Auditing the audits: evaluating methodologies for social media recommender system audits

Paul Bouchaud, Pedro Ramaciotti Morales

Bouchaud and Ramaciotti
Applied Network Science (2024) 9:59
https://doi.org/10.1007/s41109-024-00668-6

Applied Network Science

RESEARCH Open Access

Auditing the audits: evaluating methodologies for social media recommender system audits

Paul Bouchaud[1,2*] and Pedro Ramaciotti[1,3,4]

*Correspondence:
paul.bouchaud@iscpif.fr

[1] Complex Systems Institute of Paris Ile-de-France CNRS, Paris, France
[2] CAMS EHESS, Paris, France
[3] médialab Science Po, Paris, France
[4] LPI Université Paris Cité, Paris, France

Abstract

Through a simulated Twitter-like platform designed to optimize user engagement and grounded in authentic behavioral data, this study evaluates methodologies for auditing social media recommender systems. Our analysis focuses on the impact of key parameters in sock-puppet audits, the number of friends and session length, on audit outcomes. Additionally, we investigate the algorithmic amplification of political content across different levels of granularity, segmenting users based on political leanings and considering multiple political dimensions beyond declared affiliations. Our findings underscore the necessity of employing realistic parameter settings in audits and highlight the importance of nuanced political segmentation. Amid increasing regulatory scrutiny, this research contributes to enhancing methodologies for auditing social media platforms.

Keywords: Social media, Ideology, Polarization, Diversity, Algorithm audit, User models

Introduction

In recent years, social media platforms have emerged as pivotal arenas for political discourse, offering a platform for politicians, political organizations, and news outlets to interact with vast audiences (Benkler et al. 2018; Stewart and Hartmann 2020). These platforms engage in advertising activities, where their revenues are influenced by the number of ads shown to users and their subsequent interactions (Meta 2023). Consequently, the platforms' earnings are associated with the user activity on the platform; to optimize user engagement, they employ sophisticated algorithms to curate and rank content (Covington et al. 2016; Satuluri et al. 2020). The ramifications of such engagement-maximizing recommender systems have elicited intense scholarly and public scrutiny. Previous studies have highlighted the potential impact of algorithmic ranking on political discourse, amplifying emotionally-charged content (Brady et al. 2017; Bouchaud et al. 2023) and contributing to polarization (Rathje et al. 2021; Van Bavel et al. 2021; Milli et al. 2023). Part of this attention and scrutiny have translated into policy, for instance, in the European Union Digital Services Act.

While social and political implications of the massification of social media use, particularly taking place on Twitter/X (hereinafter Twitter) has been extensively studied by academics, thanks to their late openness to research, only a handful of research works focus on auditing the algorithms mediating such use. Auditing the algorithmic machinery used by social media platforms is particularly challenging as it requires access to data that is usually private, in particular, what users are being served by those platforms. Hence, to assess the impact of recommender systems, previous research either relied on volunteers providing their data (Hargreaves et al. 2018; Bouchaud et al. 2023; Milli et al. 2023), on non-public proprietary data (Huszár et al. 2021; Guess et al. 2023), or more easily using, so-called "sock-puppet" accounts, where researchers programmatically interact with the platform with fake accounts (Sandvig et al. 2014; Hussein et al. 2020; Haroon et al. 2023; Bandy and Diakopoulos 2021; Bartley et al. 2021).

In this paper, we seek to strengthen these various audit methodologies through empirically-grounded social media simulations. Specifically, using engagement predictive models trained on behavioral data obtained via a data donation program, we simulate a Twitter-like platform to examine how audit conclusions vary with changes in the number of friends, user session length, segmentation based on political leaning, and consideration of multidimensional ideological positions. We show that while computing the amplification of Members of Parliament across the entire population, audit may conclude that the "mainstream political right enjoys higher algorithmic amplification than the mainstream political left" (Huszár et al. 2021), when segmenting users based on their political leanings, we observe that engagement-based ranking merely favors content aligned with users' ideological preferences.

The article is structured as depicted on Fig. 1: We begin by introducing previous algorithmic auditing methodologies and their findings. Next, we describe our simulation of a Twitter-like platform, utilizing the engagement prediction models we developed. We then present the metrics used to analyze the generated timelines. Finally, we report and discuss how the conclusions drawn from audits depend on their methodologies

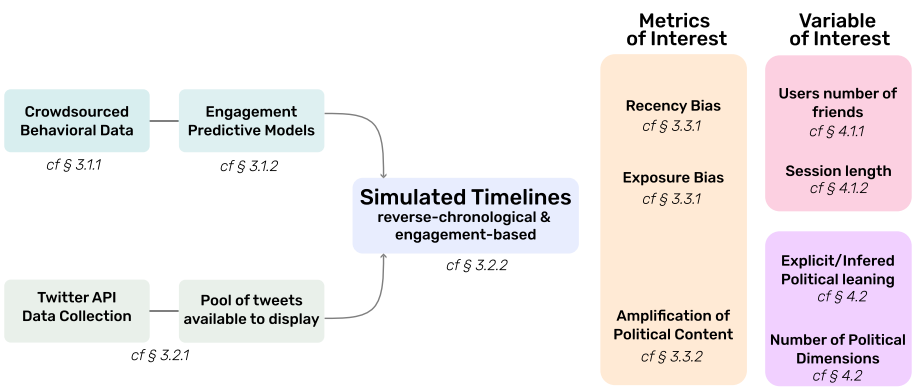

Fig. 1 Structure of this study

Related work

Algorithmic auditing studies relying on data donation have the potential to offer valuable insights into real-life effects of social media algorithms. Recent studies have shown that Twitter's recommender systems amplifies toxic and divisive content, displaying a political landscape different from the one users subscribed to (Bouchaud et al. 2023; Milli et al. 2023), contrasting with Facebook were studies fail to show such deviations (Bakshy et al. 2015; González-Bailón et al. 2023). Nevertheless, these studies can be costly, involve intrusive data-collection, and may be prone to potential selection bias as they rely on user willingness to donate their data (Kmetty et al. 2023).

On the other hand, sock-puppet audits offer a perfectly-controlled setting for researchers to understand the behavior of the platform. Researchers create fake accounts and interact with the platform, in an automated way, while trying to mimic real human behavior. Using such an approach, Boeker and Urman (2022) investigated the variables that are most determinant in personalized user profiling on TikTok, Haroon et al. (2023) showed, leveraging one hundred thousand sock-puppets, that YouTube's algorithm recommends ideologically congenial content to partisan users. However, despite the experimental control they offer, the insights into real-life effects of recommender systems provided by sock-puppets are hindered by their imperfect mimicking of human digital behavior, for instance by blindly follow YouTube videos recommendations (Haroon et al. 2023), contrary to real users (Lee et al. 2022). While sock-puppets studies showed that watching videos related to misinformation, conspiracy theories or pseudoscience causes YouTube to recommend more such content (Hussein et al. 2020; Haroon et al. 2023), recent works using real navigation logs highlights that YouTube's recommender system is not the primary driver of attention toward ideologically extreme content (Hosseinmardi et al. 2024; Chen et al. 2023). Ribeiro et al. (2023) illustrate, with a simple agent-based model, how audits failing to model how users interact with algorithms are of limited utility in determining the prevalence of phenomena like radicalization and algorithmically-driven segregation (e.g., as in the so-called "filter bubbles").

In addition to those theoretical limitations, technical implementations of sock-puppet audits are impaired by platform opacity, resulting in somewhat arbitrary methodological decisions. Should an auditor use logged-in YouTube accounts while gathering recommendations from the algorithm, or non-logged-in users accepting cookies are enough? Should sock-puppets watch YouTube videos in their entirety, or only the first few seconds? As explored by Chandio et al. (2023) on YouTube, these design choices may significantly impact the conclusions drawn from recommender audits.

Because it requires users to be logged-in (which is not enforced on all platforms, e.g., YouTube), there have been fewer sock-puppet audits of Twitter. Bartley et al. (2021) created four pairs of bot accounts, within each pair one bot is set up to see tweets in a personalized timeline and the other in reverse chronological order. These bots logged in six times a day during the Spring of 2020 and observed "at least 15 tweets" per session, without performing any actions. The sock-puppet accounts followed between 55 and 59 friends, randomly sampled from a curated list of popular anti-science and pro-science Twitter accounts related to COVID-19. With this methodology, Bartley et al. (2021) found that Twitter's algorithmic curation favored popular tweets (receiving more likes and retweets at the time of exposure), amplified few of the accounts a user follows

and recommended older tweets than in chronological timelines. The second sock-puppet audit of Twitter by Bandy and Diakopoulos (2021) also created eight sock puppet accounts, collecting 50 tweets at 9am and 9pm each day during 30 days during the Spring of 2020. The accounts sought to emulate "archetypal" real-world users, selected through a large-scale network analysis among users following US congresspeople. As a result of Twitter's algorithmic curation, Bandy and Diakopoulos (2021) found a decrease in the number of external links (i.e., linked referring to websites outside Twitter), an increase in source exposure diversity in terms of the number of Twitter accounts, and an increased exposure to partisan-specific content within US bipartisan landscape. While providing interesting insights, the mimicked sock-puppets differ from real Twitter users in several respects, for instance the diversity and number of friends or the length of scrolling sessions. In this article we explore how these parameters impact the conclusions of such audits.

Alternatively, studies conducted by, or in collaboration with, corporations provide valuable insights, leveraging access to non-public data and the capacity to conduct large-scale experiments. Through collaboration with Meta, Guess et al. (2023) found that switching users from algorithmic feeds to reverse-chronological feeds significantly reduced their platform usage and increased exposure to political content, content from moderate friends, and ideologically mixed sources on Facebook. Leveraging proprietary user information and a multi-year controlled experiment involving nearly two million users, Huszár et al. (2021) revealed how Twitter's recommender system unevenly amplified tweets from politicians based on their ideological leaning. Despite delivering unprecedented insight by its scale and access, this study computed the algorithmic amplification over the whole population, country-wise, without, for instance, segmenting users based on their political leaning. This lack of granularity in audit can be particularly impactful as the algorithmic curation precisely seeks to shape the timelines to users' unique tastes. Also, it relies on politicians' declared political groups, which may lead to imprecise assessment, particularly when a diversity of ideological leanings may co-exists within a political party, as underscored by multiple studies on ideological scaling (Ramaciotti et al. 2021; Ramaciotti Morales et al. 2022).

Methods

To answer these research questions, we developed a framework simulating the timeline curation of a social media platform such as Twitter. In particular, we trained machine learning models on behavioral data from Twitter to predict the engagement of users to a given tweet.

Engagement predictive models

Training dataset

The prediction of engagement, cornerstone of social media platforms, relies on the extensive behavioral data. Such training data are proprietary assets held by companies and are not publicly accessible. Twitter did collaborate with ACM RecSys to provide large datasets for engagement prediction challenges in 2020 and 2021, composed of publicly available information (Belli et al. 2020, 2021). However, the sole focus of these challenges has been on enhancing prediction accuracy rather than delving into the broader

ethical implications associated with the large-scale deployment of such recommender systems. For our research, the datasets released for ACM RecSys Challenges were unavailable. We constructed a training dataset through a data donation program, Horus, collecting, through a browser add-on, Twitter timelines of volunteers and their potential engagements on each displayed tweet. With 2258 installations and 739 users active on Twitter since its deployment in October 2022, our extension captured, up to January 2024, more than 16.7 millions tweets impressions. Among those impressions, we captured 91 129 likes and 20 724 retweets from our volunteers.

Furthermore, we enrich our dataset by incorporating historical and account-specific features. Among those, we collected the tweets users previously liked and retweeted as well as their number of followers, total number of likes and posts since the creation of their account. Similarly as in Twitter's recommender system pipeline (Rossi et al. 2022), we assess the similarity between Twitter accounts through a follower network. The network was constructed using a snowball sampling, using as seed accounts the data-donation volunteers, the users selected for this study (detailed below), popular accounts and randomly picked accounts from the Politoscope and Climatoscope databases (Gaumont et al. 2018). We then pruned nodes with a degree less than 150, ending up with a network large of 1.58 billions edges and 4.1 millions nodes. The network contains 88.1% of volunteers' friends and 90.5% of simulated users' friends, as well as all the authors of tweets of the training and inference datasets (introduced below). We embedded the resulting network through the *node2vec* algorithm (Grover and Leskovec 2016), which assigns low-dimensional embeddings to nodes while maximizing the likelihood of preserving neighborhood relationships through biased random walks. Our exploration of the graph further accentuated homophily, understood here as having common friends, by favoring a breadth-first sampling approach (with parameters p=1 and q=0.25). For the purposes of computational efficiency, we performed a dimensional reduction on the initial 64-dimensional *node2vec* embeddings, reducing them to just 16 dimensions using the PaCMAP algorithm (Wang et al. 2021). This approach retains local structure needed to assess homophily (Chari and Pachter 2023), with users following and followed by similar sets of users positioned in close proximity.

Models training

Following the approach of ACM RecSys Challenges winning implementations (https://blog.twitter.com/engineering/en_us/topics/insights/2020/what_twitter_learned_from_recsys2020), we trained gradient-boosting machines, specifically LightGBM (Ke et al. 2017), on a small set of features, listed in annex Table 1. Among these features, some surfaced as particularly influential during the training process: the age of the tweet, the average word length within the tweet, metrics related to the similarity between the tweet's author and the last authors liked and retweeted by the user, and the Jaccard index between the user's friends and the tweet's author's friends.We emphasize that our objective is not to claim beating the current, proprietary, state-of-the-art with our new machine learning model. Instead, our focus is on training models that effectively predict engagement for the specific purposes of conducting simulations with synthetic users. Overall, our model achieved an average precision score (summarizing the precision-recall curve) of 86.3% for predicting likes and 77.6% for predicting retweets, a gain in

cross-entropy of 55.9% for likes and 45.1% for retweets compared to a naive baseline. To assess the generalization capability of our models, we followed the methodology from Barbiero et al. (2020). Specifically, we compared the performance on data points either inside or outside the convex hull, within the feature space i.e. the vector space associated to features vectors, of the training dataset and verified that samples from the simulated timelines exhibited a high similarity with the training data, we refer to Bouchaud (2024) for further details.

Simulated timelines

Relying exclusively on publicly available data for crafting model features enables us to compute the probability that a given user, whose profile is public on Twitter, will like or retweet a particular tweet.

Data collection

For the purpose of this analysis, we simulate timelines for a set of 6363 Twitter accounts, randomly selected in the Politoscope database. This database gathers tweets authored by French Political figures and/or containing political keywords, see Gaumont et al. (2018) for the curation details. We considered the 2022 retweets network, filtering out users having retweeted or been retweeted less than five times. The network exhibits political communities (Gaumont et al. 2018), we then sample users from each, such as reflecting their respective size.

To construct the corpus of tweets that might appear in the simulated timelines, we gathered all the tweets that had been either published or retweeted by a subset of users' friends. Our pool of 6363 simulated users followed a total of 1,744,564 unique Twitter accounts, 41,186 of them are followed by at least 40 simulated users, we fetched those. Moreover, we randomly selected accounts in the pool of friends, ending up with a corpus made of 95,778 accounts. This selection of accounts covered, on average, 66.4% of the friends of the users (median 70.3%). This data collection effort spanned from March 1st, 2023, to March 15, 2023.

Timelines simulation

We simulated the timelines as follows: each user logs in every day for 15 consecutive days at randomly selected hours, following the empirical distribution of the users that installed our browser add-on. We then aggregate all tweets that have been either published or retweeted by the user's friends, as per our database, within an 18-hour time window preceding the user's session. We estimate the probability of each tweet being liked or retweeted by the user with our engagement predictive models. We then sum these two probabilities to derive an engagement score, in accordance with Twitter's design principles (https://github.com/twitter/the-algorithm-ml/blob/main/projects/home/recap/README.md). The synthetic users in our simulation are presented with two different recommender systems: (1) presenting tweets ranked in decreasing engagement score, and (2) in reverse chronological order.

To bolster our findings, we implemented two simple heuristics to mitigate potential trivial distorting effects. The first heuristic ensured that each Twitter account is represented only once within a given timelines. The second remove duplicated tweets, such as

a tweet being retweeted by multiple friends. These heuristics, implemented on Twitter in a less stringent way (https://blog.twitter.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm), prevent spamming a user from a single account or tweet.

Finally, we simulate scrolling in our synthetic population by setting the number L of tweets that a given user scrolls while reading. The session length $L \in [5, 100]$ is drawn from the empirical distributions, displayed on Fig. 4A, with a median value of 30 tweets read per session. The distribution was determined from the 256k user sessions collected through our data-donation initiative. The size of the aggregation window, set at 18 h, is a balance between ensuring a sufficiently sized tweet pool while also managing computational costs, a sensitivity analysis can be found in Bouchaud (2024). In the session collected via our data-donation initiative, the oldest tweets displayed within a session are on average 16.4 h old (filtering out tweets older than 3 days, typically displayed when associated with a recent reply).

It's important to note that this article does not delve into the intricacies of feedback loops, wherein the recommender adjusts itself based on user past engagement, as successive timelines are treated as independent. We also do not delve into higher-order effects, which could emerge from messages retweeted by users' friends. Indeed, for retweeted tweets, the intervention of recommender systems is two fold: initially, the system displays a tweet in a users' friends feed, who decided to retweet it, followed by the decision of whether to display the retweet in users' timelines.

Metrics

In this section we present the metrics through which we will analyze the generated timelines. Except mentioned otherwise, all metrics will be reported with a 95% confidence interval, determined by bootstrapping, sampling 250 times with replacement, with the same sample size as the original data. Also, the correlation will be reported with a 95% confidence interval, estimated using the Fisher transformation, with a p-value below.01.

Recency and exposure bias

Following the analysis of Bartley et al. (2021), we assess the so-called *recency bias* and *exposure bias* by comparing the age of the tweets as well as the friends being displayed in engagement-based and reverse-chronological timelines.

The *recency bias* is measured by comparing the median age of tweets, (with age measured as the time elapsed between publication and impression), in engagement-based $\widetilde{age}_{eng}$ and reverse-chronological $\widetilde{age}_{chrono}$ timelines.

To quantify *exposure bias* we compare the Gini coefficient of friends activity, i.e. the number of tweets they published, and of friends impressions, i.e. the number of tweets displayed on the timelines. A low Gini coefficient means that all friends have a similar posting/impression frequency, while a high Gini coefficient indicates that a few friends are responsible for a significant portion of the tweets/impressions. For our experimental setting, we prefer Gini coefficient to other diversity metrics such as entropy because it does not take into account the number of friends of each user, and only the degree to which they are presented uniformly in feeds (Ramaciotti et al. 2021). We compute the Gini coefficients associated to the productions of tweets by

friends, $G(P)$, to the impressions of friends in timelines, either engagement-based, $G(I_{eng})$, or reverse-chronological, $G(I_{chrono})$. In addition to the Gini coefficients, we compute the perplexity associated with the frequencies of publication, PP_{prod}, and of impressions of users' friends,PP_{chrono} & PP_{eng}, to assess the diversity in authors displayed in the timelines.

We will evaluate these metrics as a function of n_f, the number of friends of each simulated user, and as a function of L, the session lengths, two parameters usually arbitrarily fixed in sock-puppet audits.

Algorithmic amplification

Following the approach of Huszár et al. (2021), we seek to compare how the political landscape depicted in engagement-based timelines may differ from the one in reverse-chronological timelines. We then define the *reach amplification* $a_R(T, U)$ for a set T of tweets within an audience U as: the ratio between (1) the number of unique users in U who have seen at least one tweet from T in their engagement-based timelines and (2) the number of unique users in U who have seen at least one tweet from T in their reverse-chronological timelines:

$$a_R(T, U, d) = \frac{\sum_{u \in U} \mathbb{I}(feed^{eng}(u, d) \cap T \neq \emptyset)}{\sum_{u \in U} \mathbb{I}(feed^{chrono}(u, d) \cap T \neq \emptyset)} - 1$$

Here, $feed^{eng,chrono}(u, d)$ represents the set of tweets appearing in the timelines, either engagement-based or reverse-chronological, of user u at day d. This ratio is normalized so that an amplification ratio of 0% corresponds to an equal reach in engagement-based and reverse-chronological timelines.

However, by focusing on tweets' reach, the measure $a_R(T, U)$, used in Huszár et al. (2021), may overlook differences in terms of number of impressions between reverse-chronological and engagement-based timelines. To address this limitation, we introduce the impression amplification, $a_I(T, U)$, which is defined as the ratio between (1) the number of impressions of tweets from the set T in the engagement-based timelines of users in U and (2) the number of impressions of tweets from the set T in the reverse-chronological timelines of users in U:

$$a_I(T, U, d) = \frac{\sum_{u \in U} |feed^{eng}(u, d) \cap T|}{\sum_{u \in U} |feed^{chrono}(u, d) \cap T|} - 1$$

Higher amplification values indicate that engagement predictive models assign higher relevance scores to the set of tweets T, causing them to appear more frequently than they would in reverse-chronological timelines. The amplification is calculated relative to the reverse-chronological ranking baseline. This baseline was selected to ensure consistency with prior audits that we aim to replicate, in particular (Bandy and Diakopoulos 2021; Bartley et al. 2021; Huszár et al. 2021), and due to the lack of an alternative 'null' model for content selection. It is important to note, however, that reverse-chronological ranking, although historically used as the default on Twitter and other social media platforms and sometimes depicted as an alternative to 'algorithmic feeds', is not neutral. This ranking system is based on recency, which tends to favor frequent posters.

Using these definitions of amplification, we first consider the set of tweets published by elected officials members of the French Parliament, considering their declared political label as group, as done in Huszár et al. (2021). To explore more thoroughly the consequences of engagement maximization, we considered Members of Parliament (MPs) and user positions along multiple ideological dimensions, and form subgroups $\widetilde{U} \subset U$ of ideologically aligned users instead of computing the amplification over the whole population.

Political leaning

We sought to assign a granular political leaning to MPs and simulated users, beyond the mere declared political groups. To this end, we leveraged ideological embedding derived from the following relationships of MPs on Twitter (Barberá et al. 2015; Ramaciotti Morales et al. 2022). We analyzed a subset of 881 out of 925 French MPs active on Twitter, representing 10 political parties, along with their followers. Data collection took place via the Twitter API in March 2023, see privacy-compliance information in the dedicated section. To identify users engaged in political discourse while minimizing the inclusion of inactive accounts, bots, or those following MPs for non-ideological reasons, we followed the criteria outlined in Barberá (2015). Specifically, we considered only followers who follow at least 3 MPs and have a minimum of 25 followers themselves, resulting in 518.460 users.

Through Correspondance Analysis (Greenacre 2017), as an approximation of ideal point estimation (Lowe 2008), we produced a reduced dimensionality spatial representation of the MP-follower bipartite graph. This representation preserves homophily as MPs positioned closely are followed by similar sets of users, and users positioned closely follow similar sets of MPs. To compute positions along interpretable political continuous dimensions, we used party positions contained in the Chapel Hill Expert Survey, CHES (Jolly et al. 2022). In the CHES data, European parties are positioned along 53 issue and ideology dimensions using responses from 412 political science experts. To map our MPs and users to these dimensions, we use the positions of political parties. In our ideal point homophily space, we compute party positions as the centroid of MPs affiliated to that party. We then compute a map between homophily space and CHES dimensions as a Linear Least Squares problem using the position of parties. As depicted in Fig. 2, the first two principal components exhibit strong correlations with the Left-Right axis and the Anti-Elite Salience, respectively. These two political dimensions, have been identified in several studies as the two main dimensions structuring online politics in France (Ramaciotti et al. 2021; Cardon et al. 2019). In the CHES data used to compute political leanings, continuous positions range from 0 to 10, going left-most (0) and right-most (10) positions in the Left-Right dimension, and least anti-elite (0) and most anti-elite sentiment (10) in the Anti-Elite Salience dimension.

We use these two dimensions to compute the amplification of tweets of MPs. Furthermore, we will segment our simulated users based on their leaning on theses scales, considering far-left ($PC1 \in [-1, 2], PC2 \in [6, 10]$), left ($PC1 \in [2, 4.5], PC2 \in [4, 7]$), center ($PC1 \in [5, 6.5], PC2 \in [2, 4]$), right ($PC1 \in [7, 8], PC2 \in [4, 6]$) and far-right ($PC1 \in [9, 13], PC2 \in [8, 14]$) leaning users.

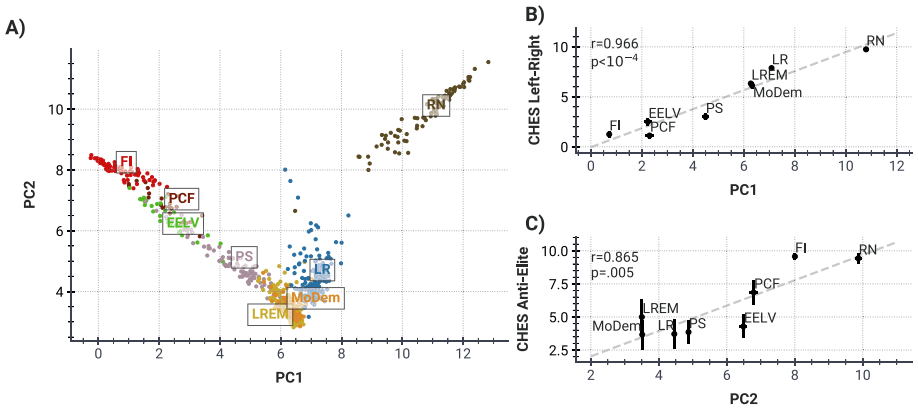

Fig. 2 **A** Positions of Members of Parliament in the reduced-dimensional space defined by the first two latent dimensions of the ideological space. The mean positions of MPs are correlated with the CHES positions on the Left-Right **B** and Anti-Elite Salience **C** axes. Error bars indicate bootstrap standard errors on the average position and expert assessments

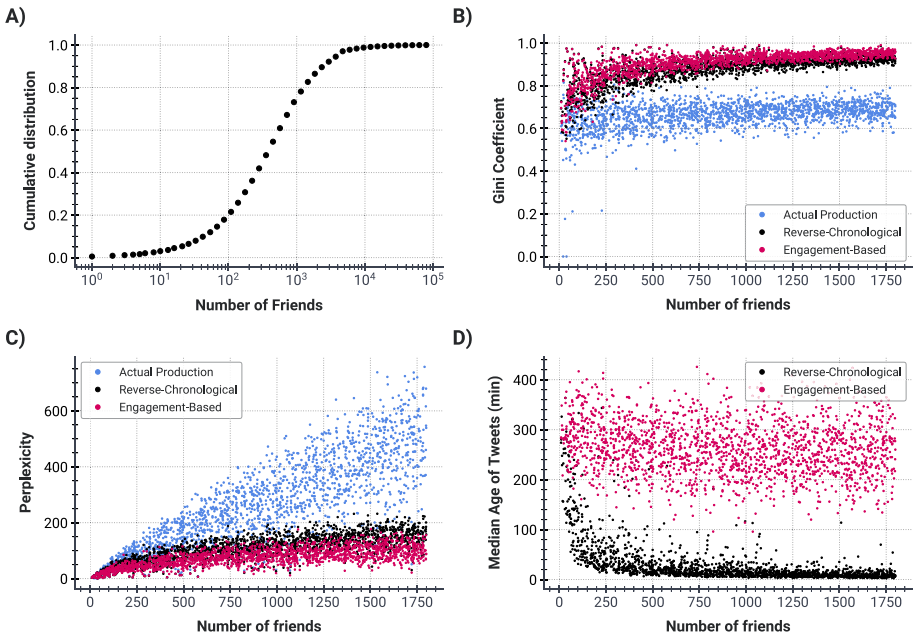

Fig. 3 **A** Empirical distribution of number of friends. Gini coefficients (**B**) and Perplexity (**C**) associated with friends' frequency of publication and impression in engagement-based and reverse-chronological timelines as a function of users' number of friends. **D** Median age of tweets in engagement-based and reverse-chronological timelines relative to users' number of friends

Results

Recency and exposure bias

Number of friends

We assess *recency bias* and *exposure bias* as a function of the number of friends of each simulated user. As depicted in Fig. 3B, both tweet production by friends, and their impressions in users' timelines exhibit increasing Gini coefficients with the number of friends. The Spearman rank correlation between users' number of friends and these Gini

coefficients are as follows: $\rho(G(P), n_f) = .39$ [.35, .43], $\rho(G(I_{chrono}), n_f) = .72$ [.70, .74], and $\rho(G(I_{eng}, n_f) = .77$ [.75, .79].

Likewise, the perplexity of the frequency of impressions of users' friends in timelines also increases as the number of friends grows. The Spearman's rank correlation is $\rho(PP_{chrono}, n_f) = .80$ [.78, .851] for reverse-chronological timelines and $\rho(PP_{eng}, n_f) = .70$ [.68, .72] for engagement-based ones. With more accounts being followed, the predictability of which account will appear in the timelines decreases. However, the ratio of perplexity decreases with the number of friends, $\rho(PP_{eng}/PP_{chrono}, n_f) = -.30$ [−.34, −.26]. An audit leveraging sock-puppet accounts following only a few accounts may therefore overestimate the disparity between engagement-based and reverse-chronological timelines.

Similarly, as displayed on Fig. 3C, as user's number of friends increases, as the tweets in the timelines tend to be recent. The Spearman rank correlations between users' number of friends and the median age of tweets in timelines are $\rho(\widetilde{age}_{chrono}, n_f) = -.72$ [−.74, −.69] for reverse-chronological timelines and $\rho(\widetilde{age}_{eng}, n_f) = -.17$[−.21, −.12] for engagement-based timelines. An audit based on accounts following between 10 and 100 accounts (uniformly distributed within this range) will observe a median tweet age in reverse-chronological timelines of 2.76 [2.53, 3.03] *h* and 4.78 [4.65, 4.92]*h* in engagement-based timelines. Conversely, an audit considering the distribution of the number of friends observed in the overall population will observe a median tweet age of 46.9 [43.9, 49.1] *min* in reverse-chronological timelines and 4.52 [4.48, 4.56] *h* in engagement-based timelines.

Session length

We examine *recency bias* and *exposure bias* with respect to session length, defined as the number of tweets displayed on users' timelines, while enforcing the empirical distribution of the number of friends. As depicted in Fig. 4, shorter simulated sessions exhibit more pronounced deviations. Specifically, for both engagement-based and reverse-chronological timelines, longer simulated sessions lead to the decreases of ratio between $G(I)$ and $G(P)$, the Gini coefficients associated with users' friends' tweets impressions/productions. Similarly, longer simulated sessions result in a decrease in the ratio between the perplexities of friends' impressions in engagement-based timelines, PP_{eng}, and reverse-chronological timelines, PP_{chrono}.

An audit that collects, session-wise, the top 50 tweets in engagement-based and reverse-chronological timelines would find that $G(I)$ is 35.6 [34.4, 37.2]% higher than $G(P)$ in engagement-based timelines compared to 30.7 [29.4, 32.0] in reverse-chronological timelines. However, when considering the empirical distribution of session length, the deviations from reverse-chronological are more pronounced, with $G(I)$ being 38.9[37.6, 40.5]% and 33.5[32.4, 35.0]% larger than $G(P)$ respectively. Similarly, the perplexity of friends' impressions in engagement-based timelines is 29.8 [28.3, 31.1]% smaller than in reverse-chronological timelines when considering the distribution of session length, while an audit considering the top 10 tweets would estimate it at 40.0 [38.2, 41.6]%.

Finally, longer simulated sessions correspond to an increase in the median age of tweets impressed in reverse-chronological timelines, while remaining relatively constant

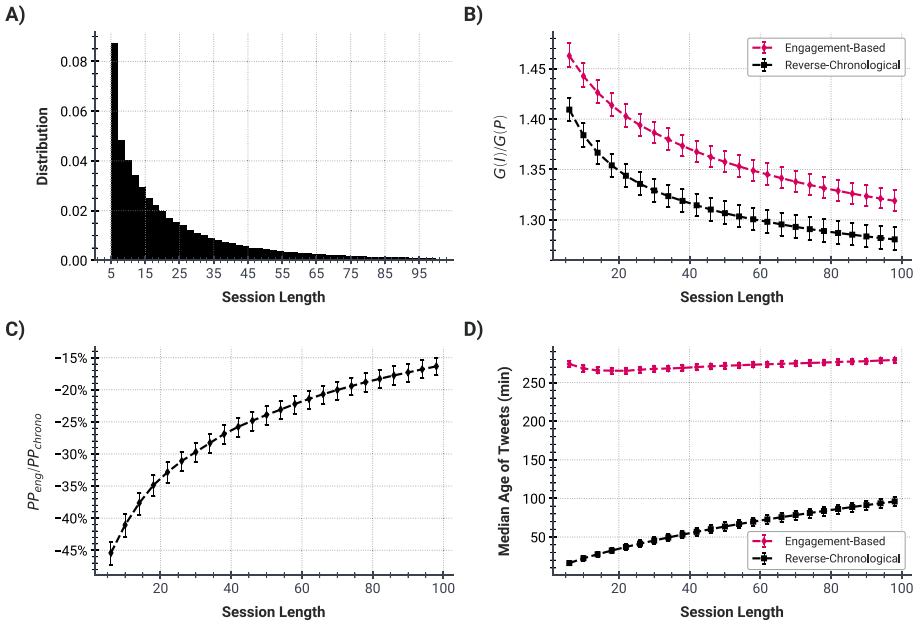

Fig. 4 **A** Distribution of session length in empirical data, with truncated tails where sessions longer than 100 tweets constitute 2.6% of observed sessions. **B** Ratio of Gini coefficients $G(I)/G(P)$ for engagement-based and reverse-chronological timelines relative to session length. **C** Ratio between perplexity in engagement-based (PP_{eng}) and reverse-chronological (PP_{chrono}) timelines relative to session length. **D** Average median age of tweets in engagement-based ($\widetilde{age}_{eng}$) and reverse-chronological ($\widetilde{age}_{chrono}$) timelines relative to session length. Error bars represent the 95% confidence interval determined by bootstrapping over users

in engagement-based timelines. For sessions made of 10 tweets, the median age of tweets is 24.8 [21.6, 28.3] minutes in reverse-chronological timelines and 268.9[265.2, 272.7] minutes in engagement-based timelines. For sessions consisting of 50 tweets, the median age of tweets is 66.5[61.1, 72.1] minutes in reverse-chronological timelines and 270.9[267.2, 274.5] minutes in engagement-based timelines.

Political leaning

Figure 5A presents the amplification of Members of Parliament's tweets, in terms of impressions in users' timelines, both across the entire population and segmented by users' political leanings. An audit at the population level, akin to that conducted by Huszár et al. (2021), reveals that tweets aligned with right-wing politics experience a higher algorithmic amplification compared to their left-wing counterparts. Specifically, tweets authored by far-right MPs are shown 69.6 [62.9, 75.3]% more in engagement-based timelines compared to reverse-chronological ones. However, upon segmenting users based on their political leanings, we observe that tweets authored by MPs ideologically aligned with the users are more prevalent in engagement-based timelines compared to reverse-chronological ones. For instance, far-left-leaning users are exposed to 42.2 [27.6, 54.0]% more tweets published by communist MPs ("PCF" group) in engagement-based timelines compared to reverse-chronological ones, while experiencing a decrease of 30.8 [28.1, 34.8]% from center-leaning "LREM" MPs tweets.

Focusing on the number of users served with at least one tweet from a given political group, the *reach* amplification, depicted in Fig. 5B, tends to underestimate the

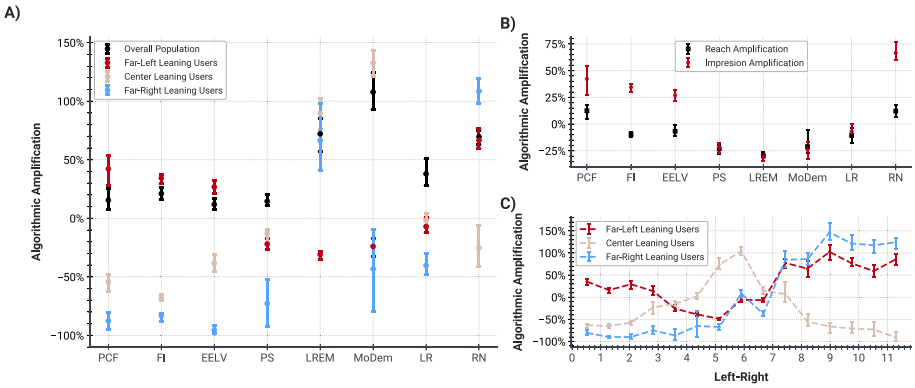

Fig. 5 Amplification of Members of Parliaments tweets impression in the timelines of Left, Center and Right leaning users as a function of their declared political groups (**A**) and of their positions on the Left-Right scale (**B**). **C** Reach and Impression Amplification of Members of Parliaments in the timelines of Far-Leaning Users. Vertical error bars correspond to 95% confidence intervals determined by bootstrapping over users

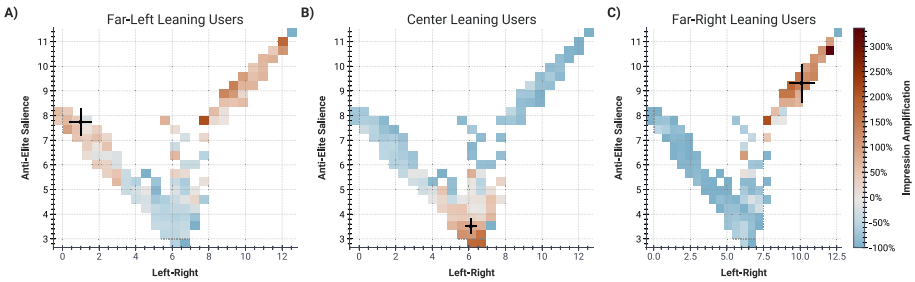

Fig. 6 Amplification of Members of Parliaments tweets impression in the timelines of Far-Left (**A**), Center (**B**), Far-right (**C**) leaning users, as a function of MPs position on the Left-Right/Anti-Elite plane. We display the average ideological position of users for each subgroup, error bars correspond to standard deviation of simulated users' ideological position

algorithmic amplification of concordant views. The number of far-left-leaning users who have seen at least one tweet from far-left "LFI" MPs in engagement-based timelines is 9.7 [7.8, 12.1]% smaller than in reverse-chronological timelines, compared to a 34.4 [29.6, 37.1]% increase in number of impressions.

To gain more granular insights into the impact on algorithmic curation over the political landscape, Fig. 5C showcases the impression amplification of MPs on the timelines of far-left, center and far-right-leaning users, as a function of their position on the Left-Right scale, rather than considering MPs declared political groups. Engagement-based timelines are made of a larger share of congenial tweets compared to reverse-chronological timelines, with an algorithmic amplification decreasing with the ideological difference. For far-right-leaning users, the amplification decreases as MPs lean more towards the left. For center-leaning users, the amplification decreases as MPs diverge from the center. Interestingly, despite this general trend, we notice that for far-left-leaning users, tweets from far-right MPs are amplified, till reaching 102.4 [85.1, 118.3]% for MPs positioned at 9 on the Left-Right axis. To elucidate this amplification, which contrasts with general trends, we consider a second political dimension, the Anti-Elite Salience.

Figure 6 displays the impression amplification of Members of Parliament as a function of their position on the Left-Right axis and Anti-Elite Salience, segmented by users

political leaning either far-left, center or far-right-leaning, we display left and right leaning users in annex Fig. 7. As observed along the Left-Right axis, for center and far-right-leaning users, we observe that engagement-based timelines favor tweets from the same ideological region as the users in the Left-Right/Anti-Elite 2D plane. The additional ideological dimension enlightens the amplification of some right-leaning MPs' tweets in the timelines of far-left-leaning users. Indeed, we notice that far-right MPs being amplified, distant in the Left-Right axis to far-left-leaning users, share a high Anti-Elite Salience with far-left-leaning users.

Discussion and conclusions

In this study, we employed machine-learning models trained on behavioral data obtained through a data-donation initiative to predict likes and retweets from users on a specific tweet. Through a comprehensive data collection effort, we constructed a simulated Twitter-like platform where timelines consist of messages posted by users' friends, ordered by decreasing likelihood of engagement and in reverse-chronological order for comparison. Departing from the analysis conducted in Bouchaud (2024), which characterized the differences between engagement-based and reverse-chronological timelines, in terms of content diversity, our study leverage this framework to examine how the conclusions drawn from audit depend on their methodology.

In particular, we examined the sensitivity of audits reliant on artificial accounts, commonly referred to as "sock-puppets", to two parameters arbitrarily defined in such audits: the number of friends the accounts follow and the session length, i.e., the number of tweets the accounts scroll through in a given session. Secondly, we evaluated how the measurement of algorithmic amplification of political content evolves across different levels of granularity. This involved segmenting users based on their political leaning, considering the position of Members of Parliament on a continuous Left-Right scale rather than their declared political groups, and incorporating additional political dimensions, notably the salience of Anti-Elite sentiment.

Sock-puppet audits: Our simulations illustrate the sensitivity of sock-puppet audits to the number of friends and session length of the mimicked accounts. As the number of friends increases, the deviation between engagement-based and reverse-chronological timelines becomes more pronounced, owing to recommender systems having a larger pool of messages to select from. With three-quarters of Twitter users following more than 150 accounts and one-third following over a thousand, audits involving sock-puppet accounts with only a few friends may offer limited insights into the effects of algorithmic curation experienced by real users.

Similarly, we observe that as artificial accounts scroll and engage in longer sessions, the deviation between engagement-based and reverse-chronological timelines diminishes. Given the absence of real log data, session length has been arbitrarily defined in previous research. For instance, Bandy and Diakopoulos (2021) report "at least 15" tweets per session, while Bartley et al. (2021) collected fifty tweets per session. Through our data-donation initiative, Horus, we have determined the empirical distribution of session length. On desktop, 74.4% of user sessions exceeded 15 tweets, while 72.8% were shorter than 50 tweets, with a median session length of 30 tweets. To mitigate the

arbitrariness associated with session length in future sock-puppet audits, we provide the empirical distribution of session lengths, in annex Table 2.

Ideological Scaling: Finally, we coupled our social media simulation with ideological scalings from the European Polarisation Observatory (Ramaciotti et al. 2021) to examine how audits at different levels of granularity yield differing conclusions. Particularly noteworthy, as outlined in Bouchaud (2024), is the significant impact of segmenting users based on their political leanings on the conclusions drawn. When computing the amplification of Members of Parliament across the entire population, the audit may conclude that the "mainstream political right enjoys higher algorithmic amplification than the mainstream political left" (Huszár et al. 2021). However, upon segmenting users based on their political leanings, we observe that engagement-based ranking merely favors content aligned with users' ideological preferences.

Moreover, in addition to considering the political groups declared by MPs, we incorporated their positions on continuous ideological dimensions identified from the MPs-follower bipartite graph. This heightened granularity confirms previously observed trends, indicating that algorithmic curation aimed at maximizing engagement tends to prioritize tweets from ideologically aligned MPs, with amplification decreasing as the ideological gap with users widens. We also underscore the necessity to move beyond the traditional Left-Right axis and consider additional political dimensions. In our simulations, we note that engagement-based timelines of far-left-leaning users contain a larger proportion of tweets from far-right-leaning MPs compared to reverse-chronological timelines. This counter-intuitive observation can be elucidated by considering a second political dimension, namely, the salience of Anti-Elite sentiment. Indeed, we observe that the far-right-leaning MPs being amplified share similar positions on this second axis.

Limitation: It is important to note that our models are not trained to replicate Twitter's recommender systems but rather to provide a framework allowing us to evaluate audit methodologies. Consequently, our simulations do present disparities with the findings of previous studies such as Huszár et al. (2021). For instance, while our simulations indicate amplification within the general population ranging from 11.8 to 107.8%, the audit made by Twitter researchers consistently reported amplification levels exceeding 100%, with figures reaching as high as 153.1%. This discrepancy may be attributed to the intricate interplay of factors within Twitter's recommender systems, which combine deep-learning models with hand-crafted heuristics. For instance, Twitter employs a reputation scoring mechanism for accounts, influenced by factors such as the number of followers and the age of the account (https://github.com/twitter/the-algorithm-ml/blob/main/projects/home/recap/README.md). Such heuristics could potentially lead to differential treatment of various account types, such as those of Members of Parliament. Moreover, our timelines were exclusively composed of tweets published or retweeted by users' immediate friends, without any injected content, unlike Twitter's timeline composition (https://blog.twitter.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm).

Additionally, our study does not account for feedback loops resulting from successive interactions between users and the platform. As a consequence, we do not aim to capture phenomena such as radicalization or the formation of echo chambers. Additionally, our simulation does not consider higher-order effects, which arise, for

instance, from retweeted tweets. Consequently, we were unable, and do not seek, to capture cascading effects in our simulations.

Furthermore, due to the insufficient amount of behavioral data for other engagement signals, we only considered likes and retweets as engagement signals for this study, weighting them equally to form the engagement score. In contrast, Twitter utilizes a significantly broader range of signals, weighted to maximize platform-wide metrics (https://github.com/twitter/the-algorithm-ml/blob/main/projects/home/recap/README.md). The present framework allows to explore in future work the effects of weightings the different engagement signals Milli et al. (2023).

Finally, to replicate the amplification metrics used in previous studies (Bandy and Diakopoulos 2021; Bartley et al. 2021; Huszár et al. 2021), we used reverse-chronological timelines as a baseline. As previously emphasized, such ranking is not neutral but tends to favor frequent posters. Future research could explore alternative ranking policies, such as bridging systems (Ovadya and Thorburn 2023).

In light of increasing regulatory scrutiny, such as that prescribed by the Digital Services Act in the EU, there is an urgent need to enhance methodologies for auditing social media platforms. This article showcases the use of social media simulation, empirically grounded through a training on real behavioral data, to explore algorithmic curation and algorithmic auditing.

Appendix

See Fig. 7, Tables 1 and 2.

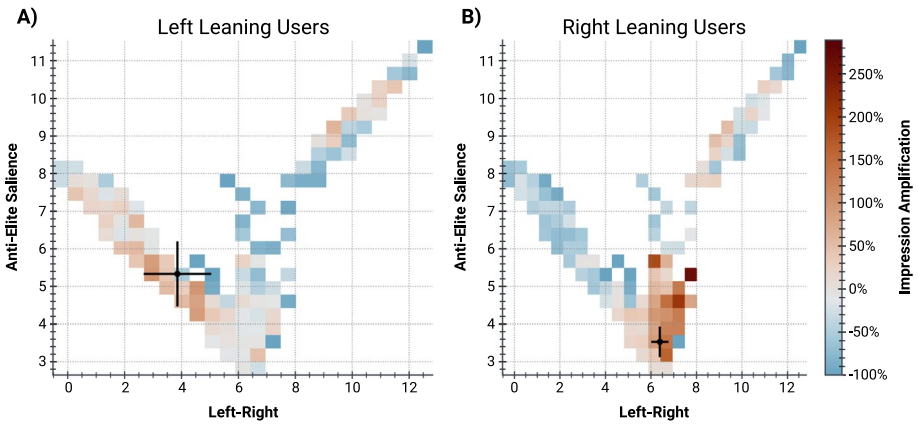

Fig. 7 Amplification of Members of Parliaments tweets impression in the timelines of left leaning (**A**) and right leaning (**B**) users as a function of MPs position on the Left-Right/Anti-Elite plane. We display the average ideological position of users for each subgroups, errorbars correspond to standard deviation of users ideological position

Table 1 Set of features used by our engagement-predictive models

Feature category	Features names and description
Impression related	'nb_min_since_publication': Number of minutes elapsed since the tweet was published
Tweets related	'tw_nb_characters': Number of characters in the tweet
	'tw_nb_words': Number of words in the tweet.
	'tw_mean_length_words': Mean length of words in the tweet
	'tw_nb_hashtag': Number of hashtags in the tweet
	'tw_nb_urls': Number of URLs included in the tweet.
	'tw_nb_mentions': Number of user mentions in the tweet
Author related	'author_created_days_ago': Number of days since the author's Twitter account was created
	'author_created_years_ago': Number of years since the author's Twitter account was created
	'author_followers_count': Number of followers the author has
	'author_friends_count': Number of accounts the author follows.
	'author_listed_count': Number of public lists the author is a part of
	'author_statuses_count': Number of tweets the author has posted
	'author_followers_count_rate': Number of followers divided by the number of days since account creation
	'author_friends_count_rate': Number of friends divided by the number of days since account creation
	'author_listed_count_rate': Number of list the authors is a part of divided by the number of days since account creation
	'author_statuses_count_rate': Number of tweet posted by the author divided by the number of days since account creation
	'author_default_profile_image': Binary indicator representing whether the author has a default profile image
	'author_verified': Binary indicator representing whether the author's Twitter account is verified
Relation to authors	'subjectFollowsAuthor': Binary indicator representing whether the subject follows the author
	'authorSubjectJaccard': Jaccard similarity coefficient between the author and subject's Twitter friends
	'authorSubjectOverlapCoef': Overlap coefficient between the author and subject's Twitter friends
Relation with past engagement	'author_pagerank_ratio_previously_rt': Ratio of PageRanks between the author of message and the last retweeted author
	'author_pagerank_ratio_previously_like':Ratio of PageRanks between the author of message and the last liked author
	'author_reduced_l2_previously_like': L2 norm between the author of message and the last liked author (in 8D follow space)
	'author_reduced_cosine_sim_previously_like': Cosine similarity between the author of message and the last liked author
	'author_reduced_l2_previously_retweet': L2 norm between the author of message and the last liked author
	'author_reduced_cosine_sim_previously_retweet': Cosine similarity between the author of message and the last retweeted author

Table 2 Distribution of session lengths, determined via data-donation

Session length	Distribution (%)
6	6.63
10	9.79
14	9.15
18	8.27
22	7.73
26	6.72
30	6.18
34	5.39
38	4.84
42	4.28
46	3.83
50	3.61
54	3.02
58	2.80
62	2.67
66	2.44
70	2.11
74	1.93
78	1.73
82	1.63
86	1.40
90	1.33
94	1.29
98	1.24

Truncated at 100 tweets, 2.6% of observed sessions are longer

Acknowledgements
Our study did not involve experimentation with human subjects, and all data used is publicly available through Twitter's API. Data declared the 19 March 2020 and 15 July 2021 at the registry of data processing at the *Fondation Nationale de Sciences Politiques* (Sciences Po) in accordance with General Data Protection Regulation 2016/679 (GDPR) and Twitter policy. For further details and the respective legal notice, please visit https://medialab.sciencespo.fr/en/activities/epo/.

Author Contributions
PB designed, developed and maintained the data donation program, and developed the research design and analysis. PR collected and treated data used of political opinion estimation. Both authors wrote the manuscript.

Funding
The data used in this study were provided by the Horus project at the Complex Systems Institute of Paris Ile-de-France (ISC-PIF) and by the "European Polarisation Observatory" (EPO) of the CIVICA Research (co-)funded by EU's Horizon 2020 programme under grant agreement No 101017201. P.B. acknowledges the Jean-Pierre Aguilar fellowship from the CFM Foundation for Research. P.R. acknowledges support by the Data Intelligence Institute of Paris (diiP) through the French National Agency for Research (ANR) grant ANR-18-IDEX-0001 "IdEx Université de Paris" and SoMe4Dem (Grant No. 101094752) Horizon Europe project.

Availability of data and materials
The data used to train the predictive models and the models themselves consist of non-public information and are not made accessible to the public in order to protect individuals' privacy according to Horus's privacy policy. The data used in the simulations was acquired via the Twitter API and cannot be made publicly accessible due to Twitter's developer policy. Code and aggregated data used to generate the figures are available from the corresponding author on reasonable request.

Declarations

Competing interests
The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

Received: 17 April 2024 Accepted: 21 August 2024
Published online: 16 September 2024

References

Bakshy E, Messing S, Adamic LA (2015) Exposure to ideologically diverse news and opinion on Facebook. Science 348(6239):1130–1132. https://doi.org/10.1126/science.aaa1160
Bandy J, Diakopoulos N (2021) More accounts, fewer links: How algorithmic curation impacts media exposure in twitter timelines. In: Proceedings of the ACM on human-computer interaction 5(CSCW1):1–28. https://doi.org/10.1145/3449152
Barberá P (2015) Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. Polit Anal 23(1):76–91. https://doi.org/10.1093/pan/mpu011
Barberá P, Jost JT, Nagler J, Tucker JA, Bonneau R (2015) Tweeting from left to right: is online political communication more than an echo chamber? Psychol Sci 26(10):1531–1542. https://doi.org/10.1177/0956797615594620
Barbiero P, Squillero G, Tonda A (2020) Modeling generalization in machine learning: a methodological and computational study
Bartley N, Abeliuk A, Ferrara E, Lerman K (2021) Auditing algorithmic bias on twitter. In: 13th ACM web science conference 2021. WebSci '21. ACM. https://doi.org/10.1145/3447535.3462491
Belli L, Ktena SI, Tejani A, Lung-Yut-Fong A, Portman F, Zhu X, Xie Y, Gupta A, Bronstein M, Deliç A, Sottocornola G, Anelli W, Andrade N, Smith J, Shi W (2020) Privacy-aware recommender systems challenge on Twitter's home timeline
Belli L, Tejani* A, Portman* F, Lung-Yut-Fong* A, Chamberlain B, Xie Y, Lum K, Hunt J, Bronstein M, Anelli VW, Kalloori S, Ferwerda B, Shi W (2021) The 2021 RecSys challenge dataset: fairness is not optional. In: RecSysChallenge '21: proceedings of the recommender systems challenge 2021. RecSysChallenge 2021. ACM. https://doi.org/10.1145/3487572.3487573
Benkler Y, Faris R, Roberts H (2018) Network propaganda: manipulation, disinformation, and radicalization in American politics. Oxford University Press, Oxford
Boeker M, Urman A (2022) An empirical investigation of personalization factors on TikTok. In: Proceedings of the ACM web conference 2022. WWW '22. ACM. https://doi.org/10.1145/3485447.3512102
Bouchaud P (2024) Skewed perspectives: examining the influence of engagement maximization on content diversity in social media feeds. J Comput Soc Sci. https://doi.org/10.1007/s42001-024-00255-w
Bouchaud P, Chavalarias D, Panahi M (2023) Crowdsourced audit of Twitter's recommender systems. Sci Rep 13(1):16815. https://doi.org/10.1038/s41598-023-43980-4
Bouchaud P (2024) Algorithmic amplification of politics and engagement maximization on social media, pp 131–142. Springer. https://doi.org/10.1007/978-3-031-53503-1_11
Brady WJ, Wills JA, Jost JT, Tucker JA, Van Bavel JJ (2017) Emotion shapes the diffusion of moralized content in social networks. Proc Natl Acad Sci 114(28):7313–7318. https://doi.org/10.1073/pnas.1618923114
Cardon D, Cointet J-P, Ooghe B, Plique G (2019) Unfolding the multi-layered structure of the French mediascape
Chandio S, Dar DP, Nithyanand R (2023) How auditing methodologies can impact our understanding of YouTube's recommendation systems
Chari T, Pachter L (2023) The specious art of single-cell genomics. PLoS Comput Biol 19(8):1011288
Chen AY, Nyhan B, Reifler J, Robertson RE, Wilson C (2023) Subscriptions and external links help drive resentful users to alternative and extremist YouTube channels. Sci Adv 9(35):eadd8080. https://doi.org/10.1126/sciadv.add8080
Covington P, Adams J, Sargin E (2016) Deep neural networks for YouTube recommendations. In: Proceedings of the 10th ACM conference on recommender systems. RecSys '16. ACM. https://doi.org/10.1145/2959100.2959190
Gaumont N, Panahi M, Chavalarias D (2018) Reconstruction of the socio-semantic dynamics of political activist twitter networks—method and application to the 2017 French presidential election. PLoS ONE 13(9):0201879. https://doi.org/10.1371/journal.pone.0201879
González-Bailón S, Lazer D, Barberá P, Zhang M, Allcott H, Brown T, Crespo-Tenorio A, Freelon D, Gentzkow M, Guess AM, Iyengar S, Kim YM, Malhotra N, Moehler D, Nyhan B, Pan J, Rivera CV, Settle J, Thorson E, Tromble R, Wilkins A, Wojcieszak M, Jonge CK, Franco A, Mason W, Stroud NJ, Tucker JA (2023) Asymmetric ideological segregation in exposure to political news on Facebook. Science 381(6656):392–398. https://doi.org/10.1126/science.ade7138
Greenacre M (2017) Correspondence analysis in practice. CRC Press, Boca Raton
Grover A, Leskovec J (2016) node2vec. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM. https://doi.org/10.1145/2939672.2939754 . https://doi.org/10.1145%2F2939672.2939754
Guess AM, Malhotra N, Pan J, Barberá P, Allcott H, Brown T, Crespo-Tenorio A, Dimmery D, Freelon D, Gentzkow M, González-Bailón S, Kennedy E, Kim YM, Lazer D, Moehler D, Nyhan B, Rivera CV, Settle J, Thomas DR, Thorson E, Tromble R, Wilkins A, Wojcieszak M, Xiong B, Jonge CK, Franco A, Mason W, Stroud NJ, Tucker JA (2023) How do social media feed algorithms affect attitudes and behavior in an election campaign? Science 381(6656):398–404. https://doi.org/10.1126/science.abp9364
Hargreaves E, Agosti C, Menasche D, Neglia G, Reiffers-Masson A, Altman E (2018) Biases in the Facebook news feed: a case study on the Italian elections. In: 2018 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM). IEEE. https://doi.org/10.1109/asonam.2018.8508659
Haroon M, Wojcieszak M, Chhabra A, Liu X, Mohapatra P, Shafiq Z (2023) Auditing YouTube's recommendation system for ideologically congenial, extreme, and problematic recommendations. In: Proceedings of the national academy of sciences 120(50). https://doi.org/10.1073/pnas.2213020120
Hosseinmardi H, Ghasemian A, Rivera-Lanas M, Horta Ribeiro M, West R, Watts DJ (2024) Causally estimating the effect of YouTube's recommender system using counterfactual bots. In: Proceedings of the national academy of sciences 121(8). https://doi.org/10.1073/pnas.2313377121

Hussein E, Juneja P, Mitra T (2020) Measuring misinformation in video search platforms: an audit study on YouTube. In: Proceedings of the ACM on human-computer interaction 4(CSCW1):1–27. https://doi.org/10.1145/3392854
Huszár F, Ktena SI, O'Brien C, Belli L, Schlaikjer A, Hardt M (2021) Algorithmic amplification of politics on Twitter. In: Proceedings of the national academy of sciences 119(1). https://doi.org/10.1073/pnas.2025334119
Jolly S, Bakker R, Hooghe L, Marks G, Polk J, Rovny J, Steenbergen M, Vachudova MA (2022) Chapel hill expert survey trend file, 1999–2019. Elect Stud 75:102420. https://doi.org/10.1016/j.electstud.2021.102420
Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y (2017) LightGBM: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst 30:3146–3154
Kmetty Z, Stefkovics A, Szamely J, Deng D, Aniko K, Omodei E, Edit P, Koltai J (2023) Determinants of willingness to donate data from social media platforms. https://doi.org/10.31219/osf.io/ncwkt
Lee AY, Mieczkowski H, Ellison NB, Hancock JT (2022) The algorithmic crystal: conceptualizing the self through algorithmic personalization on TikTok. In: Proceedings of the ACM on human-computer interaction 6(CSCW2):1–22. https://doi.org/10.1145/3555601
Lowe W (2008) Understanding wordscores. Polit Anal 16(4):356–371
Meta: Meta Reports Fourth Quarter and Full Year 2022 Results (2023) https://investor.fb.com/investor-news/press-release-details/2023/Meta-Reports-Fourth-Quarter-and-Full-Year-2022-Results/default.aspx. Accessed 24 Feb 2024
Milli S, Carroll M, Wang Y, Pandey S, Zhao S, Dragan AD (2023) Engagement, user satisfaction, and the amplification of divisive content on social media
Milli S, Pierson E, Garg N (2023) Choosing the right weights: balancing value, strategy, and noise in recommender systems
Ovadya A, Thorburn L (2023) Bridging systems: Open problems for countering destructive divisiveness across ranking, recommenders, and governance. Technical report, Knight First Amendment Institute . https://knightcolumbia.org/content/bridging-systems
Ramaciotti P, Lamarche-Perrin R, Fournier-S'Niehotta R, Poulain R, Tabourier L, Tarissan F (2021) Measuring diversity in heterogeneous information networks. Theoret Comput Sci 859:80–115
Ramaciotti Morales P, Cointet J-P, Muñoz Zolotoochin G, Fernández Peralta A, Iñiguez G, Pournaki A (2022) Inferring attitudinal spaces in social networks. Soc Netw Anal Min 13(1):14
Ramaciotti P, Cointet J-P, Zolotoochin GM8 (2021) Unfolding the dimensionality structure of social networks in ideological embeddings. In: Proceedings of the 2021 IEEE/ACM international conference on advances in social networks analysis and mining, pp 333–33
Rathje S, Van Bavel JJ, Linden S (2021) Out-group animosity drives engagement on social media. In: Proceedings of the national academy of sciences 118(26). https://doi.org/10.1073/pnas.2024292118
Ribeiro MH, Veselovsky V, West R (2023) The amplification paradox in recommender systems
Rossi WS, Polderman JW, Frasca P (2022) The closed loop between opinion formation and personalized recommendations. IEEE Trans. Control Netw. Syst. 9(3):1092–1103. https://doi.org/10.1109/tcns.2021.3105616
Sandvig C, Hamilton K, Karahalios K, Langbort C (2014) Auditing algorithms: research methods for detecting discrimination on internet platforms. Data Discrim Convert Crit Concerns Product Inq 22(2014):4349–4357
Satuluri V, Wu Y, Zheng X, Qian Y, Wichers B, Dai Q, Tang GM, Jiang J, Lin J (2020) Simclusters: community-based representations for heterogeneous recommendations at Twitter. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery &; data mining. KDD '20. ACM. https://doi.org/10.1145/3394486.3403370
Stewart E, Hartmann D (2020) The new structural transformation of the public sphere. Sociol Theory 38(2):170–191
Twitter: What Twitter learned from the RecSys 2020 challenge. Twitter. https://blog.twitter.com/engineering/en_us/topics/insights/2020/what_twitter_learned_from_recsys2020
Twitter: Twitter/the-Algorithm: Source Code for Twitter's recommendation algorithm: Heavy Ranker. https://github.com/twitter/the-algorithm-ml/blob/main/projects/home/recap/README.md
Twitter: Twitter's recommendation algorithm. Twitter. https://blog.twitter.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm
Van Bavel JJ, Rathje S, Harris E, Robertson C, Sternisko A (2021) How social media shapes polarization. Trends Cogn Sci 25(11):913–916. https://doi.org/10.1016/j.tics.2021.07.013
Wang Y, Huang H, Rudin C, Shaposhnik Y (2021) Understanding how dimension reduction tools work: an empirical approach to deciphering t-SNE, UMAP, TriMAP, and PaCMAP for data visualization. J Mach Learn Res 22(201):1–73

Publisher's Note