around seven-in-ten adult Twitter users in the U.S. get news on the site" [62], a higher rate than other social media platforms. Some studies have used these high-level findings to motivate more fine-grained analyses of news content on Twitter.

One of the most consistent findings from studies of news content on Twitter is a high degree of source concentration. An analysis of 1.8 billion tweets from 2014 found that "the top 10 handles alone account for 19 percent of news media-related activity," while the bottom "5603 of the 6103 handles account for only 5 percent of the volume of news media-related tweets" [61]. As reviewed by Orellana-Rodriguez and Keane [72], a number of studies have reproduced this pattern, with "a few tweets receiving a lot of attention and most tweets receiving no attention in a long tail." One reason for this may be Twitter's algorithmic timeline curation, which explicitly prioritizes "top tweets" based on engagement metrics such as likes, retweets, and replies [82].

2.2.2 Algorithmic Timeline Curation. Even when Twitter's algorithmic timeline was just a rumor in 2016, it attracted concerns from a variety of stakeholders. Some users espoused resistance to the changes, tweeting under the hashtag #RIPTwitter: "algorithms ruin everything," "we like chronological tweets," and more, as discussed in the content analysis by DeVito et al. [25]. The concerns about algorithmic curation persisted. As one journalist summarized, the algorithmic timeline "has been loathed and derided by users since it started" [19]. In spite of this, Twitter's timeline curation algorithm has become a cornerstone of the platform, and has allegedly attracted millions of additional users [50].

In addition to curating timelines algorithmically, Twitter has also *supplemented* users' timelines with tweets from users they do not follow, beginning as early as October 2014 [27]. This began experimentally, but is now a key feature of the platform. Twitter says it utilizes deep learning to make these recommendations [51], describing the feature as follows [89]:

> Additionally, when we identify a Tweet, an account to follow, or other content that's popular or relevant, we may add it to your timeline. This means you will sometimes see Tweets from accounts you don't follow. We select each Tweet using a variety of signals, including how popular it is and how people in your network are interacting with it. Our goal is to show you content on your Home timeline that you're most interested in and contributes to the conversation in a meaningful way, such as content that is relevant, credible, and safe.

While some recent work has explored news distribution on Twitter [36, 72, 83, 84], our study is the first to our knowledge that focuses on algorithmic curation and its specific effects. As others have noted [15, 54], studying the algorithmic timeline presents a number of challenges related to API access and user privacy, which we address with strategic sampling and a sock puppet setup. As in previous work, our analysis includes measures of source diversity due to potentially problematic source concentration on the platform. Given that algorithmic curation and "injected tweets" have become increasingly prominent on Twitter, we specifically analyze how these factors affect source concentration, source diversity, and partisan echo chambers, among other effects. By characterizing the effects of algorithmic curation and injected tweets, we hope to begin filling an important gap for understanding information distribution on Twitter.

3 METHODS AND DATA COLLECTION

To address our research questions, we followed a "sock puppet" audit method (as defined by Sandvig et al. [79]), whereby we identified a panel of users, simulated their interactions with the algorithmic Twitter timeline, and made comparisons to a baseline chronological timeline. While this audit method introduces some limitations (including a relatively small sample size similar to Haim et al. [38], Le et al. [55]), we deliberately chose it as an empirical and interpretable path to characterizing Twitter's curation algorithm. We also introduce a network-based approach for strategically identifying a set of representative users which we call "archetype puppets." By emulating representative real-world users, we intended this approach to sock puppet auditing would not only provide a "great deal of control over the manipulation and data collection" [79], but also a degree of ecological validity that often eludes algorithm audit studies. Notably, this ecological validity applies to the scoped context we chose for our study, which includes the particularities of U.S. politics as well as a distinctive news cycle amidst the COVID-19 pandemic. We elaborate specific implications of these limitations in section 6.1, as well as potential methods to mitigate them in future work.

This section first introduces the network-based method for measuring salient behavior within Twitter communities and identifying archetypal users from these communities (section 3.1). After identifying these users and creating archetype puppets, we collected their algorithmic and chronological timelines for a period of about four weeks (section 3.2). Finally, we analyzed how the puppets' algorithmically curated timelines differed from their chronological timelines (section 4).

3.1 Archetype Puppets

Since Twitter's policies constrained the number of puppet accounts we could create and emulate, we devised a method for selecting representative users for our study. We refer to the resulting accounts as "archetype puppets," and here present a process and rationale for creating these puppets. The general process requires the definition of one initial input (a pool of users) and then proceeds in three subsequent steps: detecting communities, selecting an archetype from each community, and validating the archetype selections. Future researchers may utilize and adapt this framework, outlined in Table 1, as a way of strategically sampling the space of potential sock puppets. Here we explain our specific choices for these steps and how they supported our research questions.

3.1.1 Initial Pool of Twitter Users. To identify an initial pool of Twitter users, we collected over 20 million accounts that followed U.S. congresspeople[1] in February 2020. We intended this pool to capture a large number of users engaged in U.S. politics and the accompanying news cycle, aligning with the kind of users needed to address our research question about partisan differences in news exposure. Wihbey et al. [96] used a similar approach to identify politically active users in

[1] https://github.com/unitedstates/congress-legislators/

Step in Process	Specific Choice of Method/Data
Step 1: Define Initial Pool of Users	All users following U.S. congresspeople
Step 2: Community Detection	Standard Louvain algorithm [12], yielding four left-leaning and four right-leaning communities
Step 3: Archetype Selection	Normalized degree centrality in community co-following network
Step 4. Archetype Validation	Archetype user bot score and friends list

Table 1. Our network-based archetype selection pipeline proceeds in four main steps, starting with an initial input of potential users.

their study of ideological patterns among journalists on Twitter. Still, this pool of users scopes our analysis to the United States, and even there it is not representative of all U.S. Twitter users.

Due to rate limits on Twitter's API, we took a random sample of 10,000 users to analyze from the initial pool of 20 million. We took this sampling step out of necessity, as collecting the friends[2] even for these 10,000 users took approximately one week under the Twitter API rate limits. More than 90% of users in this sample were public, leaving us with 9,020 users from which to detect communities and select archetypes.

3.1.2 Community Detection. The next step in our framework involves community detection to identify Twitter communities across the political spectrum. We started by assigning users a heuristic ideological lean score based on the ratio of republican and democratic congresspeople followed. We used a scale of [-1,1], similar to ideological scales in related work [8, 45, 96], with negative values representing left-leaning ideology and positive values representing right-leaning ideology:

$$\text{Lean} = \frac{-(\#\,Democrats\,Followed) + (\#\,Republicans\,Followed)}{\#\,Legislators\,Followed}$$

Then, we created two networks and assigned users to them based on lean score. These separate networks were intended to help address our research questions about partisan differences in exposure, thus, one network was created for users on the left (lean <= 0), and one for users on the right (lean >= 0). Users with a lean score of 0 were included in both networks to help identify more moderate communities. Still, there were more left-leaning users (N=7,217) than right-leaning users (N=2,055), reflecting the fact that people who use Twitter are more likely to be Democrats compared to the general public, and that Democratic legislators on Twitter have more followers on average [97].

After scoring and assigning users, we created network edges, weighted based on co-following behavior between users (i.e. two users are connected if they both follow the same user). This choice stemmed from our focus on exposure and the assumption that users who have similar friends will be exposed to similar tweets. Edge weights were calculated using the Jaccard coefficient of the connected user's friends list. Because it supports weighted edges by default and has been applied successfully in similar studies of Twitter communities, [37, 86] we ran the standard Louvain community detection algorithm [12], which identified four major communities on both sides. In the right-leaning network, 98.28% of users were assigned to the major communities, and in the left-leaning network, 99.79% of users were assigned to the major communities. While we did not anticipate that the Louvain algorithm would identify four major communities on each side, it conveniently provided a balance for the rest of our analysis. Further information about each community is shown in Table 2, suggesting meaningful distinctions across and within partisan lines. If this were not the case, we may have adjusted the modularity parameter in the Louvain algorithm to explore alternative clusterings of communities.

[2]While some papers [36, 93] have called these accounts "followees," in this paper we use Twitter's vernacular, so "friends" refers to *accounts followed by a user.*

	left-1	left-2	left-3	left-4	right-1	right-2	right-3	right-4
Member Count	1718	1801	1887	1811	869	357	460	369
Avg. Congresspeople Followed	1.58	6.26	2.03	2.3	2.84	1.61	6.75	2.88
Avg. Democrats Followed	1.53	5.6	1.78	1.93	0.78	0.28	0.88	0.21
Avg. Republicans Followed	0.05	0.63	0.23	0.36	2.05	1.32	5.86	2.67
Democrats:Republicans Ratio	29.78	8.84	7.63	5.39	0.38	0.21	0.15	0.08
Republicans:Democrats Ratio	0.03	0.11	0.13	0.19	2.63	4.77	6.64	12.82
Friends Count Q1	57	45	62	60	94	15	250	34
Friends Count Median	140	125	112	135	227	31	452	61
Friends Count Mean	337	456	382	471	551	45	1249	72
Friends Count Q3	396	404	332	475	591	60	909	102

Table 2. The eight major communities detected with the Louvain algorithm, along with descriptive attributes which align with previous findings from Pew Research [97]. For example, Pew found that Twitter users are more likely to be Democrats and that Democrats tend to follow more accounts, as reflected in larger member counts and higher average friend counts in communities on the left.

	left-1	left-2	left-3	left-4	right-1	right-2	right-3	right-4
Friends	580	430	510	120	440	60	690	100
Coverage	98%	96%	95%	99%	93%	96%	98%	97%
Followers	70	760	150	10	150	820	80	30
Statuses	1000	13700	600	100	7300	200	700	200
#1	@nytimes	@nytimes	@CNN	@nytimes	@realDonaldTrump	@realDonaldTrump	@BreitbartNews	@FoxNews
#2	@Reuters	@washingtonpost	@SportsCenter	@Reuters	@NASA	@POTUS	@steph93065	@realDonaldTrump
#3	@SenSanders	@Reuters	@TMZ	@guardian	@BarackObama	@VancityReynolds*	@Reuters	@benshapiro
#4	@washingtonpost	@AP	@chrissyteigen	@BBCWorld	@TheDailyShow	@GeorgesStPierre*	@JackPosobiec	@SebGorka
#5	@BernieSanders	@jaketapper	@ArianaGrande	@WSJ	@HistoryInPics	@UnderArmour*	@SebGorka	@WhiteHouse
#6	@TheOnion	@MalcolmNance	@espn	@Forbes	@SethMacFarlane	@DanCrenshawTX*	@realDailyWire	@charliekirk11
#7	@TheDailyShow	@JoyAnnReid	@UberFacts	@TheEconomist	@ConanOBrien	@RepDanCrenshaw*	@realDonaldTrump	@DLoesch
#8	@joerogan	@NPR	@wizkhalifa	@FT	@neiltyson	@HockeyCanada*	@nytimes	@DineshDSouza
#9	@SarahKSilverman	@SethAbramson	@khloekardashian	@cnnbrk	@prattprattpratt	@TRXtraining*	@cparham65	@NEWS_MAKER
#10	@ewarren	@FoxNews	@SenSanders	@TechCrunch	@tomhanks	@mriles4*	@mitchellvii	@marklevinshow
#11	@Sethrogen	@TeaPainUSA	@TheEllenShow	@UN	@BillNye	@NavalAcademy*	@washingtonpost	@RealJamesWoods
#12	@SethMacFarlane	@Forbes	@KevinHart4real	@NatGeo	@SteveCarell	@IMGAcademy*	@HuffPost	@ericbolling
#13	@StephenKing	@cnnbrk	@Diddy	@WIRED	@Xbox*	@NavyAthletics*	@dbongino	@BillOReilly
#14	@danieltosh	@RedTRaccoon	@NASA	@SenSanders	@NASAHubble*	@TGlass15*	@NEWS_MAKER	@RealCandaceO
#15	@mcuban*	@KeithOlbermann	@tylerthecreator	@ObamaWhiteHouse	@starwars*	@BodyRockTV*	@Rockprincess818	@SenateGOP
#16	@WhatTheFFacts*	@yashar	@TheWeirdWorld	@BarackObama	@lancearmstrong*	@ToddDurkin*	@CNN	@wikileaks
#17	@Metallica*	@soledadobrien	@BarackObama	@HarvardBiz	@foofighters*	@elliotthetrainr*	@politico	@AllenWest
#18	@OzzyOsbourne*	@tribelaw	@BernieSanders	@DeptofDefense	@kumailn*	@TRXTrainingCntr*	@WhiteHouse	@Jim_Jordan
#19	@chrisrock*	@AuschwitzMuseum	@iamcardib	@narendramodi	@AdamSandler*	@TPtherapy*	@ScottPresler	@RT_com*
#20	@thehill*	@BarackObama	@SnoopDogg	@JustinTrudeau	@johnkrasinski*	@djones5454*	@RT_com	@jordanbpeterson*

Table 3. Archetype account attributes, with counts rounded to the nearest 10 (Friends and Followers) or 100 (Statuses) to preserve anonymity. Below, community influencers followed by the archetypes, sorted by *average potential exposures per day* in the community, as detailed in the Methods section. Accounts denoted * are not community influencers, and are sorted by average potential exposures per day in all of Twitter. The archetype from the "right-2" community only followed two community influencers, otherwise, the archetypes tend to follow many top community influencers.

3.1.3 Archetype Selection. Next, we used network centrality to select one archetype user from each of the communities detected by the Louvain algorithm. There are many potential methods for defining and operationalizing archetype puppets, but for this work, we use the notion that *an archetypal user should follow many of the same accounts as other users in its community*. This characteristic is captured well in the edges of the co-following network, which connects users based on shared friends. Importantly, the network's edge weights are calculated using the Jaccard coefficient of two users' friends lists (the size of the intersection divided by the size of the union). By accounting for the total number of friends, the Jaccard coefficient helps ensure selection does not favor more active accounts, unless those accounts resemble others in the community. In short, higher normalized degree centrality within a community indicates a user follows more accounts which are also followed by other users in the community, and thus the user with the *highest* degree centrality aligns well with our notion of a community archetype.

	left-1	left-2	left-3	left-4	right-1	right-2	right-3	right-4
#1	@nytimes	@nytimes	@nytimes	@nytimes	@nytimes	@FoxNews	@FoxNews	@FoxNews
#2	@guardian	@HuffPost	@CNN	@Reuters	@Reuters	@realDonaldTrump	@realTT2020	@realDonaldTrump
#3	@HuffPost	@washingtonpost	@HuffPost	@HuffPost	@FoxNews	@nytimes	@JulieReichwein1	@benshapiro
#4	@CNN	@thehill	@ABC	@CNN	@CNN	@CNN	@ReneeCarrollAZ	@BreitbartNews
#5	@Reuters	@CNN	@NBA	@washingtonpost	@HuffPost	@VP	@KatTheHammer1	@dbongino
#6	@BBCWorld	@Reuters	@FoxNews	@guardian	@washingtonpost	@WhiteHouse	@wwwillstand	@SebGorka
#7	@SenSanders	@AP	@washingtonpost	@FoxNews	@BBCWorld	@RandPaul	@jjauthor	@IngrahamAngle
#8	@washingtonpost	@politico	@BBCWorld	@BBCWorld	@TIME	@POTUS	@BlueSea1964	@greta
#9	@AP	@jaketapper	@Reuters	@ABC	@ABC	@CoryBooker	@JoeFreedomLove	@WhiteHouse
#10	@realDonaldTrump	@guardian	@NFL	@business	@business	@NASA	@HLAurora63	@AnnCoulter
#11	@NPR	@maggieNYT	@AP	@TIME	@guardian	@tedcruz	@JanetTXBlessed	@nytimes
#12	@BernieSanders	@TheRickWilson	@SportsCenter	@AP	@AP	@SenTedCruz	@LindaSuhler	@DonaldJTrumpJr
#13	@TheOnion	@BBCWorld	@TIME	@WSJ	@WSJ	@marcorubio	@DallasBrownin16	@DailyCaller
#14	@NASA	@MalcolmNance	@MTV	@thehill	@Independent	@cnnbrk	@superyayadize	@charliekirk11
#15	@TheEconomist	@TIME	@billboard	@Independent	@Forbes	@NatGeo	@QTAnon1	@JudicialWatch
#16	@NatGeo	@JoyAnnReid	@WSJ	@Forbes	@businessinsider	@SpeakerRyan	@bbusa617	@DLoesch
#17	@GeorgeTakei	@WSJ	@TMZ	@TheEconomist	@thehill	@Interior	@GaetaSusan	@BretBaier
#18	@chrissyteigen	@ABC	@realDonaldTrump	@BBCNews	@NBCNews	@DeptofDefense	@AnnaApp91838450	@POTUS
#19	@rickygervais	@NPR	@chrissyteigen	@NBCNews	@CBSNews	@DonaldJTrumpJr	@9975Ts	@foxandfriends
#20	@BarackObama	@business	@NBCNews	@CBSNews	@NBA	@BarackObama	@inittowinit007	@WSJ

Table 4. The twenty most-influential users in each community, as determined by the average potential exposures per day within the community. Community influencers reflect previous findings regarding partisan news consumption [47, 48], for example, @FoxNews is the most-influential account for three of the four communities on the right.

3.1.4 Archetype Validation. To validate archetype puppet selection and evaluate our community detection, we identified influential users within each community. Based on a 2019 study by Pew Research [97], approximately 80% of all tweets in the United States come from an influential set of just 10% of Twitter accounts. We approximated influence by measuring the *average potential exposures per day* for popular accounts within each community. Other studies have used similar approximations to influence, alongside alternatives such as retweets and mentions [7, 18]. To measure average potential exposures per day, we first selected all accounts followed by at least 5% of community members, then calculated the average number of tweets sent per day by each account. Within a community, an account's average potential exposures per day is simply its average number of tweets per day times the number of community members who follow the account. The result approximates the number of times community members could see tweets from the account on an average day, which we used as a proxy for community influence.

Accounts with high potential exposures per day (Table 4) show a high degree of face validity, both in differences between left-leaning and right-leaning communities *and* differences between communities of the same partisan leaning. For instance, influential accounts in left-leaning communities included @nytimes, @HuffPost, and @CNN, news outlets with left-leaning audiences according to a 2020 survey by Pew Research [48]. The most influential account for three communities on the right was @FoxNews, which is the most-used and most-trusted news outlet among conservatives (based on the same survey [48]). The influencers also show differences between communities of the same partisan leaning. For example, in the left-3 and left-4 communities, members are more likely to follow republican congresspeople compared to left-1 and left-2 (see also Table 2), and accounts that have more influence in right-leaning communities (ex. @FoxNews and @realDonaldTrump) are among the top 20 influencers. Conversely, in the right-1 and right-2 communities, which are more likely to follow democratic congresspeople compared to right-3 and right-4, @nytimes and @BarackObama are among the top 20 influencers.

Finally, we used these community influencers to validate the selected archetypes. As shown in Table 3, the archetype accounts tend to follow many community influencers. We also validated archetypes by requiring a Botometer score [100] of less than 0.5, improving the chances of selecting non-automated users. The Botometer score was developed by Yang et al. [100], and uses Twitter metadata along with a number of derived features to detect social bots. While the authors note

several limitations and potential causes of misclassifications, here we simply use it to increase the likelihood that archetype users are real people. The 0.5 threshold has been suggested and applied effectively in prior work [4, 22].

With the archetypes selected and validated, we created one puppet account on Twitter for each archetype and requested to follow all of the archetype's friends. In doing so, we took several measures to balance the informativeness of our study with other ethical considerations such as user privacy and aggravation. The biographic text for each puppet account explained its function as a research bot, and provided contact information for the first author. Some private accounts did not accept the follow request, and several public accounts blocked our puppet account. On average, the puppets were able to follow over 96% of archetypes' friends. Table 3 includes the proportional coverage of friends we achieved for each individual puppet account. Setting up the puppet accounts was the final step before collecting data.

3.2 Data Collection for Timeline Analysis

We collected and analyzed data at three levels: (1) tweets in the chronological timeline, (2) tweets that appeared in the algorithmic timeline, and (3) tweets in the algorithmic timeline that were "injected." Data was collected each day beginning April 10th and ending May 11th, 2020. One important limitation of our data collection was its lack of dynamic interaction. While Twitter's timeline algorithm adjusts based on dynamic behavior such as engagement with tweets and web browsing history [89], our audit examines the Twitter algorithm "out of the box," without any personalization beyond the accounts followed.

3.2.1 Chronological Baseline. As a baseline with which to compare the algorithmic timeline, we scraped the chronological timeline for each puppet. As Twitter describes [89], "you can toggle between seeing the top Tweets first and the latest Tweets first" by selecting "see latest Tweets instead" from the home screen, which is the feature we used to collect chronological timelines.

We also explored a "potential from network" baseline, as in Bakshy et al. [8], collecting all tweets from the puppets' friends (via the Twitter API) and comparing them to tweets in the chronological timeline. Our measurements found that this baseline was essentially indistinguishable from the chronological baseline, so we focus on results comparing the chronological and algorithmic timelines. However, we still used the "potential from network" sample to identify "injected" tweets, as described in section 3.2.3.

To determine scraping frequency, we used metrics from a survey by Pew Research Center [97], in which 36% of Twitter users checked the platform "several times a day," 14% checked "once a day," 21% checked "a few times a week," and 30% checked less often. Since approximately 50% of users checked at least once per day, and Twitter usage reportedly increased amid the COVID-19 pandemic [60], we decided to schedule timeline scrapes twice per day: one at 9am and one at 9pm, Central Daylight Time (CDT). The scheduled scrapes ran for 30 days, with the first timeline collected at 9pm on April 10th, and the last collected at 9am on May 11th. During the 9pm collection on April 22nd, Twitter requested account verification for one puppet, causing the automated login to fail. This data point is thus excluded from all puppets in our analysis.

We also decided to collect tweets from the critical window of the first fifty items displayed in the timeline. This choice stems from the correlation between position and click-through rates, also know as "position bias" [5, 35]. Users engage more often with content positioned toward the top of a list, whether the list includes search results [23], advertisements [1], or social media content [8]. For example, in their study of news exposure on Facebook, Bakshy et al. [8] report click through rates around 20% for the first item in the News Feed, dropping to 5-6% for the 40th item. By collecting

and analyzing the first fifty tweets displayed, we focus our analysis on a critical window of content exposure.

To collect tweets from each puppet's timeline, we developed a scraping program implemented in Selenium and Python[3], which we ran on two Amazon EC2 instances located in Ohio. We reduced potential temporal variation by assigning four accounts to each server, and randomly walking through the accounts at each scheduled scrape (in the worst case, the last puppet's timeline was collected four minutes after the first puppet's).

3.2.2 Algorithmic Timeline. Algorithmic timelines, the focal point of this study, were collected using the same schedule and apparatus as the chronological timelines: we collected the first fifty tweets that appeared in the algorithmic timeline, and did so at 9am and 9pm CDT each day.

Algorithmic timelines are displayed by default in the "home" timeline, and the user interface refers to them as "Top Tweets." Twitter states that "Top Tweets are ones you are likely to care about most, and we choose them based on accounts you interact with most, Tweets you engage with, and much more" [89]. Twitter's blog has disclosed that Top Tweets are ranked with deep neural networks [51], internally referred to as "DeepBird" [57]. Again, these rankings are opaque and inaccessible via Twitter's API or other means, leading us to use sock puppet methods for our audit.

3.2.3 Injected Tweets. The third and final level of data collection comprised the "injected tweets" that appear in user timelines. While Twitter experimented with this feature as early as 2014 [27] and has touted its ability to attract new users [50], little is known about its behavior outside of this description [89] from Twitter:

> [W]hen we identify a Tweet, an account to follow, or other content that's popular or relevant, we may add it to your timeline. This means you will sometimes see Tweets from accounts you don't follow. We select each Tweet using a variety of signals, including how popular it is and how people in your network are interacting with it. Our goal is to show you content on your Home timeline that you're most interested in and contributes to the conversation in a meaningful way, such as content that is relevant, credible, and safe.

To identify injected tweets, we cross-referenced all statuses that appeared in the algorithmic timeline with a set of all tweets from a user's friends (i.e. the "potential from network" sample referenced above), using the unique identifier provided by Twitter. Any tweet from the algorithmic timeline that was not in this set was considered injected.

4 ANALYTIC FRAMEWORK

The puppets' aggregated algorithmic and chronological timelines comprised 39,352 unique tweets. We first conducted high-level analyses on these tweets to characterize the data, then proceeded with our main research questions. Our high-level analyses sought to clarify the prevalence of injected tweets, the makeup of tweets (i.e. external links, pictures, etc.), and the category of external links shared. In addressing our research questions, we apply the same analytical process to the left-leaning puppets and the right-leaning puppets. This symmetry allows us to make meaningful comparisons, while corroborating and accounting for some differences in behavior between left-leaning Twitter users and right-leaning Twitter users.

Some analyses, such as topic clustering, required canonical tweet details (i.e. full text, media type, etc.) from Twitter and/or external URL details (namely for shortened URLs). To collect canonical tweet details, we used the "status lookup" endpoint in Twitter's API. This provided details for 99.5% of tweets. The remaining tweets were either deleted by users or removed by Twitter. There were

[3] https://github.com/comp-journalism/twitter-timeline-scraper

more than 12,600 unique external URLs within the tweets, and we resolved each one to collect its final domain.

4.1 High-Level Patterns

We first conducted exploratory analysis to find high-level patterns in the data and inform analyses for the two main research questions. The exploratory analysis addressed three questions:

- What fraction of tweets in the algorithmic timeline are injected (i.e. not from friends)?
- What fraction of tweets contain external links and internal links?
- Which categories of domains are prominent in external links?

For domain categories, we used Comscore's category labels, which have proven useful in prior related work on news exposure [43, 95]. Comscore's labels include 29 unique categories, assigned to over 41,000 unique web domains that covered about 86% of unique links in the data.

4.2 RQ1: How does the curation algorithm impact timeline diversity?

To determine how algorithmic curation affects source diversity on Twitter (RQ1a), we analyzed Twitter accounts and web domains in the puppets' timelines. Measurements included the total number of unique sources within the timeline, the percentage of tweets observed from the most common sources, and the Gini coefficient to quantify and compare source inequality.

To analyze content diversity (RQ1b), we aimed to characterize the distribution of topics within different timelines. As in related work analyzing content diversity [10, 67] and twitter topics [66, 74, 93], we utilized topic modeling to extract common topics from the data.

Standard topic modeling techniques such as latent Dirichlet allocation (LDA) are often less coherent on short texts, due to sparse word co-occurrence [46]. As a solution, some researchers have proposed aggregating (or "pooling") all tweets from a user or hashtag into one document [44, 58, 63], rather than treating each tweet as a separate document. Another solution involves restrictions on the document-topic distribution, namely, using a Gibbs Sampling Dirichlet Multinomial Mixture (GSDMM) model [58, 101, 102] that assumes each text belongs to a single topic.

Since we aimed to characterize content at a fine granularity, we chose to use a GSDMM model to classify each individual tweet. In terms of preprocessing, we used a Twitter-specific tokenizer from the Natural Language Toolkit (NLTK) to preserve emoji and emoticons (as done in related work [25]). We also filtered out all URLs, removed stop words in NLTK's standard English list (e.g. "you", "them", "was", "as", etc.), and removed tokens that only occurred in one tweet. Following these steps, 97% of the tweets (38,061) contained at least one token and were included in topic modeling.

The GSDMM algorithm functions to automatically infer an appropriate number of clusters in the data. Experiments with ground truth labeled content have shown that GSDMM's inferences effectively balance completeness (all members of a ground truth group are assigned to the same cluster) and homogeneity (each cluster contains only members from one ground truth group) [101]. The inferred number of clusters is almost always smaller than an upper limit specified upon initialization (set to 300 here), which we also found to be true in our analysis. We performed 10 training iterations, and optimized hyper-parameters with a grid search on the alpha and beta values. The final model used an alpha of 0.3 and a beta of 0.1, from which GSDMM inferred 134 unique topic clusters. We manually validated the topic clusters by reading a sample of tweets within each cluster and ensuring they were meaningfully related within the cluster (i.e. completeness) and sufficiently independent from other clusters (i.e. homogeneity).

We measured the overall distribution of all topic clusters from the model, and also conducted a focused analysis on a subset of important clusters. Important clusters were manually selected based on the number of tweets that fell into it as well as the relevance of its topic to current events. Four

Label	Top Words	Example Tweet
Health	"coronavirus", "covid", "19", "new", "trump", "people", "health", "pandemic", "says", "us"	If you have diabetes, you are at higher risk for getting seriously ill from #coronavirus. Learn how you can take steps to protect yourself and slow the spread at (external_link) #COVID19
Economy	"%", "$", "coronavirus", "new", "million", "us", "pandemic", "people", "oil", "covid"	U.S. GDP falls by 4.8% in the first quarter of 2020 as the coronavirus pandemic continues to hit the economy.
Politics	"trump", "president", "people", "us", "coronavirus", "like", "china", "one", "would", "get"	We only have one president at a time. I continue to hold out hope that Donald Trump will make America #1 for testing per person, will pressure governors to meet White House guidelines before reopening their states, and will try to bring Americans together in this pandemic.
Fatalities	"coronavirus", "new", "cases", "covid", "19", "deaths", "death", "pandemic", "died", "people"	Mississippi's death toll rose to 11 early Monday, the state's emergency management agency tweeted. (external_link)

Table 5. Selected topic clusters that we focused on for analysis. Top words are words in the vocabulary that occurred most frequently among tweets within the cluster.

large clusters focused on COVID-19 news, which dominated the overall U.S. news cycle during data collection and thus provided a significant and important subject on which to focus the analysis. The four clusters emphasized (1) the economy, (2) health, (3) politics, and (4) fatalities. Each cluster represents a topic which concerned many U.S. citizens based on a Pew Research study during the COVID-19 pandemic [65]. The top words from each of these clusters, along with an exemplary tweet, are shown in Table 5.

4.3 RQ2: How does the curation algorithm impact partisan differences in sources and content?

Our second research question focused on the algorithm's effect on partisan differences, both in terms of sources and content in the timeline. This question addresses whether the algorithm has an "echo chamber" effect, operationalized as *increased exposure to partisan-specific sources and content*, based on the definition from Bruns [16]. This definition comes from a synthesis of many studies and position pieces which present wide-ranging definitions for what constitutes an echo chamber. Within our definition and using a framework of source diversity, an echo chamber effect is less severe if Twitter shows more accounts and/or domains to at least one left-leaning puppet and at least one right-leaning puppet.

To measure this, we first examined different sources (RQ2a) that puppets encountered in their timeline. For Twitter accounts, we used the community influencers identified via network analysis (as detailed in section 3.1, any account followed by more than 5% of community members was considered influential). Each account in our data was assigned one of five labels. If an account was influential in both a left-leaning and a right-leaning community, or if it ever appeared in at least one left-leaning and one right-leaning timeline, it was labeled "Bipartisan" (N=424). This notion of a "bipartisan" account is specific to our research focus on partisan differences in media consumption in the U.S. context, and differs from other notions of bipartisanship based on political ideology or voting records. Other influential accounts were labeled "Left" (N=215) or "Right" (N=382) according to the community. "Niche Left" (N=5,453) and Niche Right (N=4,586) accounts were not identified as influencers in the network analysis phase, and appeared exclusively to left-leaning puppets or right-leaning puppets, respectively. These niche accounts primarily comprised those that Twitter

injected into algorithmic timelines, as well as accounts from the archetypes' smaller networks, which had fewer followers. Table 6 shows accounts from each label.

Left Influencers	Niche Left	Bipartisan	Niche Right	Right Influencers
@chrislhayes	@BroadwayWorld	@BarackObama	@GordonGChang	@RealJamesWoods
@JoyAnnReid,	@ABCPolitics	@realDonaldTrump	@TeamTrump	@benshapiro
@tedlieu	@CTVKitchener	@BernieSanders	@BretBaier	@JudicialWatch
@PeteButtigieg	@realTuckFrumper	@WSJ	@brithume	@Jim_Jordan
@ananavarro	@itsJeffTiedrich	@FoxNews	@NEWS_MAKER	@BreitbartNews
@PreetBharara	@SBarlow_ROB	@HillaryClinton	@LouDobbs	@BillOReilly
@SenGillibrand	@kingmanmarie39	@TIME	@SkyNewsAust	@SenTedCruz
@Acosta	@wryly721	@Forbes	@ElevatedMonkey	@TomiLahren
@SenKamalaHarris	@jposhaughnessy	@elonmusk	@MrStache9	@GovMikeHuckabee
@maggieNYT	@BNNBloomberg	@UN	@MollerDennis	@DanCrenshawTX

Table 6. Accounts from the five partisan account labels.

To analyze partisan source differences in terms of news domains, we used data from Robertson et al. [77] that includes partisan audience scores for more than 19,000 websites. The scores were calculated by linking U.S. voter registration records to Twitter users, then scoring each web domain based on the registered political affiliation of users who shared the domain. The resulting scores exhibit high correlation with similar partisan classifications [8, 17], including nationally representative surveys from Pew Research [48]. For example, websites such as breitbart.com and foxnews.com receive right-leaning scores, while huffingtonpost.com and nytimes.com receive left-leaning scores.

We converted the audience scores to partisan labels for all domains contained in the Comscore news/information category, dividing the full range of scores (from maximum to minimum value) into five bins of equal range. The score range and four most-popular domains from each bin are shown in Table 7.

Most of our echo chamber analysis used exposure rates to partisan sources, however, we also looked at the phenomenon by tabulating sources that merely appeared to puppets of both partisan affiliations. Rather than measuring how *much* content from bipartisan sources appeared in the timelines, this analysis measures whether bipartisan sources appeared in the first place. In this case, the echo chamber effect becomes greater as fewer sources reach left-leaning and right-leaning timelines. A source was counted as having bipartisan appearances if it appeared in at least one left-leaning timeline and at least one right-leaning timeline.

Label	Audience is Mostly Left	Audience Leans Left	Audience is Moderate	Audience Leans Right	Audience is Mostly Right
Score Range	(-0.938, -0.556]	(-0.556,-0.177]	(-0.177, 0.203]	(0.203, 0.582]	(0.582, 0.962]
Domain #1	newyorker.com	nytimes.com	cnn.com	dailymail.co.uk	foxnews.com
Domain #2	nymag.com	washingtonpost.com	wsj.com	washingtonexaminer.com	the blaze.com
Domain #3	salon.com	huffingtonpost.com	usatoday.com	market-watch.com	breitbart.com
Domain #4	motherjones.com	npr.org	forbes.com	realclearpolitics.com	dailycaller.com

Table 7. Top domains from the five bins of partisan audience labels.

To analyze partisan differences in *content* (RQ2b), we used the same topic clusters identified for the initial analysis of content diversity. Again, the important clusters were manually selected based on the number of tweets that fell into each as well as the relevance of the cluster's topic

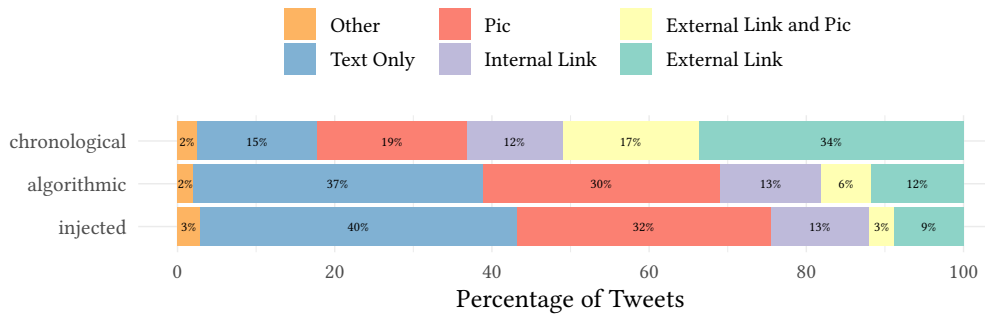

Fig. 1. Algorithmic curation decreased the rate of external links in the timeline, while increasing the rate of tweets containing internal media (namely pictures and links to other tweets). On average, only 18% of tweets in the algorithmic timeline contained an external link, compared to 51% in the chronological timeline. "Other" tweets contained mixed media, for example, an internal link and a picture.

to current events. The four clusters emphasized (1) the economy, (2) health, (3) politics, and (4) fatalities (see Table 5. Analyzing these topic clusters allowed us to glimpse whether there were partisan differences in the prominence of various news topics, and whether the algorithmic timeline increased or decreased any differences.

5 RESULTS

5.1 High-Level Patterns

The effect of algorithmic timeline curation was substantial even in high-level exploratory analyses, first in terms of injected tweets and second in terms of external links in the timeline. On average, injected tweets comprised 55% of the algorithmic timeline (min over eight puppets: 47%, max over eight puppets: 65%), meaning less than half of all tweets in the algorithmic timeline came from followed accounts. In terms of external media, 51% of tweets in the chronological timeline contained an external link (min: 31%, max: 81%), but only 18% of tweets in the algorithmic timeline contained an external link (min: 7%, max: 43%). This effect was consistent, with each of the eight puppets experienced a reduction in external links from the chronological timeline to the algorithmic timeline. The rate was even lower among injected tweets, at 12% (min: 4%, max: 26%). In contrast, text tweets with no media captured 15% of the average chronological timeline (min: 6%, max: 21%), increasing to 36% in the average algorithmic timeline (min: 26%, max: 54%). The rate of tweets with pictures also increased, from 19% for the chronological timeline (min: 8%, max: 29%) to 30% for the algorithmic timeline (min: 18%, max: 43%). Figure 1 visualizes the reduced exposure rate to external links in algorithmic timelines (the supplementary materials contain figures showing these effects for each individual puppet).

We then used the Comscore category labels to examine how different categories of websites were affected by algorithmic curation. Consistent with previous research analyzing Twitter content [53, 72], the most salient finding was the prevalence of news media. In the mean chronological timeline, 50% of links (min: 19%, max: 80%) directed to domains in Comscore's "News and Information" category, while in the algorithmic timeline the rate was 48% (min: 19%, max: 74%). That is, despite decreasing the fraction of tweets with external links, the algorithmic timeline generally maintained the distribution of domain categories. The slight changes were likely driven by injected tweets, of which a smaller percentage linked to "News and Information" domains (mean: 40%, min: 22%, max: 75%).

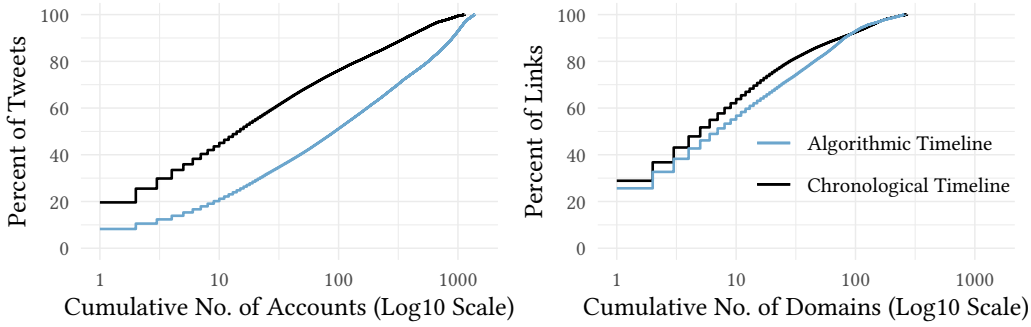

Fig. 2. Averaged Cumulative Distribution Functions (CDFs) to analyze source diversity in terms of accounts and domains in the timeline. **Left**: the algorithmic timeline (blue) distributes exposure more evenly across accounts, as indicated by its position beneath the chronological timeline (black). **Right**: the algorithmic timeline decreases the percentage of links claimed by top domains. Still, the domains exhibit high concentration, with the top ten claiming 64% in the chronological timeline, and 56% in the algorithmic timeline.

5.2 RQ1: Diversity

5.2.1 RQ1a: Source Diversity. Based on our data, Twitter's algorithmic timeline curation *increases* source diversity in a number of ways. For example, compared to the chronological timeline, the algorithmic timeline nearly doubled the number of unique accounts: across all eight puppets, there were an average of 663 unique accounts in the chronological timeline (min=379, median=662, max=1141), and an average of 1,169 unique accounts in the algorithmic timeline (min=638, median=1197, max=1388).

In addition to drastically increasing the number of accounts in the timeline, algorithmic curation also mitigated source concentration in terms of the average number of tweets per account observed. On average, the ten most common accounts in the chronological timeline made up 52% of tweets, but the ten most common accounts in the algorithmic timeline made up just 24% of tweets. The average Gini coefficient was 0.59 in the algorithmic timeline and 0.72 in the chronological timeline, indicating greater inequality of source exposure when the timeline was sorted chronologically. All eight puppets saw a lower Gini coefficient (i.e. less inequality) in their algorithmic timelines than in their chronological timelines, suggesting a consistent effect across these users' algorithmic timelines.

The increased exposure diversity disproportionately affected some specific accounts. For example, @CNN was a top account for the puppet from the left-3 community, claiming 10% of tweets in the chronological timeline. But when the puppet checked the algorithmic timeline, only 3% of tweets encountered were from @CNN. Similarly, @BroadwayWorld claimed 8% of tweets in the chronological timeline, and only 2% of tweets in the algorithmic timeline.

While some of these effects are more nuanced, the cumulative distribution function (Figure 2) shows a clear pattern of diversification: compared to the chronological timeline the algorithmic timeline features more unique accounts, and distributes more evenly across accounts.

Algorithmic curation had a similar but less pronounced effect on the domains of external links. The ten most common domains claimed 64% of all links in the chronological timeline, but only 56% in the algorithmic timeline. These rates also show that domains were more concentrated than accounts, as the top ten domains claimed a greater share than the top ten accounts. Figure 2 shows further evidence of this pattern.

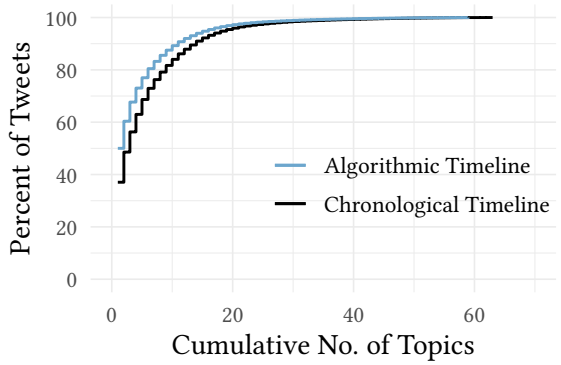

Fig. 3. CDF for the topic clusters from the GSDMM model, showing higher topic concentration in the algorithmic timeline. The most common topic was a generic cluster (common words: like, get, know, today, going, good).

5.2.2 RQ1b: Topic Diversity. Overall, algorithmic timelines exhibited slightly higher topic concentration compared to the chronological timelines. The most common topic (a generic cluster with common words including: like, get, know, today, going, good) accounted for 37% of tweets in the chronological timeline and 50% of tweets in the algorithmic timeline. The ten most common clusters in the chronological timeline made up 84% of the tweets, while the ten most common clusters in the algorithmic timeline made up 89% of the tweets. The pattern can be observed in Figure 3, and figures in the supplementary materials show the effect was consistent across all eight puppets.

Algorithmic timeline curation decreased the prevalence of the COVID-19 topics we examined, except for politics. On average, the four COVID-19 news clusters made up 39% of the chronological timeline (min=13%, max=53%), but only 32% of the algorithmic timeline (min=11%, max=49%). The average rate of fatality-focused tweets decreased from 5% (min=1%, max=8%) to 2% (min=1%, max=4%) in the algorithmic timeline, and the rate of health-related tweets also decreased from 12% (min=4%, max=18%) to 8% (min=3%, max=12%). The rate of politically-focused tweets increased slightly from 15% (min=4%, max=28%) to 17% (min=5%, max=34%). The average effects can be seen in Figure 4, and are further explored when measuring partisan differences.

5.3 RQ2: Partisan Differences

Our results show some partisan differences in exposure to Twitter accounts, news domains, and tweet topics, which the algorithm exacerbated in some ways. This "echo chamber" effect is defined and operationalized as *increased exposure to partisan-specific sources or content.* Similar to the study of selective exposure on Facebook by Bakshy et al. [8], the echo chamber effects we observe appear to stem primarily from accounts followed, and do not originate from algorithmic curation.

5.3.1 RQ2a: Source Differences. In terms of exposure to partisan accounts, the main effect of the algorithmic timeline was increased exposure to niche accounts that were exclusive to one partisan affiliation. Conversely, bipartisan accounts (which included bipartisan influencers and any account that appeared in at least one left-leaning and one right-leaning timeline) were *less* common in the algorithmic timeline (mean=18%, min=7%, max=33%) and injected tweets (mean=9%, min=6%, max=15%), compared to the chronological timeline (mean=31%, min=5%, max=59%). The overall effect is shown in Figure 5, and the reduced exposure to bipartisan and ideologically cross-cutting

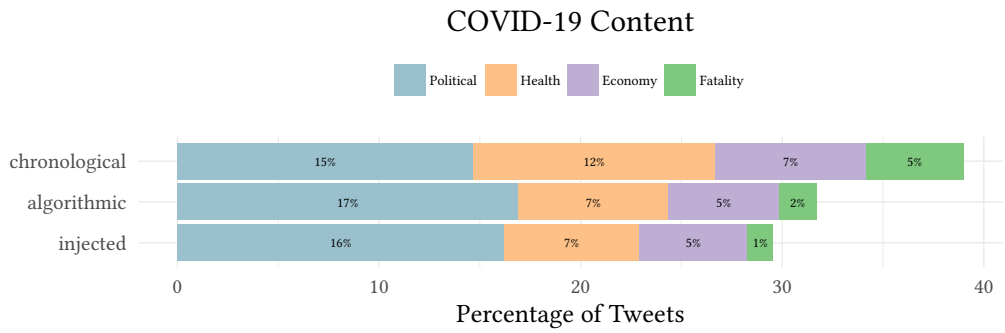

Fig. 4. Fraction of tweets from common topic clusters, based on the GSDMM topic model detailed in the methods section section 4.2. Overall, the topics studied are more prevalent in the chronological timelines, although tweets in the COVID-19 political cluster are slightly more prevalent in the algorithmic timelines.

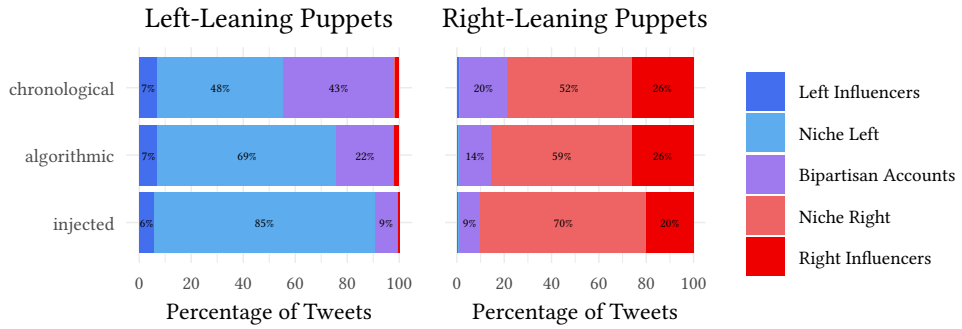

Fig. 5. The algorithmic timeline increased exposure to niche partisan accounts, decreased exposure to bipartisan accounts, and did not have a significant effect on exposure to more popular, influential accounts. The effect was consistent for all but one puppet.

accounts was consistent across seven of the eight puppets, the exception being the puppet from the "right-4" community. In that one case, which was the most right-leaning puppet based on our metrics, the algorithmic timeline *increased* exposure rates to bipartisan and ideologically cross-cutting accounts, from 5% in the chronological timeline to 9% in the algorithmic timeline.

The algorithm's effect on news domain exposure was more subtle. Notably, chronological timelines for both left-leaning and right leaning-puppets featured a sizable portion of news domains (a mean of 60% and 34%, respectively) considered to have "moderate" audiences, based on the classification detailed in section 4.3. Generally, news domains in the algorithmic timeline did not exhibit the echo chamber effects observed for Twitter accounts. However, news links in *injected* tweets contained a slightly higher rate of ideologically cross-cutting domains, compared to the chronological and algorithmic timelines. Also injected tweets contained a slightly higher rate of ideologically consistent domains when excluding "leaning" domains (i.e. including only "Left Domains" for left-leaning puppets, and only "Right Domains" for right-leaning puppets). Note these results apply only to the eight puppet accounts examined in the study and may not generalize across the platform. Figure 6 depicts these findings visually.

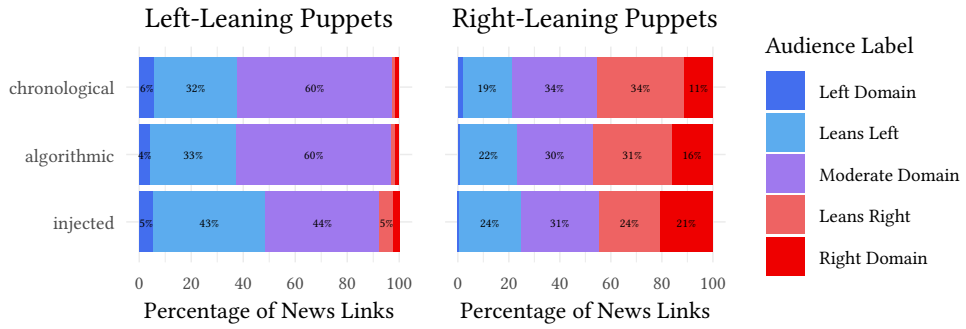

Fig. 6. Partisan differences in news domains, using partisan audience scores from Robertson et al. [77]. The timeline curation algorithm had less of an echo chamber effect on news domains than it did on Twitter accounts. The rate of partisan domains among injected tweets was slightly different, increasing ideologically cross-cutting domains in some cases (more "Leans Right" domains for left-leaning puppets and more "Leans Left" domains for right-leaning puppets), and increasing ideologically consistent domains in other cases (more "Leans Left" domains for left-leaning puppets).

There is also a notable asymmetry in chronological baselines along partisan lines. In the average chronological timeline for left-leaning puppets, right and right-leaning news domains claimed less than 3% of news links. But for right-leaning puppets, 21% of all news links were from left and left-leaning news domains in the chronological timeline. While limited to the eight puppets in our study, these results may corroborate previous findings that right-leaning users are exposed to more ideologically cross-cutting sources than left-leaning users [8]. This could be related to the right's distrust toward mainstream media organizations [48] and tendency to share these sources for critique, as well as greater underlying volume of sources with left-leaning audiences, as suggested by previous studies [77, 87].

The partisan echo chambers can also be observed by analyzing which sources had bipartisan appearances, reaching both left-leaning and right-leaning timelines. Of all 8,436 unique accounts in the puppets' algorithmic timelines, only 3% (N=287) appeared in at least one left-leaning timeline and at least one right-leaning timeline, including some popular accounts that tweet in high volumes (e.g. @Reuters, @CNN, @realDonaldTrump, @espn). The remaining 97% of accounts in the dataset represent a significant partisan chasm in the accounts that appeared in the timelines, despite a slight improvement compared to the chronological timelines (2% bipartisan accounts, N=110). Furthermore, only 10% (N=97) of all unique domains in the algorithmic timelines appeared to at least one puppet from both partisan affiliations. As with the accounts, the domains included popular, high-volume websites such as youtube.com, cnn.com, instagram.com, nytimes.com, and reuters.com. In the case of domains, the algorithm exacerbated the echo chamber effect in the chronological timelines, where a greater percentage (12%) and a greater number of sources (N=146) made appearances to at least one puppet from each of the partisan affiliations.

5.3.2 RQ2b: Topic Differences. Finally, moving to partisan differences in topics (RQ2b), we found a mixed effect: the algorithmic timeline amplified some topics and squelched others. For puppets on both sides, political COVID-19 content was the most common topic cluster in chronological timelines, accounting for 13% of tweets for left-leaning puppets and 16% of tweets for right-leaning puppets (overall mean=15%, min=4%, max=28%). But in the algorithmic timeline, right-leaning

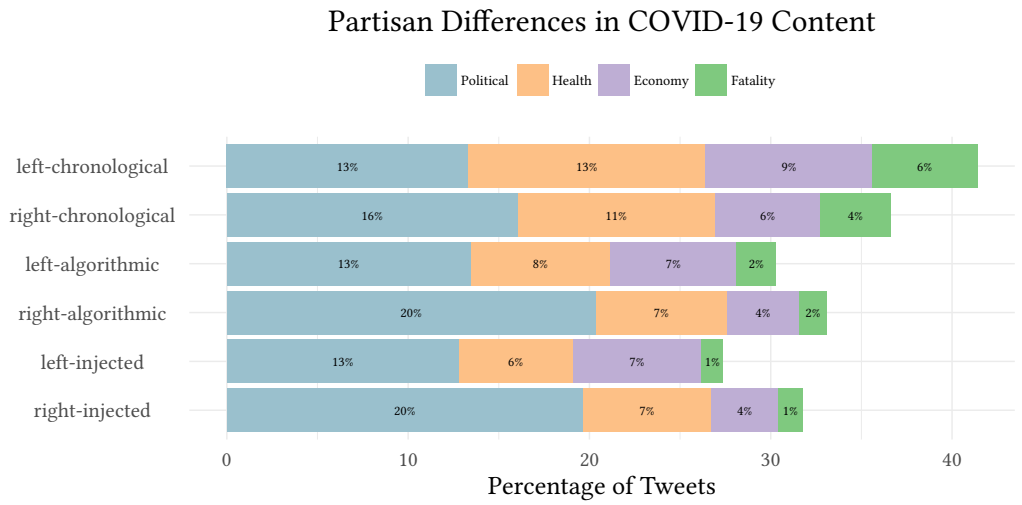

Fig. 7. Partisan differences in tweet clusters related to COVID-19 topics, identified via topic modeling. Puppets' algorithmic timelines included less COVID-19 content related to fatalities, health, and the economy, regardless of political leaning.

puppets saw tweets from this cluster even more often (20%), even though the rate remained at 13% for left-leaning puppets (overall mean=17%, min=5%, max=34%).

For the health, fatality, and economy clusters, the effect of algorithmic curation had a similar reduction effect for both left-leaning and right-leaning puppets. In other words, puppets' algorithmic timelines included less COVID-19 content related to fatalities, health, and the economy, regardless of political leaning. One notable exception is that right-leaning algorithmic timelines saw tweets in the fatality cluster at roughly the same rate as left-leaning algorithmic timelines (overall mean=1%, min=1%, max=4%), even though this cluster was more common for left-leaning puppets in the chronological timeline (6% for left-leaning puppets compared to 4% for right-leaning puppets). See Figure 7.

6 DISCUSSION

This work sought to characterize the effects of algorithmic curation on source and content diversity in the Twitter timeline. Notably, we found evidence that the algorithmic timeline substantially increases the number of unique accounts that a user sees in the timeline, and decreases the dominance of top accounts. In our data, the algorithm also exacerbated partisan echo chambers in terms of accounts and domains, while having a mixed effect on content echo chambers. Before discussing the implications of these findings, we first note some important limitations.

6.1 Limitations

As with many technical audits, we faced several constraints and limitations in our study related to scale and ecological validity. First, the initial input to our archetype selection pipeline only included users who follow U.S. congresspeople, which intentionally scoped the study to U.S. politics, but also may have excluded people who engage differently with the U.S. political system and the accompanying news cycle. Also, due to Twitter's policies, we analyzed just eight different accounts, which is a relatively small sample size and yet still yielded notable variation in some metrics. The

main effects we observed—increased source diversity, partisan amplification, and topic shifts—were consistent across puppets, although the size of these effects could have been more precise with a larger sample size and/or different accounts. Finally, we also collected data at a time with unusual news dynamics due to the COVID-19 pandemic. Before the pandemic, topic clusters would have likely exhibited a different profile, although clusters such as political news may be relatively stable. Future work may further explore our findings using parts of our data collection apparatus, for instance, collecting data from additional puppets, real-world users, and/or collect data during a different time period with different news dynamics.

Furthermore, our study did not account for personalization and dynamic user activity. In real-world use, Twitter's timeline algorithm adjusts based on personalized attributes such as engagement with tweets and web browsing history [89]. Since we did not engage with any tweets (e.g. clicking, responding to, or retweeting) and collected data on remote servers, our study does not address the effects of this personalization. We also limited data collection to fifty tweets at 9am and 9pm each day, while real-world usage would be more dynamic.

Our study introduced a network-based approach to identifying representative sock puppets, which we refer to as "archetype puppets." However, we selected archetype puppets based on a specific operationalization that used co-following behavior and degree centrality due to our focus on exposure. Future work may benefit from alternative selection methods. For example, some selection methods may not require network analysis, and select "typical" accounts based on friend or activity patterns. Alternative network-based approaches could select archetype accounts with friend networks, retweet networks, like networks, or some combination of these, rather than the co-following network we used in this work. Selecting different archetype users could lead to greater variation in results, especially given that our archetype selection process selected from just over 9,000 accounts compared to the 150 million daily active accounts reported by Twitter [32].

One final limitation to note is that while we found some evidence that the curation algorithm has echo chamber effects, we used a specific operationalization of these effects in our analysis. We identified hundreds of partisan influencers through a large-scale network analysis, but thousands of "niche" accounts were labeled heuristically, based on their appearance in the eight puppets' timelines. We also do not account for exposure diversity that may arise due to engagement via other platforms, which is a common limitation in related work [16]. Future work could use alternative approaches for scoring and labeling partisanship of Twitter accounts, and may benefit from studying real-world users who encounter media on multiple platforms.

6.2 Potential Impacts on Journalism

One of the largest effects of algorithmic curation we observed in our timeline was a decrease in the fraction of tweets with external links, from 51% in the average chronological timeline to 18% in the average algorithmic timeline. Of those links, roughly half came from the "News and Information" category defined by Comscore. This observation is concerning for journalism organizations hoping to drive website traffic via Twitter, and for users who aim to use the platform to discover news content from those organizations.

Journalists often use Twitter for more than just sharing links to news stories. For example, many journalists communicate with colleagues and crowdsource news tips on the platform [21], which may be largely unaffected by algorithmic curation. Other journalistic uses, such as live-tweet breaking events, may even benefit from the increased algorithmic distribution of text-only and text+pic tweets we observed. On the flip side, algorithmic distribution of unvetted text-only or text+pic tweets in a breaking news context could also increase the risk of spreading unverified information from non-journalists, potentially exacerbating challenges to journalists related to monitoring and correcting misinformation. Based on our findings, we can say somewhat more

definitively that algorithmic curation does impact the common and important use case of sharing external links to journalistic media. Despite the preponderance of such news links on Twitter (approximately one in four tweets in the chronological timelines had links directing to domains in Comscore's "News and Information" category), the curation algorithm seems to substantially undermine their distribution in the timeline. To support the journalism industry as well as news discovery, future work might seek ways to mitigate such effects.

We posit two potential explanations for this observation. First, since Twitter prioritizes engagement, it is possible that tweets with external links simply do not generate as much engagement and thus do not rank well in algorithmic curation. For example, if users visit external links and do not return to Twitter to like or comment on them, then tweets with external links may accumulate less engagement, and would be less likely to rank as "top tweets" prioritized by the algorithmic timeline.

A second potential explanation is that Twitter's deep learning algorithm optimizes on-site time, and suppresses external links that take users away from the platform. Twitter's own documentation states their algorithm optimizes user engagement, and they have touted its ability to attract users to the platform [50]. In light of this, their systems may have gradually learned to keep users on the platform by showing fewer tweets that could send them elsewhere.

Whatever the cause, our evidence highlights how algorithmic values now play a gatekeeping role in curating news information and other content, with substantial effects. As detailed by DeVito [24] in a study of the Facebook News Feed, these algorithmic values tend to diverge from traditional news values such as importance, impact, and timeliness. Instead, they rely on a variety of optimization metrics that directly impact the information encountered by end users, sometimes with unintended consequences [26]. The timelines in our study further show that algorithmic curation has complex effects, such as increasing "incidental exposure" [31] (for example, the number of unique accounts that users see), yet exacerbating partisan differences in source exposure. Further exploring and clarifying these effects may help journalists effectively use the platform for distributing news.

6.3 Algorithmic Curation and Partisan Echo Chambers

While we found that the algorithmic timeline increased the overall diversity of sources for user timelines, we did not find a mainstreaming effect in the sources. Instead, we found that partisan accounts made up the vast majority of the algorithmic timeline and injected tweets, and tended to exacerbate baseline differences in the chronological timeline. Only 3% of all unique sources in the algorithmic timelines appeared to at least one left-leaning puppet and at least one right-leaning puppet, though these sources were popular and accounted for 18% of all tweets in the algorithmic timelines.

The echo chamber effect was somewhat surprising. Based on Twitter's description [89], a tweet is added to a timeline based on "how popular it is and how people in your network are interacting with it." If injected tweets were based on popularity alone, one would expect them to be added to timelines across partisan lines, potentially mitigating echo chamber effects. But given that injected tweets were largely from niche accounts, and 97% of all accounts did *not* appear to both partisan affiliations, "how people in your network are interacting" with a tweet is likely more of a determinant than overall popularity.

Adding tweets to user timelines is an especially opportune mechanism to mitigate the potential negative impacts of social media platforms. Added content could be curated using more traditional editorial values, in service to both journalists and Twitter users. Importantly, this would involve keeping users simultaneously engaged, informed, and connected to shared experiences in their cultural context, rather than maximizing engagement and on-site time [40, 49, 59]. In some settings this curation may include more traditional news information that provides users with important,

timely, non-partisan updates, though even showing internet memes could serve to create shared cultural experiences [81]. Considering this potential impact brings us to the broader implications of our findings.

6.4 Broader Implications and Future Work

Our results intersect with a broad conversation and research agenda surrounding the gatekeeping power exercised by social media curation algorithms. This topic is often discussed in terms of polarization, filter bubbles, echo chambers, selective exposure, and/or partisan bias [8, 10, 76]. While we did find that the curation algorithm exacerbated these effects in some ways, it also had notable effects on the accounts, links, and topics that appeared in the timeline. Future work should test whether these findings generalize to other locations, social contexts, and time periods. One way to test the algorithm in a different context is to use a different set of input users in the puppet selection pipeline. For example, a study might start by collecting Twitter users who report a specific location or who follow certain accounts. Collecting data at another time, and potentially for a longer time period, could test whether the results apply during other news cycles. Going a different route, future research could also audit algorithmic timelines with real-world users via crowdsourcing. Still, some findings from our study already have important implications for researchers, including the prevalence of "injected tweets," the reduced exposure to external links, and the shift in topics.

First, given that more than half of all tweets in the puppets' algorithmic timelines came from accounts they did not follow, future research should consider these "injected tweets" an integral part of the platform that affects exposure just as much as the accounts users actually follow. For example, measuring exposures for users as "any tweet shared by one of their followees [i.e. friends]" [37] is likely no longer an accurate representation of the tweets users see in their real-world timelines. Future research might explore injected tweets with more scrutiny to determine their network effects and their effect on user experience, beyond the increased retention reported by Twitter.

Second, the reduced exposure to external links is particularly troubling for news organizations and news-seeking users. More broadly, it is also concerning for any content producer aiming to use Twitter for distribution, and any user wanting to discover content outside of Twitter. In terms of gatekeeping power, the timeline curation algorithm appears to dramatically increase exposure to content *within* the platform, a behavior which deserves critical attention from researchers. Exploring the exact cause of this phenomenon is an especially promising area for future research.

Finally, the algorithm's impact on topics within the timeline carries implications for media exposure and news ecology at large, particularly during a public health emergency such as the COVID-19 pandemic. We found that the algorithmic timeline amplified exposure to tweets in a politically-focused topic cluster, while decreasing exposure to tweets in fatality-focused and health-focused topic clusters. This means that echo chamber effects can be observed not only in amplifying partisan sources, but also in shifting exposure to specific topics. The reason for this could be that tweets about COVID-19 fatalities and health information generated less engagement: the median tweet in the political cluster received 247 retweets, far more than the median tweet from the health cluster (73) or the fatality cluster (54). However, given the opacity of Twitter's curation algorithms, this is only a preliminary hypothesis, and the exact cause for the difference remains unknown.

Regardless of the cause, the stakes are high when it comes to curation algorithms amplifying some topics and burying others. During the COVID-19 pandemic, informative tweets about fatality rates and preventative measures are often more important than tweets related to political events. This is true regardless of the engagement these tweets attract. With Twitter now relying on human curation for some features such as "Twitter Moments" [90], it remains to be seen whether the

platform's overall curation strategy can shift toward more traditional news values with more transparency.

7 CONCLUSION

This work makes two contributions, first by introducing a network-based method for identifying "archetype puppets," and second by applying this method to audit Twitter's timeline curation algorithm in terms of source diversity, topic diversity, and partisan echo chamber effects. Based on evidence from a sample of eight sock puppets in the United States, salient effects of the algorithm include a decrease in the number of external links, an increase in source exposure diversity (in terms of Twitter accounts), and a slight shift in topic exposure (including reduced exposure to clusters of tweets about COVID-19 fatalities and health information). Findings also suggest that algorithmic curation may have a slight partisan echo chamber effect rather than a mainstreaming effect, in terms of exposure rates to partisan sources. The findings highlight how social media curation algorithms may diverge from traditional curation values used in journalism (e.g. [2]), and contribute to an important ongoing discussion about algorithmic gatekeeping and its implications for the public.

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation Grant, Award IIS-1717330.

REFERENCES

[1] Ashish Agarwal, Kartik Hosanagar, and Michael D. Smith. 2011. Location, location, location: An analysis of profitability of position in online advertising markets. *Journal of Marketing Research* 48, 6 (dec 2011), 1057–1073. https://doi.org/10.1509/jmr.08.0468

[2] American Press Institute. 2020. What is the purpose of journalism? https://www.americanpressinstitute.org/journalism-essentials/what-is-journalism/purpose-journalism/

[3] Anne Marie Apanovitch, Danielle McCarthy, and Peter Salovey. 2003. Using message framing to motivate HIV testing among low-income, ethnic minority women. *Health Psychology* 22, 1 (2003), 60–67. https://doi.org/10.1037/0278-6133.22.1.60

[4] Adam Badawy, Kristina Lerman, and Emilio Ferrara. 2019. Who falls for online political manipulation?. In *The Web Conference 2019 - Companion of the World Wide Web Conference, WWW 2019*. Association for Computing Machinery, Inc, New York, New York, USA, 162–168. https://doi.org/10.1145/3308560.3316494

[5] Ricardo Baeza-Yates. 2018. Bias on the web. *Commun. ACM* 61, 6 (2018), 54–61. https://doi.org/10.1145/3209581

[6] Ben H Bagdikian. 1983. *The media monopoly*. Boston.

[7] Eytan Bakshy, Winter A. Mason, Jake M. Hofman, and Duncan J. Watts. 2011. Everyone's an influencer: Quantifying influence on twitter. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining, WSDM 2011*. 65–74. https://doi.org/10.1145/1935826.1935845

[8] Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (jun 2015), 1130–1132. https://doi.org/10.1126/science.aaa1160

[9] Jack Bandy and Nicholas Diakopoulos. 2020. Auditing News Curation Systems: A Case Study Examining Algorithmic and Editorial Logic in Apple News. *International conference on web and social media (ICWSM)* (2020).

[10] Anja Bechmann and Kristoffer L. Nielbo. 2018. Are We Exposed to the Same "News" in the News Feed?: An empirical analysis of filter bubbles as information similarity for Danish Facebook users. *Digital Journalism* 6, 8 (2018), 990–1002. https://doi.org/10.1080/21670811.2018.1510741

[11] Marianne Bertrand and Sendhil Mullainathan. 2003. Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. (2003). https://doi.org/10.2139/ssrn.422902

[12] Vincent D Blondel, Jean Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008, 10 (oct 2008), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

[13] Engin Bozdag and Jeroen van den Hoven. 2015. Breaking the filter bubble: democracy and design. *Ethics and Information Technology* 17, 4 (2015), 249–265. https://doi.org/10.1007/s10676-015-9380-y

[14] Petter Bae Brandtzæg. 2010. Towards a unified Media-User Typology (MUT): A meta-analysis and review of the research literature on media-user typologies. *Computers in Human Behavior* 26, 5 (sep 2010), 940–956. https:

//doi.org/10.1016/j.chb.2010.02.008
[15] Axel Bruns. 2019. After the 'APIcalypse': social media platforms and their fight against critical scholarly research. *Information Communication and Society* 22, 11 (2019), 1544–1566. https://doi.org/10.1080/1369118X.2019.1637447
[16] Axel Bruns. 2019. *Are Filter Bubbles Real?* John Wiley & Sons.
[17] Ceren Budak, Sharad Goel, and Justin M. Rao. 2016. Fair and balanced? Quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly* 80, Specialissue1 (2016), 250–271. https://doi.org/10.1093/poq/nfw007
[18] Meeyoung Cha, Hamed Haddadi, Fabrício Benevenuto, and Krishna P Gummadi. 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. In *ICWSM*.
[19] Samantha Cole. 2020. How to Take Your Twitter Feed Back From the Algorithm. https://www.vice.com/en{_}us/article/7kz9ez/go-into-2020-by-taking-your-twitter-feed-back-from-the-algorithms
[20] Cédric Courtois, Laura Slechten, and Lennert Coenen. 2018. Challenging Google Search filter bubbles in social and political information: Disconforming evidence from a digital methods case study. *Telematics and Informatics* 35, 7 (oct 2018), 2006–2015. https://doi.org/10.1016/j.tele.2018.07.004
[21] Dharma Dailey and Kate Starbird. 2014. Journalists as Crowdsourcerers: Responding to Crisis by Reporting with a Crowd. *Computer Supported Cooperative Work: CSCW: An International Journal* 23, 4-6 (2014), 445–481. https://doi.org/10.1007/s10606-014-9208-z
[22] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. BotOrNot. In *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion*. Association for Computing Machinery (ACM), New York, New York, USA, 273–274. https://doi.org/10.1145/2872518.2889302
[23] Thorsten Joachims Dept, Computer Science, Laura Granka Google Inc, Helene Hembrooke Dept, Information Science, Filip Radlinski Dept, and Geri G A Y Dept. 2007. Evaluating the Accuracy of Implicit Feedback from Clicks and Query Reformulations in Web Search. *ACM Transactions on Information Systems* 25, 2 (2007). https://dl.acm.org/doi/abs/10.1145/1229179.1229181
[24] Michael A. DeVito. 2017. From Editors to Algorithms: A values-based approach to understanding story selection in the Facebook news feed. *Digital Journalism* 5, 6 (jul 2017), 753–773. https://doi.org/10.1080/21670811.2016.1178592
[25] Michael A. DeVito, Darren Gergle, and Jeremy Birnholtz. 2017. "Algorithms ruin everything". In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3025453.3025659
[26] Nicholas Diakopoulos. 2019. *Automating the News: How Algorithms Are Rewriting the Media*. Harvard University Press.
[27] Stuart Dredge. 2014. Yes, Twitter is putting tweets in your timeline from people you don't follow. https://www.theguardian.com/technology/2014/oct/17/twitter-tweets-timeline-dont-follow
[28] Robert Entman. 1993. Framing - Toward Clarification of a Fractured Paradigm. *Journal of Communication* 43, 4 (1993), 51–58. https://doi.org/10.1111/j.1460-2466.1993.tb01304.x
[29] Motahhare Eslami, Amirhossein Aleyasen, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. FeedVis: A Path for Exploring News Feed Curation Algorithms. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing - CSCW'15 Companion*. 65–68. https://doi.org/10.1145/2685553.2702690
[30] Sean Fischer, Kokil Jaidka, and Yphtach Lelkes. 2020. Auditing local news presence on Google News. *Nature Human Behaviour* (sep 2020). https://doi.org/10.1038/s41562-020-00954-0
[31] Richard Fletcher and Rasmus Kleis Nielsen. 2018. Are people incidentally exposed to news on social media? A comparative analysis. *New Media and Society* 20, 7 (2018), 2450–2468. https://doi.org/10.1177/1461444817724170
[32] Agence France-Presse. 2020. Twitter shares surge on millions new users. https://tribune.net.ph/index.php/2020/02/07/twitter-shares-surge-on-millions-new-users/
[33] S Michael Gaddis. 2017. *An Introduction to Audit Studies in the Social Sciences*. Technical Report. www.auditstudies.com
[34] Mary A. Gerend and Jon K. Maner. 2011. Fear, Anger, Fruits, and Veggies: Interactive Effects of Emotion and Message Framing on Health Behavior. *Health Psychology* 30, 4 (jul 2011), 420–423. https://doi.org/10.1037/a0021981
[35] Laura A. Granka, Thorsten Joachims, and Geri Gay. 2004. Eye-tracking analysis of user behavior in WWW search. In *Proceedings of Sheffield SIGIR - Twenty-Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, New York, USA, 478–479. https://doi.org/10.1145/1008992.1009079
[36] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Political science: Fake news on Twitter during the 2016 U.S. presidential election. *Science* 363, 6425 (2019), 374–378. https://doi.org/10.1126/science.aau2706
[37] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Supplementary Materials for Fake news on Twitter during the 2016 U.S. presidential election. *Science* 363, 6425 (2019), 374–378. https://doi.org/10.1126/science.aau2706

[38] Mario Haim, Andreas Graefe, and Hans Bernd Brosius. 2018. Burst of the Filter Bubble?: Effects of personalization on the diversity of Google News. *Digital Journalism* 6, 3 (2018), 330–343. https://doi.org/10.1080/21670811.2017.1338145

[39] Anikó Hannák, Piotr Sapieżyński, Arash Molavi Khaki, David Lazer, Alan Mislove, and Christo Wilson. 2013. Measuring Personalization of Web Search. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 527—-538. https://doi.org/10.1145/2488388.2488435

[40] Natali Helberger. 2019. On the Democratic Role of News Recommenders. *Digital Journalism* 7, 8 (sep 2019), 993–1012. https://doi.org/10.1080/21670811.2019.1623700

[41] Natali Helberger, Kari Karppinen, and Lucia D'Acunto. 2018. Exposure diversity as a design principle for recommender systems. *Information Communication and Society* 21, 2 (2018), 191–207. https://doi.org/10.1080/1369118X.2016.1271900

[42] Alfred Hermida. 2010. Twittering the news: The emergence of ambient journalism. *Journalism Practice* 4, 3 (2010), 297–308. https://doi.org/10.1080/17512781003640703

[43] Matthew Hindman. 2011. Less of the same: The lack of local news on the Internet. *Prepared for the FCC* (2011).

[44] Liangjie Hong and Brian D Davison. 2010. Empirical study of topic modeling in Twitter. In *SOMA 2010 - Proceedings of the 1st Workshop on Social Media Analytics*. 80–88. https://doi.org/10.1145/1964858.1964870

[45] Desheng Hu, Ronald E Robertson, Shan Jiang, and Christo Wilson. 2019. Auditing the partisanship of Google search snippets. In *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019*. 693–704. https://doi.org/10.1145/3308558.3313654

[46] Qiang Jipeng, Qian Zhenyu, Li Yun, Yuan Yunhao, and Wu Xindong. 2019. Short Text Topic Modeling Techniques, Applications, and Performance: A Survey. (apr 2019). arXiv:1904.07695 http://arxiv.org/abs/1904.07695

[47] Mark Jurkowitz and Amy Mitchell. 2020. A fifth of Democrats, Republicans get news only from outlets with like-minded audiences. https://www.journalism.org/2020/03/04/about-one-fifth-of-democrats-and-republicans-get-political-news-in-a-kind-of-media-bubble/

[48] Mark Jurkowitz, Amy Mitchell, Elisa Shearer, and Mason Walker. 2020. *U.S. Media Polarization and the 2020 Election: A Nation Devided.* Technical Report. 67 pages. https://www.journalism.org/2020/01/24/americans-are-divided-by-party-in-the-sources-they-turn-to-for-political-news/

[49] Marius Kaminskas and Derek Bridge. 2016. Diversity, Serendipity, Novelty, and Coverage. *ACM Transactions on Interactive Intelligent Systems* 7, 1 (2016), 1–42. https://doi.org/10.1145/2926720

[50] Jacob Kastrenakes. 2020. Twitter says AI tweet recommendations helped it add millions of users. https://www.theverge.com/2020/2/6/21125431/twitter-q4-2019-earnings-daily-user-growth-machine-learning

[51] Nicolas Koumchatzky and Anton Andryeyev. 2017. Using Deep Learning at Scale in Twitter's Timelines. https://blog.twitter.com/engineering/en{_}us/topics/insights/2017/using-deep-learning-at-scale-in-twitters-timelines.html

[52] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P. Gummadi, and Karrie Karahalios. 2019. Search bias quantification: investigating political bias in social media and web search. *Information Retrieval Journal* 22, 1-2 (apr 2019), 188–227. https://doi.org/10.1007/s10791-018-9341-2

[53] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a Social Network or a News Media?. In *Proceedings of the 19th international conference on World wide web*. ACM, 591–600. http://bit.ly

[54] David Lazer. 2020. Studying human attention on the Internet. , 21–22 pages. https://doi.org/10.1073/pnas.1919348117

[55] Huyen Le, Raven Maragh, Brian Ekdale, Andrew High, Timothy Havens, and Zubair Shafiq. 2019. Measuring Political Personalization of Google News Search. Association for Computing Machinery (ACM), 2957–2963. https://doi.org/10.1145/3308558.3313682

[56] Jayeon Lee and Weiai Xu. 2018. The more attacks, the more retweets: Trump's and Clinton's agenda setting on Twitter. *Public Relations Review* 44, 2 (jun 2018), 201–213. https://doi.org/10.1016/j.pubrev.2017.10.002

[57] Nicholas Léonard and Cibele Montez Halasz. 2018. Twitter meets TensorFlow. https://blog.twitter.com/engineering/en{_}us/topics/insights/2018/twittertensorflow.html

[58] Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. 2016. Topic modeling for short texts with auxiliary word embeddings. In *SIGIR 2016 - Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, Inc, New York, New York, USA, 165–174. https://doi.org/10.1145/2911451.2911499

[59] Feng Lu, Anca Dumitrache, and David Graus. 2020. Beyond Optimizing for Clicks : Incorporating Editorial Values in News Recommendation. In *UMAP*. https://doi.org/10.1145/3340631.3394864 arXiv:2004.09980

[60] Ingrid Lunden. 2020. Twitter Q1: sales up 3% to $808M as it swings to a loss on COVID-19, mDAUS hit record 166M. https://techcrunch.com/2020/04/30/twitter-q1-sales-up-3-to-808m-as-it-swigs-to-a-loss-on-covid-19-mdaus-hit-record-166m/

[61] Momin M. Malik and Jürgen Pfeffer. 2016. A Macroscopic Analysis of News Content in Twitter. *Digital Journalism* 4, 8 (nov 2016), 955–979. https://doi.org/10.1080/21670811.2015.1133249

[62] Katerina Eva Matsa and Elisa Shearer. 2018. *News Use Across Social Media Platforms | Pew Research Center.* Technical Report. http://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/

[63] Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. 2013. Improving LDA topic models for microblogs via tweet pooling and automatic labeling. In *SIGIR 2013 - Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, New York, USA, 889–892. https://doi.org/10.1145/2484028.2484166

[64] Danaë Metaxa, Joon Sung Park, James A Landay, and Jeff Hancock. 2019. Search media and elections: A longitudinal investigation of political search results in the 2018 U.S. Elections. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019). https://doi.org/10.1145/3359231

[65] Amy Mitchell and J Baxter Oliphant. 2020. *Americans Immersed in COVID-19 News; Most Think Media Are Doing Fairly Well Covering It*. Technical Report. 1–4 pages. https://www.journalism.org/2020/03/18/americans-immersed-in-covid-19-news-most-think-media-are-doing-fairly-well-covering-it/

[66] Tanushree Mitra and Eric Gilbert. 2015. CREDBANK: A large-scale social media corpus with associated credibility annotations. In *Proceedings of the 9th International Conference on Web and Social Media, ICWSM 2015*. 258–267. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/viewPaper/10582

[67] Judith Möller, Damian Trilling, Natali Helberger, and Bram van Es. 2018. Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity. *Information Communication and Society* 21, 7 (2018), 959–977. https://doi.org/10.1080/1369118X.2018.1444076

[68] Philip M. Napoli. 2011. Exposure Diversity Reconsidered. *Journal of Information Policy* 1, 4 (2011), 246–259. https://doi.org/10.5325/jinfopoli.1.2011.0246

[69] Philip M. Napoli. 2015. Social media and the public interest: Governance of news platforms in the realm of individual and algorithmic gatekeepers. *Telecommunications Policy* 39, 9 (2015), 751–760. https://doi.org/10.1016/j.telpol.2014.12.003

[70] Philip M. Napoli. 2015. Social Media and the Public Interest: The Rise of Algorithmic News and the Future of the Marketplace of Ideas. *Telecommunications Policy* (2015).

[71] Efrat Nechushtai and Seth C. Lewis. 2019. What kind of news gatekeepers do we want machines to be? Filter bubbles, fragmentation, and the normative dimensions of algorithmic recommendations. *Computers in Human Behavior* 90 (jan 2019), 298–307. https://doi.org/10.1016/j.chb.2018.07.043

[72] Claudia Orellana-Rodriguez and Mark T. Keane. 2018. Attention to news and its dissemination on Twitter: A survey. https://doi.org/10.1016/j.cosrev.2018.07.001

[73] Eli Pariser. 2011. *The filter bubble: how the new personalized Web is changing what we read and how we think*. Penguin. https://doi.org/10.5860/CHOICE.50-0926

[74] Marco Pennacchiotti and Ana-Maria Popescu. 2011. A Machine Learning Approach to Twitter User Classification. In *Fifth international AAAI conference on weblogs and social media (ICWSM)*.

[75] Chelsea Peterson-Salahuddin and Nicholas Diakopoulos. 2020. Negotiated autonomy: The role of social media algorithms in editorial decision making. *Media and Communication* 8, 3 (jul 2020), 27–38. https://doi.org/10.17645/mac.v8i3.3001

[76] Paul Resnick, R. Kelly Garrett, Travis Kriplean, Sean A Munson, and Natalie Jomini Stroud. 2013. Bursting Your (Filter) Bubble: Strategies for Promoting Diverse Exposure. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW* (2013), 95–100.

[77] Ronald E Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. Auditing partisan audience bias within Google search. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 22. https://doi.org/10.1145/3274417

[78] Peter Salovey and Pamela Williams-Piehota. 2004. Field Experiments in Social Psychology: Message Framing and the Promotion of Health Protective Behaviors. In *American Behavioral Scientist*, Vol. 47. SAGE Publications, 488–505. https://doi.org/10.1177/0002764203259293

[79] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. In *Data and discrimination: converting critical concerns into productive inquiry*. 1—-23.

[80] Michael. Schudson. 1995. *The Power of News*. Harvard University Press. 269 pages.

[81] Limor Shifman. 2013. Memes in a digital world: Reconciling with a conceptual troublemaker. *Journal of Computer-Mediated Communication* 18, 3 (2013), 362–377. https://doi.org/10.1111/jcc4.12013

[82] Sprout Social. 2020. How the Twitter Algorithm Works in 2020. https://sproutsocial.com/insights/twitter-algorithm/

[83] Kate Starbird. 2017. Examining the Alternative Media Ecosystem through the Production of Alternative Narratives of Mass Shooting Events on Twitter. *ICWSM* (2017).

[84] Leo G. Stewart, Ahmer Arif, A. Conrad Nied, Emma S. Spiro, and Kate Starbird. 2017. Drawing the lines of contention: Networked frame contests within #BlackLivesMatter discourse. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (nov 2017), 1–23. https://doi.org/10.1145/3134920

[85] Natalie Jomini Stroud. 2010. Polarization and partisan selective exposure. *Journal of Communication* 60, 3 (aug 2010), 556–576. https://doi.org/10.1111/j.1460-2466.2010.01497.x
[86] Didi Surian, Dat Quoc Nguyen, Georgina Kennedy, Mark Johnson, Enrico Coiera, and Adam G. Dunn. 2016. Characterizing twitter discussions about HPV vaccines using topic modeling and community detection. *Journal of Medical Internet Research* 18, 8 (aug 2016), e232. https://doi.org/10.2196/jmir.6045
[87] Daniel Trielli and Nicholas Diakopoulos. 2019. Search as News Curator: The Role of Google in Shaping Attention to News Information. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM.
[88] Daniel Trielli and Nicholas Diakopoulos. 2020. Partisan search behavior and Google results in the 2018 U.S. midterm elections. *Information, Communication & Society* (may 2020), 1–17. https://doi.org/10.1080/1369118X.2020.1764605
[89] Twitter. 2020. About your Twitter timeline. https://help.twitter.com/en/using-twitter/twitter-timeline
[90] Twitter. 2020. Twitter moments: Guidelines and principles. https://help.twitter.com/en/rules-and-policies/twitter-moments-guidelines-and-principleshttps://about.twitter.com/company/moments-guidelines
[91] Sebastián Valenzuela, Soledad Puente, and Pablo M. Flores. 2017. Comparing Disaster News on Twitter and Television: an Intermedia Agenda Setting Perspective. *Journal of Broadcasting and Electronic Media* 61, 4 (2017), 615–637. https://doi.org/10.1080/08838151.2017.1344673
[92] Nicholas Vincent, Isaac Johnson, Patrick Sheehan, and Brent Hecht. 2019. Measuring the Importance of User-Generated Content to Search Engines. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*.
[93] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (mar 2018), 1146–1151. https://doi.org/10.1126/science.aap9559
[94] David H. Weaver and Lars Willnat. 2016. Changes in U.S. Journalism: How do journalists think about social media? *Journalism Practice* 10, 7 (oct 2016), 844–855. https://doi.org/10.1080/17512786.2016.1171162
[95] James G Webster and Harsh Taneja. 2018. Building and interpreting audience networks: A response to mukerjee, majo-vazquez & gonzalez-bailon. , E11–E14 pages. https://doi.org/10.1093/joc/jqy024
[96] John Wihbey, Kenneth Joseph, and David Lazer. 2019. The social silos of journalism? Twitter, news media and partisan segregation. *New Media and Society* 21, 4 (2019), 815–835. https://doi.org/10.1177/1461444818807133
[97] Stefan Wojcik and Adam Hughes. 2019. Sizing Up Twitter Users. *Pew Research Center* (2019), 1–23.
[98] Shaomei Wu, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. 2011. Who says what to whom on twitter. In *Proceedings of the 20th International Conference on World Wide Web, WWW 2011*. 705–714. https://doi.org/10.1145/1963405.1963504
[99] Yan Yan and Wanjiang Zhang. 2020. Gossip at one's fingertips: Predictors of celebrity news on Twitter. *Journalism* 21, 5 (may 2020). https://doi.org/10.1177/1464884918791349
[100] Kai-Cheng Yang, Onur Varol, Pik-Mai Hui, and Filippo Menczer. 2020. Scalable and Generalizable Social Bot Detection through Data Selection. In *Association for the Advancement of Artificial Intelligence (AAAI)*.
[101] Jianhua Yin and Jianyong Wang. 2014. A Dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, New York, New York, USA, 233–242. https://doi.org/10.1145/2623330.2623715
[102] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. Comparing twitter and traditional media using topic models. In *European conference on information retrieval*. Springer. https://doi.org/10.1007/978-3-642-20161-5_34

Received June 2020; revised October 2020; accepted December 2020.