




Is Twitter’s recommender biased ? An audit.

Paul Bouchaud A,B,* 
David Chavalarias A,B 
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

* corresponding author: paul.bouchaud@iscpif.fr

March 8, 2023

Abstract

Combining crowd-sourced data donation and a large-scale server-side data collection, we provide quantitative experimental evidence of Twitter recommender distortion of users’ environment reality. Twitter’s algorithmically curated *home feed* amplifies toxic and sentimentally valenced tweets, distorts the political landscape perceived by the users, and favors small and/or usually quiet accounts. We argue the need of independent audits of social media platforms with access to large-scale data.

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 by recommender systems. If these recommendation systems are biased, the access to information of millions of citizens could be biased as well, which could lead to systemic risks for society. The study of Huszár et al [9] 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” [2, 4], creating artificial users and scrapping the platform content. While providing interesting insights into the distortion caused by rec-

ommenders, such audits are limited by the number of fake accounts [4] that researchers can create and their ability to realistically mimic human digital behavior [2]. Enlisting volunteers to provide their data seems to gradually become a promising avenue in digital services independent external audits [8, 10]. Yet, the relative lack of control is a common drawback of a purely crowd-sourced audits.

To circumvent these limitations, we combined a lightweight desktop browser extension, capturing the content of volunteers’ Twitter *feed*, and a large-scale data collection through Twitter API to reconstruct the set of messages the participants could have been exposed to. Rather than taking the perspective of the sender, as [9] have done, we have taken the perspective of the receiver to investigate to what extent algorithmic *feed* curation distort the environment perceived by the user, compared to what they have subscribed to. We focus on three biases and show that Twitter’s recommender system 1) highly amplifies small accounts and/or those scarcely publishing content 2) distort the political landscape perceived by the users—reinforcing users’ beliefs for some, confronting them to radically different views for others— and 3) amplifies *tweets* being toxic and/or being sentimentally valenced at the expense of neutral ones. This approach could be generalized to test a myriad of other biases in a fully-controlled way.

Results

Overall, our participants (N=267) had 650 ^[550,761] friends (95% confidence interval determined via bootstraps), each publishing on average 8.3 ^[0,35.2] *tweets* per day. We investigate if the few tweets selected by Twitter for impression faithfully depicts what has been published by participants’ friends.

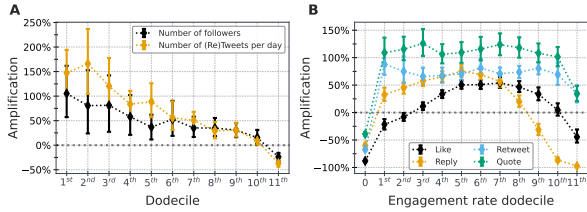


Figure 1. Algorithmic amplification of accounts depending of their number of followers and their daily average of published tweets, binned into 12 quantiles (A). Amplification of *tweets* depending of their of engagement rate (B), in addition to the 11 dodeciles, we report the amplification for tweets having 0 engagement. Error-bars correspond to 95% bootstrap confidence interval of the amplification.

Small accounts benefit from higher algorithmic amplification

Accounts having less than 354 followers (first dodecile) have their *tweets* appearing more than twice as often as if the recommender would provide a faithful representation of users’ friends activity (+105.3 [57.2,161.8] %). Conversely, tweets’ reach of accounts with a number of followers larger than 150k is lessened by -23.8 [-32.0,-15.7] %. *Tweets* of accounts having published in average less than three *tweets* per week since its creation are significantly amplified when they do publish, with an amplification of +147.3 [103.2,194.1] %. The *tweets* of highly active accounts, more than 2 *tweets* per day in average, are shown -36.8 [-43.7,-29.6] % less than if Twitter recommender was faithful.

Algorithmic curation distort the political landscape

Twitter recommender distorts the political landscape by amplifying the *tweets*’ reach unevenly depending on the political leaning of its authors and of the user to whom the messages could be shown to. After having determined the political orientation of participants’ friends (see SI.Methods), and having segmented our participants by political orientation (self-declaration via a form, crossed with their Twitter friends opinion), we notice that for far-left-leaning (N=20) and left-leaning (N=66) participants, the distortion favors the opinion of the participants, see figures 1.D, 1.E. As display on figure 1.D, far-left participants are exposed to +64.6 [6.6,155.5] % more messages published by far-left accounts than what they would have if their *feed* curation was neutral. The amplification decreases as the opinion difference increases, until it

reaches -67.7 [-81.7,-51.2] % for center-right accounts’ *tweets*. Interestingly, for center-right participants (N=48), see figure 1.F, the opposite effect is noticed, the exposition to ideologically aligned accounts *tweets* is lessened by -8.4 [-28.2,12.9] %, while far-left and further right *tweets* are highly amplified, respectively by +122.3 [72.8,212.25] % and +80.7 [28.4,126.0] %. The statistical significance of the differences observed according to participants’ political leaning has been tested through permutation tests. We do not have enough far-right participants at this point to derive meaningful statistics. Once we will, an extended analysis will be performed.

Algorithmic curation amplify toxic and sentimentally valence tweets

The proportion of toxic *tweets* —such as insults, threats or obscenity— published by participants’ friends is around 2.2%, but this small fraction of *tweets* is amplified by Twitter recommender. Participants (N=110) were exposed to +48.7 [37.6,60.8] % more toxic *tweets* than what they would have if Twitter recommender was unbiased. We notice a large inter-participant variability in the amplification, some being exposed to more than twice the proportion of toxic tweets. Platform-wide, toxic *tweets* get more than twice as many *replies* and *likes* per impression than non toxic ones, and only 10-20% less *retweets* and *quote* per impression than non toxic tweets. In a lesser extent, sentimentally valenced *tweets*, either labelled as positive or negative by [3], are amplified —respectively, by +2.0 [-0.3,4.4] % and +5.8 [3.5,8.1] %— at the expense of neutral ones, -8.7 [-11.5,-5.9] %. The amplification of toxic *tweets* has increased since the change of Twitter *feed* structure on January 13, 2023; during the period: December 9, 2022 and January 9, 2023 (N=101), toxic *tweets* were amplified “only” by +32.0 [21.7,42.7] % (the stochastic differences has been confirmed through Mann-Whitney U tests).

Algorithmic curation distort perceived tweets popularity

Figure 1.B displays the amplification of *tweets* in function of their platform-wide engagement rate, computed weeks after their publication (such as being stabilized). First, the reach of *tweets* getting no engagements is strongly lessened and is engagement specific: -88.1 [-90.5,-85.6] % for *tweets* with a null like rate but only -39.2 [-42.2,-36.2] % for *tweets* with a null quote rate. While for *quote* and *retweet* the amplification remains roughly around 110% and 75%, for the first 10th dodeciles; before finally decreasing to 34.2 [18.9,49.5] % and 34.1 [20.9,48.6] % at the eleventh; the amplification is drastically more sensible to the number of *likes* and *replies*. For both, the amplification increases

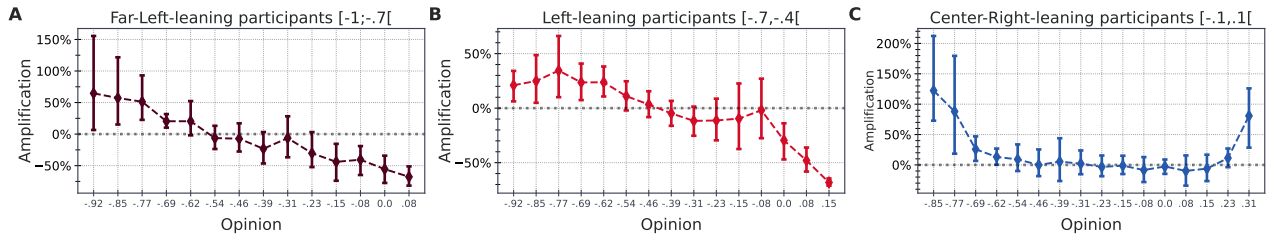


Figure 2. Algorithmic amplification of accounts depending of their political leaning (aggregation windows of 0.15, with successive half overlap), segmenting participants by political orientation, either far-left (D), left (E), or center-right (F). Error-bars correspond to 95% bootstrap confidence interval of the amplification.

with the engagement rate before decreasing after the fifth and seventh dodecile respectively. In the last dodecile, the reach of *tweets* is lessened by -97.4 $[-98.5, -96.1]$ % and -44.5 $[-54.6, -30.8]$ %, corresponding respectively to tweets having a reply/quote rate higher than 1.67% and 2.83%.

Discussion

At the accounts level, Twitter recommender favors *tweets* stemming from small accounts and/or accounts scarcely publishing content. Such a behavior could be a heuristic design to avoid having *feeds* populated by spamming accounts and extremely popular ones; on the one hand giving every users the possibility of being “heard” —or at least the hope to be— but on the other hand, giving an advantage to actors resorting to astroturfing, i.e. amplifying their online presence via fake accounts artificially relaying their ideas. Despite the algorithm’s stated goal of maximizing engagement, it seems to stop recommending tweets once they reach a certain level of popularity. We suggest that this may be because Twitter is designed to promote new content and keep users engaged, rather than to promote ongoing conversations. This tension between promoting popular content and promoting a diverse range of recent content on Twitter, may lead to situations where popular tweets are no longer recommended and may be seen as “shadow-banned”.

Also, we notice that Twitter recommender unevenly broadcast political tweets —distorting the political landscape perceived by users— and shed light on amplification patterns, specific to the different political communities. Keeping in mind the overall objective function of Twitter recommender, one may hypothesise that these patterns are the ones found to be maximizing the engagement on the platform. At the *tweet* level, we can reasonably hypothesise that it is because toxic *tweets* have higher reply and like engagement rate that they are preferably selected by the rec-

ommender, leading to the observed +48% amplification of such toxic content.

Our audit leads us to conclude that Twitter has systemic effects on information ecosystems by making them more toxic and influencing how political groups may perceive each other. Moreover, the amplification of small accounts could make this digital space more manipulable via astroturfing practices. However, recommender systems remain black boxes with a tremendous amount of features and data-points and further studies are necessary to untangle such intricate system. Confounding factors are everywhere, we only captured some of the end product unfairness. While making the design of these recommender systems more transparent may shed light on the internal mechanisms leading to biased suggestions, independent audits with access to large-scale data will remain essential in the regulation of digital services, as specified by the 40th article of the European law on digital services [1].

Materials & Methods

Data Collection

The crowd-sourced extension, named “Horus” has been advertised on social media and newspaper in fall 2022. After being informed of the goal of the study and their informed consent gathered, the browser extensions, named “Horus”, start collecting participants’ Twitter *feed*. As an incentive to install the extension, in addition, to helping scientific research, a personalized report on the political diversity of their Twitter friends and of their curated *feed* was sent to the volunteers after having collected enough data to be reliable. 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. Taking the participants

have been active on —the desktop version of— Twitter as our only objects of study, the previous analysis has been performed on 267 participants, after collecting their Twitter *feed* between December 9, 2022, and February 7, 2023. On January 13, 2023, Twitter removed the ability for users to consume content in a reverse-chronological way, constraining them to be subject to algorithmically curated feeds either solely confined to the accounts they are following or more broadly to accounts and topics they may be “interested in” [11]. We then split our analysis on this change, and perform most of the analysis on the 24 days windows between January 14, 2023 and February 7, 2023; using the period from December 9, 2022 and January 9, 2023 for comparison. Unless stated otherwise, we perform the analysis on the 24 days windows, between January 14, 2023 and February 7, 2023, following the change of Twitter *feed*, removing the ability for users to consume content in a reverse-chronological way, constraining them to be subject to algorithmically curated feeds.

In addition to this crowd-sourced data collection, we requested through Twitter API the number of *tweets* published, during the considered timeframe, by 45k accounts — the most popular ones among the 120k unique accounts followed by our participants— and the set of 3 millions *tweets* published by the 14k accounts followed by at least three participants. These two datasets allow us to compute the algorithmic amplification either based on account or tweet-related features.

Account Political Orientation

The political orientations was estimated leveraging the Polioscope database; embedding through *node2vec* [7] the graph of *retweet* associated to French political tweets, *retweets* being a reliable signal of ideological alignment as shown in [6]. Computing the angular similarity in the latent space between the three main French political figures —Jean-Luc Melenchon (far-left), Emmanuel Macron (center), Marine Le Pen (far-right)— and each of the 1.2 millions Twitter accounts having repeatedly published and/or *retweeted* political content during the 2022 (during which both the French presidential and legislative elections occurred) we assigned a numerical political leaning in $[-1, 1]$. French political arena having the particularity of being circular [5], antisystem activists bridging far-left and far-right militants, we implemented a periodic boundary condition at ± 1 in our numerical opinion estimates. The resulting numerical scale is easily interpretable —negative value for left-leaning account and a positive one for right-leaning ones, supporter of the current French president Emmanuel Macron around zero. The numerical scale matches with both members of parliament political group and a clusters analysis [6]. Accounts displaying comparable angular sim-

ilarity to the three anchors are considered as “nonpartisan” and have been ignored in the present study. We restricted our analysis to the participants’ friends having interacted at least twice with at least 5 different accounts in our database, resulting in a set of 16k consider strongly politically active friends.

Quantification of algorithmic amplification

For a given participant having a set of friends F , we defined the algorithmic amplification of a subset of accounts $G \subseteq F$, exhibiting a given characteristic, as:

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

the ratio between, the fraction of messages published by members of $G \subseteq F$ having been impressed on the participant screen, and a neutral baseline a_F . This neutral baseline is simply the overall fraction of messages published by the participant friends having been impressed on their screen. As in [9], we excluded from our study *retweets*, the attribution of a potential amplification being ambiguous. Our server-side data collection coverage being partial, we systemically perform bootstrapping on both participants’ friends and on participants’ individual estimates; we provided 95% confidence interval on every amplification measures. The algorithmic amplification of a tweet-related features is define analogously, substituting $G \subseteq F$ by the set of tweets, published by accounts in F , exhibiting a given characteristic e.g. having a high engagement rate. When determining the algorithmic amplification at the participant level, we also compute the amplification of a random subset of friends having the cardinal of the subset of friend having the considered characteristic. We systemically performed Mann–Whitney U tests to confirm that a potential recommender unfairness was differing from a simple selection bias.

The *tweets* displayed on users *feed* are extremely recent —three fourth of the displayed content is less than 12 hours old— and while the collected data could allow us to perform a fine temporal analysis, we decided to compute the algorithmic amplification after aggregating participants sessions over a couple of weeks. Since, ultimately, the preference for recent content is an arbitrary heuristic that should not be ignore when determining if the algorithmic curation distort the content production landscape. Among the myriad of algorithmic biases we could test, we focus in this brief report, on answering long standing question while providing a demonstration of the data richness of impressions at the users level.

Data Availability

In strict compliance with both Twitter developer policy and Horus’s privacy policy we provide, upon request from the corresponding author, aggregated data, to allow reproduction of our findings.

Acknowledgment

This work was supported via a doctoral fellowship from the “Fondation CFM pour la Recherche”, as well as the support of the Complex Systems Institute of Paris Île-de-France and the Region Île-de-France.

References

- [1] 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.
- [2] J. Bandy and N. Diakopoulos. More accounts, fewer links. *Proc. ACM Hum.-Comput. Interact. on Human-Computer Interaction*, 5(CSCW1):1–28, apr 2021.
- [3] F. Barbieri, L. Espinosa Anke, and J. Camacho-Collados. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France, June 2022. European Language Resources Association.
- [4] N. Bartley, A. Abeliuk, E. Ferrara, and K. Lerman. Auditing algorithmic bias on twitter. In *13th ACM Web Science Conference 2021*. ACM, jun 2021.
- [5] D. Chavalarias, P. Bouchaud, and M. Panahi. Can few lines of code change society? beyond fact-checking and moderation: how recommender systems toxifies social networking sites. *submitted to PNAS*, march 2023.
- [6] N. Gaumont, M. Panahi, and D. Chavalarias. Reconstruction of the socio-semantic dynamics of political activist twitter networks—method and application to the 2017 french presidential election. *PLoS ONE ONE*, 13(9):e0201879, sep 2018.
- [7] A. Grover and J. Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, aug 2016.
- [8] E. Hargreaves, C. Agosti, D. Menasche, G. Neglia, A. Reiffers-Masson, and E. Altman. 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, aug 2018.
- [9] F. Huszár, S. I. Ktena, C. O’Brien, L. Belli, A. Schlaikjer, and M. Hardt. Algorithmic amplification of politics on twitter. *Proc. Natl. Acad. Sci. U.S.A.*, 119(1), dec 2021.
- [10] L. Sanna, S. Romano, G. Corona, and C. Agosti. YTTREX: Crowdsourced analysis of YouTube’s recommender system during COVID-19 pandemic. In *Information Management and Big Data*, pages 107–121. Springer International Publishing, 2021.
- [11] Twitter. About your for you timeline on twitter, Jan 2023.