# Crowdsourced Audit of Twitter's Recommender Systems Impact on Information Landscapes.

Paul Bouchaud, David Chavalarias, Maziyar Panahi

# Crowdsourced Audit of Twitter's Recommender Systems Impact on Information Landscapes.

**Paul Bouchaud** [a,b,*], **David Chavalarias** [a,b], **and Maziyar Panahi** [a]

[a]CNRS, Complex Systems Institute of Paris Île-de-France (ISC-PIF), 75013 Paris, France
[b]EHESS, Center for Social Analysis and Mathematics (CAMS), 75006 Paris, France
[*]paul.bouchaud@iscpif.fr

## ABSTRACT

This study conducts an audit of Twitter's recommender system and its impact on the information landscapes. Using a desktop browser extension and large-scale data collection, we compare the information landscapes depicted by the recommender system to the information landscape the users decided to subscribed to. Our findings reveal algorithmic distortions, including uneven amplification based on political leaning, amplification of friends from the same community, and preferential amplification of toxic and sentimentally valenced tweets. This audit emphasizes the need for transparency, regulation, and awareness of the implications of recommender systems in digital services.

## Introduction

28% of the global population has adopted social media as its main gateway for online news in 2022. On these platforms, *Newsfeeds* have become the main entry point for their users: a place where the information coming from their social environment is curated and displayed. The large volume of produced content and the wish for the platforms to maximize profit through user engagement led to the deployment of recommender systems, selecting the content to be shown. If such recommendation systems, acting as attention-allocators[1], are biased, the access to information of millions of citizens could be biased as well, leading to systemic risks for society. The study of Huszár et al[2] precisely revealed that in the case of Twitter, its recommender unevenly amplified politician tweets' reach depending on their ideological leaning. To achieve this demonstration, they leveraged proprietary information on Twitter users and a years-long experiment, with a controlled group —not exposed to Twitter recommender— of nearly two million users. For lack of direct access to data, independent audits of social networks recommender from academia has been addressed mostly through so-called "sock-puppet audit", creating artificial users and scrapping the platform content. While providing interesting insights into the distortion caused by recommenders, such audits are limited by the number of fake accounts that researchers can create —[3,4] used 8 accounts in their demonstration— and their ability to realistically mimic human digital behavior —usually ad hoc heuristics. Enlisting volunteers to provide their data —made more easily accessible to them in recent years thanks to legislative progress such as with the GDPR— seems to gradually become a promising avenue in digital services independent external audits[5,6]. Yet, the relative lack of control is a common drawback of a purely crowd-sourced audits.

In order to overcome these constraints, we employed a combination of a desktop browser extension, which captured the content of Twitter *feeds* from volunteers, and a comprehensive data collection using the Twitter API. This approach enabled us to reconstruct the set of messages that participants' friends had published and to which participants could have potentially been exposed to. In a prior investigation, Huszár et al[2] conducted a study that explored the influence of Twitter recommender systems on the reach of *tweets* authored by political figures; by comparing algorithmically curated timelines versus those ranked in reverse-chronological order. On the contrary, our focus is on the information landscape depicted by the Twitter recommender system, through its selection of *tweets* from the pool of messages published by users' friends. The aim of the Twitter algorithm is to select "relevant" content that users are likely to engage with[7]. We do not claim to comprehensively investigate the intricate socio-technical aspects or unravel feedback loops[8], our goal is to gauge the disparities between the information landscape portrayed by the recommender system and the user's subscription choices. In this paper, we specifically address a few algorithmic distortions, explore longstanding questions, and emphasize the importance of collecting user impressions. Our findings reveal that Twitter's recommender system: 1) significantly amplifies small accounts and those with limited content production, 2) unequally amplifies *tweets* from users' friends based on their political leaning, distorting the political landscape perceived by the users compared to their subscriptions, 3) greatly amplifies friends who belong to the same community as the user, and 4) amplifies toxic and sentimentally valenced *tweets* while diminishing the visibility of neutral ones.

## Methods

### Context

Our analysis was conducted prior to the release of the overall architecture and partial source codes of Twitter's recommender systems on March 31, 2023, a preliminary pre-print can be found at hal-04036232v2. The current analysis has been re-performed over a more recent timeframe, specifically 07/03/23-06/04/23 for account-level features and 14/01/23-07/02/23 for tweet-level features, as explained below. Among the numerous hard-coded heuristics and general insights, Twitter engineers have specifically highlighted the significance of community detection in the recommendation process[9]. In light of this, we have conducted an additional analysis to examine whether *tweets* from friends belonging to the same community as the participant are amplified compared to those from different communities.

### Data Collection

We developed a browser-extension called "Horus", compatible with Chrome and Firefox, that allows us to capture various data, including Twitter feeds, displayed on participants' desktop screens. The participants were self-selected, they chose to take part in the study after becoming aware of it through newspaper articles and radio broadcasts in the fall of 2022. To expand our reach and attract a wider range of individuals, we also employed online advertising on Twitter, presenting the initiative and encouraging individuals to participate. In addition to contributing to scientific research, the primary incentive for participants to install the extension was receiving a personalized report on the political diversity of their Twitter friends and their curated *feed*. This report was sent to volunteers once a sufficient amount of data had been collected. Prior to the data collection process, participants were fully informed about the study's objectives, the specific data that would be collected, and their informed consent was obtained. Following this, the extension started gathering the participants' Twitter *feed*.

Our cohort is, by design, not representative of the Twitter audience; the data collection, performed through a desktop browser extension, is filtering out mobile users. Nevertheless, the sanity of a platform should be maintained across devices and users' behavior, justifying external audits —even partial ones like ours. Nevertheless, by comparing the distribution of political leaning among participants' friends with the distribution among friends of 5k randomly selected French Twitter accounts, we verified that, political-group-wise, our set of participant does not significantly differ from a random sampling of Twitter French users. We present in the Supplementary Information various statistics on our cohort of participant

Taking the participants having been active on the desktop version of Twitter between March 3, 2023, and April 6, 2023 as our only objects of study, the analysis has been performed on $N = 463$ participants.On average, our participants followed 682 [22,2712] accounts (5-95 percentiles). In conjunction with the crowd-sourced data collection, we leveraged the Twitter API to fetch additional information. This included the number of *tweets* published, within the considered timeframe, by the 42k accounts followed by at least two participants. This subset of accounts represented 61.8% of the participants' friends. The accounts under consideration have a median weekly *tweet* count of three, we present the cumulative distribution function of their publishing rate in Supplementary Information. Additionally, we retrieved the set of 3 million *tweets* published by the 14k accounts followed by at least three participants. This setting aims to balance the data collection burden while considering a large fraction of the participants' friends.

Finally, the partial open sourcing of the Twitter recommender system[9] having revealed the importance of follow-graph-derived features[10], we sampled the Twitter follow graph using the Twitter API and a snowball approach. Starting with our participants' friends and a large random sample of Twitter French accounts as seeds, we fetched up to 5k followers and followees for each account. We repeated this process iteratively, including the newly fetched accounts, until we obtained more than 220k seeds. The resulting follow network consisted of 41 million nodes and 360 million edges. After pruning nodes with a degree less than 5, we were left with a follow network comprising 6 million nodes and 303 million edges. To analyze the graph, we employed the *node2vec* algorithm[11] for node embedding, performed cluster detection using *HDBSCAN*[12] after applying dimensionality reduction with *UMAP*[13]. For a visual representation of the clustered network, please refer to the Supplementary Information.

### Quantification of algorithmic amplification

We defined the algorithmic amplification to measure the extent to which the *tweets* authored by accounts in a subset $G \subseteq F$ of a participant's set of friends $F$ are selected for display by Twitter recommender systems, compared to the overall set of friends' tweets. We adopted a formulation equivalent to the one of Huszár et al[2] but considering the point of view of the receiver, i.e. the participant, instead of the authors'. The algorithmic amplification is computed as follow:

$$a(G) = \left( \frac{N^{\text{impressed}}_{G \subseteq F}}{N^{\text{published}}_{G \subseteq F} \times a_F} - 1 \right) \times 100\%$$

Here, $N^{\text{impressed}}_{G \subseteq F}$ represents the number of messages published by accounts in the subset $G \subseteq F$ that have been displayed on the participant's desktop screen. $N^{\text{published}}_{G \subseteq F}$ denotes the total number of messages published by accounts in the subset $G \subseteq F$. $a_F$ is a neutral baseline, it represents the fraction of messages published by the participant's friends having been displayed on the participant's screen, $N^{\text{impressed}}_F / N^{\text{published}}_F$, as captured by the extension.

Regarding this metric, it is important to consider the following points: 1) In order to avoid potential ambiguity, we excluded re*tweets* from our study. This decision, also took by Huszár et al[2], was made because the attribution of amplification in re*tweets* can be challenging. 2) Although the collected data would have allowed for a fine temporal analysis, we chose to calculate the algorithmic amplification after aggregating participants' sessions over a span of two weeks. This was done because the preference for recent content (three-fourths of the displayed content is less than 12 hours old) is ultimately an arbitrary heuristic that should not be ignored when assessing whether algorithmic curation distorts the landscape of content production. 3) Twitter's "For You" timelines consist, on average, of 50% *tweets* from friends $F$ and 50% from friends of friends (second neighbors)[9]. However, the user having actively decided to follow only its first neighbors, we then consider them as natural baseline, and restrict the computation of algorithmic amplification to them. Future works will investigate this out-of-network content.

We normalize the ratio such that an amplification value of $a(G) = 0\%$ indicates that the *tweets* from accounts in $G \subseteq F$ are displayed in proportion to their representation in the pool of *tweets* published by one's friends. An amplification ratio of $a(G) = 50\%$ means that the fraction of *tweets* appearing in the timelines, authored by accounts in $G \subseteq F$, is 50% higher than the fraction they represent in the pool of messages authored by one's friends.

To distinguish between the effect of the recommender system and the effect of considering only a subset of $|G \subseteq F|$ accounts when computing the algorithmic amplification, we compute the algorithmic amplification associated to random subsets $\tilde{G} \subseteq F$ of same cardinality ($|\tilde{G} \subseteq F| = |G \subseteq F|$); Mann–Whitney U tests between the bootstrapped amplification distributions assess the significance of the algorithmic amplification $a(G)$. Our data collection coverage being partial, we performed bootstrapping over participants' friends to generate robust estimates at the participant level and a bootstrapping over participants to derive robust collective measures. We provided 95% bootstrap confidence intervals for all amplification measures.

Finally, we define the algorithmic amplification of tweet-related features analogously, substituting $G \subseteq F$ with the set of *tweets* published by accounts in $F$ that exhibited a particular characteristic, such as a high engagement rate or being "toxic".

#### *Toxicity and Sentiment Analysis of tweets*

We examined the algorithmic amplification of "toxic" *tweets* , such as insults, threats, and obscenities. To identify these toxic *tweets* , we utilized Detoxify[14], an open-source natural language processing model trained on Google's Jigsaw toxic comments database[15]. Significant attention was dedicated to minimizing unintended biases during the training process, biases such as racial ones being present in the training corpuses as highlighted by Davidson et al[16]. Additionally, we employed XLM-T[17], a multilingual language model, to perform sentiment analysis on the tweets. This analysis categorized the *tweets* as either positive, negative, or neutral.

#### *Account Political Orientation*

We consider the algorithmic amplification of friends'*tweets* as a function of the political leaning of their authors. The estimation of political orientations was conducted using the Politoscope database, having gathered more than 700 million *tweets* related to French politics since 2016.

Shortly, the French political landscape is characterized by a multipolar structure, with the current French president occupying a centrist position. The opposition consists of a broad left coalition on one side and a far-right group on the other. Notably, anti-system activists play a role in bridging the divide between far-left and far-right militants, resulting in a circular political landscape[18]. It is crucial to consider this circularity when analyzing potential patterns of amplification. For instance, an account with a far-left leaning may exhibit stronger affinities with far-right content rather than with centrist content.

To determine the political leanings of Twitter accounts, we leveraged the re*tweet* graph associated with *tweets*, published in 2022, from French political figures and/or containing French political keywords. The collection procedures for this dataset are detailed in Gaumont et al[19], re*tweets* were shown to be reliable indicators of ideological alignment. We employed the node2vec

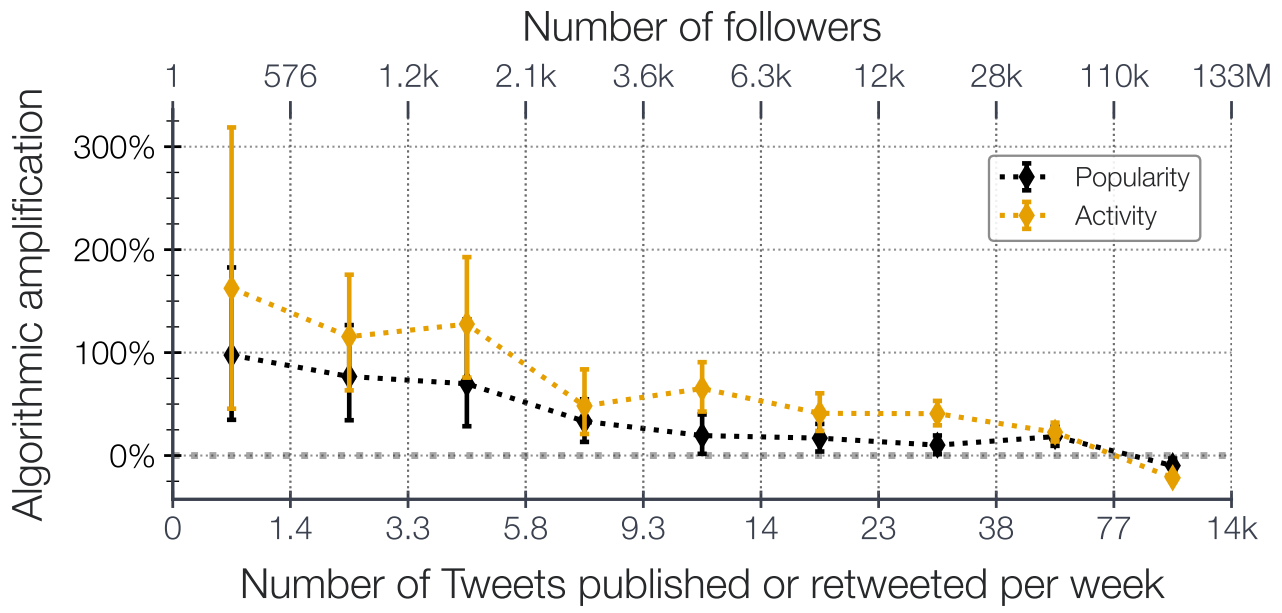


**Figure 1.** Algorithmic amplification of accounts depending of their number of followers, and of their activity in terms of number of *tweets* published or retweeted weekly (on average since the account creation), binned into deciles. Error-bars correspond to 95% bootstrap confidence interval of the amplification. The bold line corresponds to zero amplification.

algorithm[11] to generate embeddings of the nodes, capturing the underlying structure of the re*tweet* network. Kojaku et al[20] shown that such approach can indeed capture the community structure down to the theoretical community detectability limit. Subsequently, we calculated the angular similarity in the latent space between the nodes embedding of French political figures and the ones of the 1.2 million Twitter accounts that have either repeatedly published or retweeted political content, during the 2022 period (during which the French presidential and legislative elections occurred). Based on these similarities, we assigned a numerical political leaning to each account, ranging from -1 (for left-leaning accounts) to +1 (for right-leaning accounts); supporters of the current French President Emmanuel Macron cluster around zero. The numerical scale aligns with both the political group of members of parliament and the assessment of political experts (2019 Chapel Hill Expert Survey[21]), as well as a clustering analysis[19]. The computation details and validation tests are provided in the Supplementary Information.

### Data Availability

We adhere strictly to both Twitter's developer policy and Horus's privacy policy. As per these policies, we offer aggregated data upon request from the corresponding author. This data is provided to enable the reproduction of our Figures.

## Results

### Twitter algorithmic curation increases authors' representation inequality

As a first, high level, illustration of the shaping power of the recommender on what is seen by Twitter users, we evaluated the Gini coefficient over the number of *tweets* published and impressed by participants' friends. We find that despite an already highly unequal situation (Gini coefficient of friends' publications of .72 [.56,.87]) in which the 10% most active participants' friends publishes more than half of the entire set of participants' friend tweets, Twitter recommender system amplifies these inequalities, increasing the Gini coefficient by 14% on average (Gini coefficient of friends' impression of .83 [.59,.99]). The *tweets* from the 10% most shown friends represent more than 70% of participants' friends impressions in their timelines. We display in the Supplementary Information the associated Lorentz curves.

### Small accounts benefit from higher algorithmic amplification

As displayed on Figure 2, *tweets* of accounts having less than 576 followers (first decile) are amplified by +97.5 [34.7.2,182.7] %. Put differently, the proportion of *tweets* authored by small accounts is twice larger in the timelines than in the overall pool of

messages authored by participants' friends. Conversely, *tweets* of accounts with a number of followers larger than 110k are lessen by $-9.8$ [−16.9,−2.9] %. Similarly, *tweets* of accounts having published on average less than 1.4 *tweets* per week since their creation (first decile) are significantly amplified when they do publish, with an amplification of $+162.4$ [45.5.2,318.7] %. The *tweets* of highly active accounts, more than 10 *tweets* per day on average, are lessen by $-21.8$ [−25.3,−17.9] %.

### Algorithmic curation distort the political landscape

After having estimated the political leaning of participants' friends by analyzing their re*tweets* of political content, we segmented the participants based on their own political orientation; as self-declared through a form crossed with their Twitter friends orientation. Our findings reveal that for participants leaning towards the far-left (N=51) and left/center-left (N=92), Twitter's recommender system amplifies ideologically aligned friends, see Figures 2.A & 2.B.

As display on Figure 2.A, for far-left participants, the messages published by far-left friends are amplified by $+21.8$ [5.0,42.6] %. The amplification decreases as the opinion difference increases, until it reaches $-11.2$ [−32.8,12.4] % for right leaning accounts' tweets. Similarly, for left/center-left participants the *tweets* stemming from further-left or from right-leaning friends are algorithmically lessen, respectively by $-10.2$ [−27.7,7.4] % and $-31.4$ [−45.7,−19.9] %, while *tweets* from ideologically aligned friends are amplified $+21.1$ [10.6,32.6] %, see Figure 2.B. Interestingly, for center-right participants (N=33), see Figure 2.C, the opposite effect is noticed, the exposition to ideologically aligned accounts *tweets* is lessened by $-23.8$ [−43.9,−2.6] %, while far-left and further right *tweets* are highly amplified, respectively by $+44.1$ [−28.1,145.6] % and $+88.0$ [31.3,168.0] %. We assessed the statistical significance of the different amplification patterns, across participants' political leaning, through permutation tests. We have an insufficient number of far-right participants to derive meaningful statistical insights, once an adequate number of participants is obtained, we will conduct an extended analysis to explore this group further.

### Algorithmic curation prevents diversity

The proportion of *tweets* impressed in the timelines, stemming from friends belonging to the same community (the same cluster in the network of follow) is significantly greater than in the overall pool of published messages. On the other hand, *tweets* authored by accounts from different communities are fairly displayed. This behavior holds when considering a variety of community detection resolutions, namely by changing the smallest size grouping we consider as a cluster in *HDBSCAN*[12]. When considering a minimum cluster size of 100, we detect 352 communities in our follow graph, only 4.1% of our participants' friends belonged to the same cluster as the participant; the *tweets* stemming from this small fraction of friends are shown twice as much as, in proportion, in the timelines than what they represent in the pool of friends' messages ($+100.2$ [43.6,175.4]%). The algorithmic amplification decays as the clusters are getting larger; when considering a minimum cluster size of 600, we detect 82 communities, 10.8% of our participants' friends belonged to the same cluster as the participant and their *tweets* are amplified by $+40.8$[17.9,67.6] %. We present in the Supplementary Information the amplification for various intermediary community detection resolutions.

### Algorithmic curation amplify toxic and sentimatally valence tweets

The fraction of toxic *tweets* (insults, threats, or obscenities) published by participants' friends is approximately 2.2%. During the period 14/01/23-07/02/23, the proportion of toxic *tweets* in participants' timelines is significantly higher, with an amplification of $+48.7$ [37.6,60.8] % compared to the overall pool of messages published by participants' friends. It is important to note that there is considerable variability in the amplification among participants, with some individuals being exposed to more than twice the proportion of toxic tweets. However, there is no significant correlation between the amplification at the participant level and the proportion of toxic *tweets* published by their friends.

It is worth mentioning that platform-wide, toxic *tweets* receive more than twice the number of replies and likes per impression compared to non-toxic tweets, and experience only a 10-20% decrease in re*tweets* and quotes per impression. Sentimentally valenced tweets, labeled by XLM-T[17] as either positive or negative, experience amplification of $+2.0$ [−0.3,4.4] % and $+5.8$[3.5,8.1] %, respectively, while neutral *tweets* are reduced by $-8.7$[−11.5,−5.9] %. During the period from December 9, 2022, to January 9, 2023 (N=101), toxic *tweets* were amplified by $+32.0$ [21.7,42.7] % (the statistical significance of these differences was confirmed through Mann-Whitney U tests).

### Algorithmic curation distort perceived *tweets* popularity

Figure 3 illustrates the amplification of *tweets* published between 14/01/23 and 07/02/23, based on their platform-wide engagement rate calculated weeks after their publication to ensure metric stability. We report in the Supplementary Information *tweets* statistics and the engagement rate for each quantile. We observe distinct patterns based on different types of engagements.

First, *tweets* with no engagements are significantly quieten, with an amplification of $-88.1$ [−90.5,−85.6]% for null like rate *tweets* and $-39.2$[−42.2,−36.2] % for null quote rate ones. For *tweets* with quote and re*tweet* engagements, the amplification

remains relatively constant at around 110% and 75%, respectively, for the first ten dodeciles. However, it decreases to 34.2 [18.9,49.5] % and 34.1[20.9,48.6] % at the eleventh dodecile. Notably, the algorithmic amplification is more sensitive to the like and reply rates. The amplification increases with the engagement rate, reaching a peak at the fifth dodecile for like rate and the seventh dodecile for reply rate. In the last dodecile, which corresponds to *tweets* with reply rates higher than 1.67% and quote rates higher than 2.83%, the amplification is significantly reduced. *tweets* with high reply rates experience a decrease of $-97.4$ [−98.5,−96.1] %, while *tweets* with high quote rates see a decrease of $-44.5$[−54.6,−30.8] %.

## Discussion

At the accounts level, Twitter recommender system favors *tweets* stemming from small accounts and/or accounts scarcely publishing content. This behavior may be a heuristic approach to prevent *feeds* from being dominated by spamming or excessively popular accounts. On one hand, it gives every user the opportunity to be heard or at least have the hope of being heard. On the other hand, it provides an advantage to entities employing astroturfing tactics, where their online presence is artificially amplified through small fake accounts that promote their ideas. Also, Twitter recommender favors *tweets* stemming from accounts in the same community than the user.Within Twitter follow graph, accounts belonging to the same community exhibit a tendency to share common interests[9], tendency leveraged by Twitter in its recommendation process[10]. Our intention is not to express a normative stance on the intrinsic value of cross-cutting exposure, an extensive literature provides a comprehensive and detailed understanding of such exposure[22–24]. Rather, we are highlighting the observation that the observed algorithmic amplification, go against users' choice to be exposed to a diverse range of views, if not by concealing "dissonant" content, by overwhelmingly amplifying consonant one.

Similarly, we observe that Twitter's recommender system presents a political landscape that differs from the one users have actively chosen to subscribe to. We shed light on the amplification patterns specific to different political communities. Considering the overall objective function of the Twitter recommender, it can be hypothesized that these patterns are the one found to maximize user engagement on the platform.

At the *tweet* level, we can reasonably hypothesise that it is because toxic *tweets* have higher reply and like engagement rates that they are preferably selected by the recommender, leading to the observed +48% amplification of such toxic content. However, despite the recommender's goal of maximizing engagement, it appears that once *tweets* reach a certain level of popularity, they are no longer recommended. We suggest that this may be because Twitter is designed to promote new content and maintain user engagement, rather than prolonging ongoing conversations. This tension between promoting popular content and promoting a diverse range of recent content on Twitter can result in situations where popular *tweets* are no longer recommended and may be perceived as being "shadow-banned".

Our audit show that Twitter recommender system has systemic effects on the information landscape. It tends to display these ecosystems as more toxic than they actually are, and it distort the representation of political groups among each other. Additionally, the amplification of small accounts can make the digital space more susceptible to manipulative practices such as astroturfing. However, recommender systems remain complex entities with numerous features and data points, further studies are necessary to unravel the intricacies of these systems. Confounding factors abound, and our audit only captures some of the resulting distortions in the end product. While enhancing the transparency of recommender system designs, such as through open sourcing the algorithms, may offer insights into the internal mechanisms behind skewed suggestions, independent audits with access to large-scale data will remain indispensable in regulating digital services, as specified by the 40th article of the European law on digital services[25].

## References


1. Ovadya, A. & Thorburn, L. Bridging systems: Open problems for countering destructive divisiveness across ranking, recommenders, and governance (2023). 2301.09976.

2. Huszár, F. *et al.* Algorithmic amplification of politics on twitter. *Proc. Natl. Acad. Sci. U.S.A.* **119**, DOI: 10.1073/pnas.2025334119 (2021).

3. Bandy, J. & Diakopoulos, N. More accounts, fewer links. *Proc. ACM Hum.-Comput. Interact. on Human-Computer Interact.* **5**, 1–28, DOI: 10.1145/3449152 (2021).

4. Bartley, N., Abeliuk, A., Ferrara, E. & Lerman, K. Auditing algorithmic bias on twitter. In *13th ACM Web Science Conference 2021*, DOI: 10.1145/3447535.3462491 (ACM, 2021).

5. Hargreaves, E. *et al.* 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)*, DOI: 10.1109/asonam.2018.8508659 (IEEE, 2018).

6. Sanna, L., Romano, S., Corona, G. & Agosti, C. YTTREX: Crowdsourced analysis of YouTube's recommender system during COVID-19 pandemic. In *Information Management and Big Data*, 107–121, DOI: 10.1007/978-3-030-76228-5_8 (Springer International Publishing, 2021).

7. Twitter. About your for you timeline on twitter (2023).

8. Wagner, C. *et al.* Measuring algorithmically infused societies. *Nature* **595**, 197–204, DOI: 10.1038/s41586-021-03666-1 (2021).

9. Twitter. Twitter's recommendation algorithm.

10. Satuluri, V. *et al.* Simclusters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, DOI: 10.1145/3394486.3403370 (ACM, 2020).

11. Grover, A. & Leskovec, J. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, DOI: 10.1145/2939672.2939754 (ACM, 2016).

12. McInnes, L., Healy, J. & Astels, S. hdbscan: Hierarchical density based clustering. *The J. Open Source Softw.* **2**, DOI: 10.21105/joss.00205 (2017).

13. McInnes, L., Healy, J., Saul, N. & Großberger, L. Umap: Uniform manifold approximation and projection. *J. Open Source Softw.* **3**, 861, DOI: 10.21105/joss.00861 (2018).

14. Hanu, L. & Unitary team. Detoxify. Github. https://github.com/unitaryai/detoxify (2020).

15. Ian Kivlichan, J. E. L. V. M. G. P. C., Jeffrey Sorensen. Jigsaw multilingual toxic comment classification (2020).

16. Davidson, T., Bhattacharya, D. & Weber, I. Racial bias in hate speech and abusive language detection datasets (2019). 1905.12516.

17. Barbieri, F., Espinosa Anke, L. & Camacho-Collados, J. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 258–266 (European Language Resources Association, Marseille, France, 2022).

18. Chavalarias, D., Bouchaud, P. & Panahi, M. Can few lines of code change society? beyond fack-checking and moderation: how recommender systems toxifies social networking sites. *(under review)* DOI: https://doi.org/10.48550/arXiv.2303.15035 (2023).

19. Gaumont, N., Panahi, M. & Chavalarias, D. Reconstruction of the socio-semantic dynamics of political activist twitter networks—method and application to the 2017 french presidential election. *PLoS ONE ONE* **13**, e0201879, DOI: 10.1371/journal.pone.0201879 (2018).

20. Kojaku, S., Radicchi, F., Ahn, Y.-Y. & Fortunato, S. Network community detection via neural embeddings (2023). 2306.13400.

21. Jolly, S. *et al.* Chapel hill expert survey trend file, 1999–2019. *Elect. Stud.* **75**, 102420, DOI: 10.1016/j.electstud.2021.102420 (2022).

22. Lu, Y. & Myrick, J. G. Cross-cutting exposure on facebook and political participation. *J. Media Psychol.* **28**, 100–110, DOI: 10.1027/1864-1105/a000203 (2016).

23. Bail, C. A. *et al.* Exposure to opposing views on social media can increase political polarization. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 9216–9221, DOI: 10.1073/pnas.1804840115 (2018).

24. Schneider, F. M. & Weinmann, C. In need of the devil's advocate? the impact of cross-cutting exposure on political discussion. *Polit Behav* **45**, 373–394, DOI: 10.1007/s11109-021-09706-w (2021).

25. Regulation (eu) 2022/2065 of the european parliament and of the council of 19 october 2022 on a single market for digital services and amending directive 2000/31/ec (digital services act). *OJ* **L 277**, 1–102 (27/10/2022).


## Acknowledgements


Paul Bouchaud acknowledges the Jean-Pierre Aguilar fellowship from the CFM Foundation for Research. This work was supported by the Complex Systems Institute of Paris Île-de-France and the Region Île-de-France.

## Author contributions statement

P.B. designed the study, carried out the crowdsourced data collection, the data analysis, and wrote the article. M.P. designed the Politoscope big data platform. All authors contributed to the final version of the article.

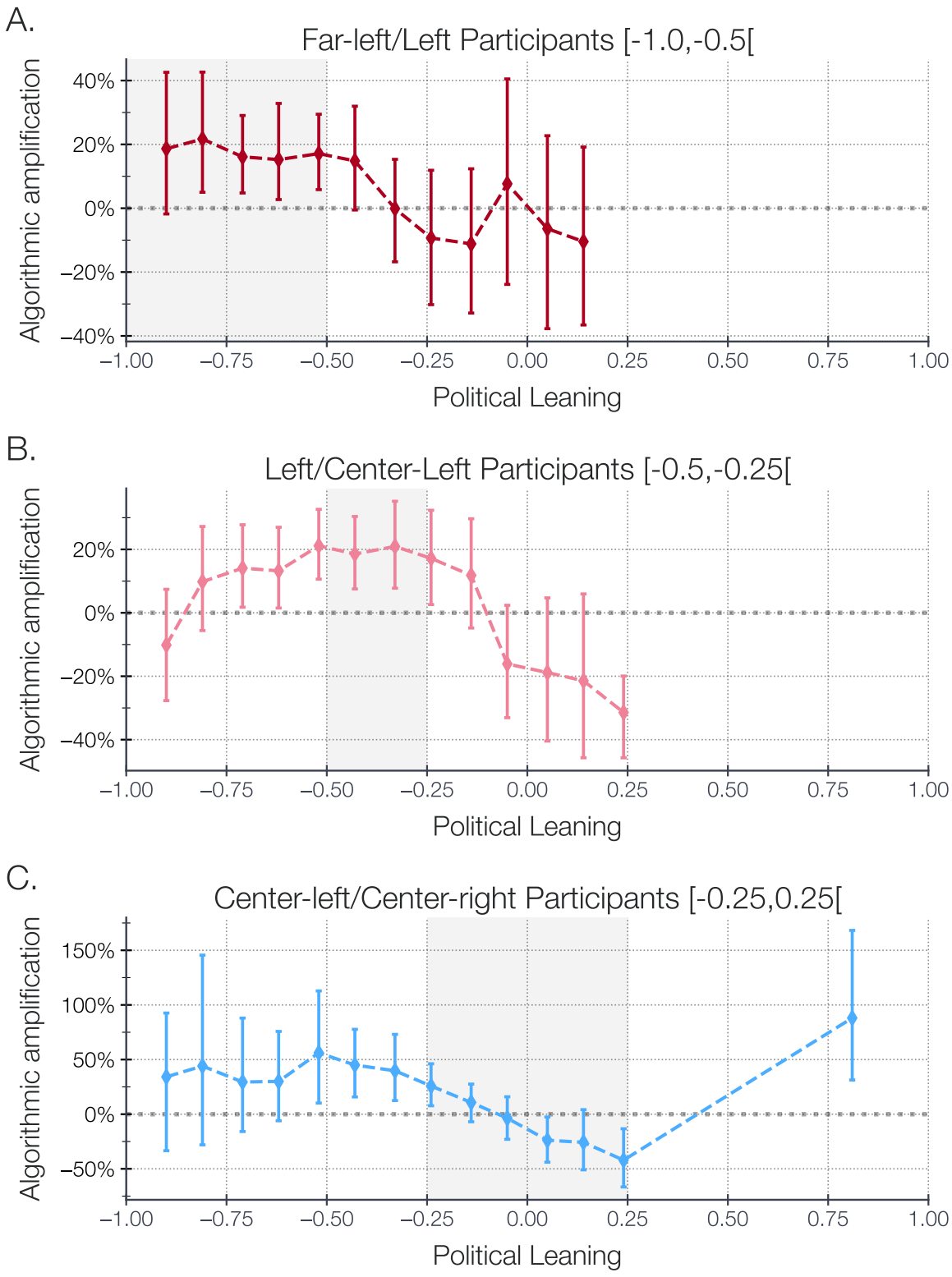


**Figure 2.** Algorithmic amplification of accounts depending of their political leaning (aggregation windows of 0.2, with successive half overlap), segmenting participants by political orientation. Error-bars correspond to 95% bootstrap confidence interval of the amplification. We shade the range of participants' opinion for each political leaning. The bold line corresponds to zero amplification. Only statistically significant points are displayed.

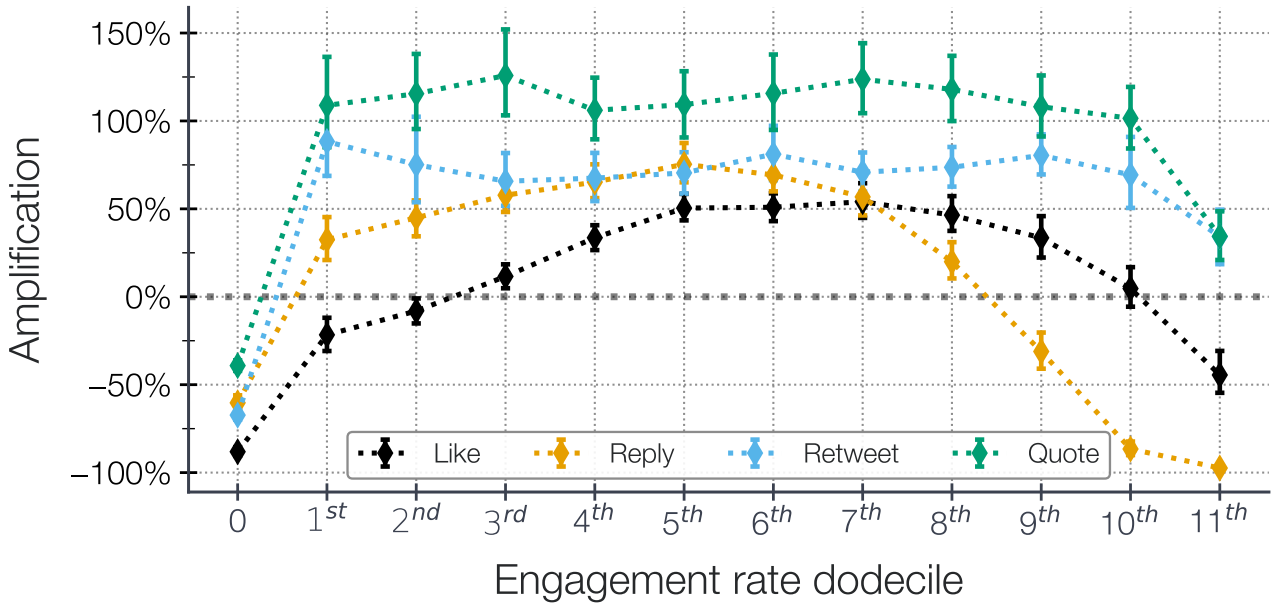


**Figure 3.** Amplification of *tweets* depending of their of engagement rate (B). We display the amplification for *tweets* having no engagement and binned in dodeciles the remainging engagement rates. Error-bars correspond to 95% bootstrap confidence interval of the amplification. The bold line corresponds to zero amplification.