Can a single line of code change society? The systemic risks for global information flow, opinion dynamics and social structures of recommender systems optimizing engagement

David Chavalarias, Paul Bouchaud, Maziyar Panahi

CAN FEW LINES OF CODE CHANGE SOCIETY ? Beyond fack-checking and moderation : how recommender systems toxifies social networking sites

David Chavalarias[1,a,b,*], Paul Bouchaud[a,b,*] et Maziyar Panahi[a]

Abstract As the last few years have seen an increase in online hostility and polarization both, we need to move beyond the fack-checking reflex or the praise for better moderation on social networking sites (SNS) and investigate their impact on social structures and social cohesion. In particular, the role of recommender systems deployed at large scale by digital platforms such as Facebook or Twitter has been overlooked. This paper draws on the literature on cognitive science, digital media, and opinion dynamics to propose a faithful replica of the entanglement between recommender systems, opinion dynamics and users' cognitive biais on SNSs like Twitter that is calibrated over a large scale longitudinal database of tweets from political activists. This model makes it possible to compare the consequences of various recommendation algorithms on the social fabric and to quantify their interaction with some major cognitive bias. In particular, we demonstrate that the recommender systems that seek to solely maximize users' engagement *necessarily* lead to an overexposure of users to negative content (up to 300% for some of them), a phenomenon called algorithmic negativity bias, to a polarization of the opinion landscape, and to a concentration of social power in the hands of the most toxic users. The latter are more than twice as numerous in the top 1% of the most influential users than in the overall population. Overall, our findings highlight the urgency to identify harmful implementations of recommender systems to individuals and society in order better regulate their deployment on systemic SNSs.

Significance statement January 6, 2021 was a shock to democracies. Everything suggests that it was not a fad and social networks played their role. However, to counter the relentless worldwide polarization of public opinion, we need to go beyond "fack-checking" and the moderation of harmful content. This paper studies the role of self-learning recommendation systems on systemic platforms such as Facebook or Twitter and their interaction with users' cognitive bias. We show that their most likely current implementation *necessarily* leads to harmful consequences for individuals and society. Unless BigTech companies prove otherwise, this is not a user behavior problem but a technology problem. This implies that systemic digital platforms currently pose systemic risks to social cohesion. Keys to evidence-based regulation are provided.

[a] CNRS, Complex Systems Institute of Paris Île-de-France (ISC-PIF), 113 rue Nationale, 75013 Paris, France

[b] EHESS, Centre d'Analyse et de Mathématique Sociales (CAMS), 75006 Paris, France

[*] These authors have equally contributed to the study

[1] Corresponding author : David Chavalarias

1 Rationale

In January 2018, Facebook announced a change in its *news feed*, a recommender systems which is the main information source of its 2.2 billion users. The aim is to favor content that generates the most engagement : shares, comments, likes, etc. Unfortunately for the public debate, research in psychology shows that this content is, on average, the most negative, a phenomenon called *negativity bias* (Rozin & Royzman, 2001). The effects of this change are not long in coming. According to leaked internal Facebook documents (Hagey & Horwitz, 2021; Zubrow, 2021), exchanges between users have since then become more confrontational and misinformation more widespread. Meanwhile, the political polarization of the users increased due to the platform (Allcott *et al.*, 2020). These changes were so radical and profound that both journalists and political parties felt forced to "skew negative in their communications on Facebook, with the downstream effect of leading them into more extreme policy positions".

This increase in polarization and hostility in on-line discussions has been observed on other platforms. On Twitter for example, where user's home timeline is by default governed by a recommender system since 2016, the proportion of negative tweets among French political messages raised from 31% in 2012 to more than 50% in 2022 (Mestre, 2022). It has also been demonstrated (Vosoughi *et al.*, 2018) that falsehood diffuses "significantly farther, faster, deeper, and more broadly than the trut" on this platform while having the strongest *echo chamber effect* (Gaumont *et al.*, 2018), and consequently the stronger polarization effect.

Can changing few lines of code of a global recommender system qualitatively change human relationships and society as a whole ? To what extent social media recommender systems are changing the structure of online public debates and social group formation processes on a global scale ? These are fundamental questions for the sanity of our democracies at a time when polarization in on-line environments is known to spill-over off-line (Doherty *et al.*, 2016). Moreover, at a time where countries like the European Union start to regulate the sector of digital services [1] a scientific answer to these questions is also required to implement evidence-based policies.

Previous studies have explored the societal impact of on-line social networking sites (SNSs) such as the impact of recommender systems on on-line social groups formation (Ramaciotti Morales & Cointet, 2021; Santos *et al.*, 2021), on-line social networks polarization (Tokita *et al.*, 2021) or the impact of networks topologies on opinions dynamics (Baumann *et al.*, 2020). But the impact of recommender systems on the coupling between opinion dynamics and social network formation is hardly addressed in literature.

This paper fills this gap and provides a methodological framework that takes into account the entanglement between personalized recommender systems, human cognitive bias, opinion dynamics and social networks evolution. It makes it possible to explore the consequences of various design of recommender systems on the social fabric and to quantify their interaction with some major cognitive bias.

As a case study, we apply this framework to a Twitter-like social network model. We build a state-of-the-art opinion dynamics model and perform an empirical calibration and empirical validation of different components of this framework on a 500M political tweets database, published between 2016 and 2022. Next, we illustrate the impact of

1. Digital Service Act (DSA) https ://ec.europa.eu/commission/presscorner/detail/en/ip_22_2545

recommender systems on society by comparing four differently designed recommender systems based on behavioral, opinion, and network models calibrated on our Twitter data. Perspectives are given to extend this approach to other types of social networks.

This case study highlights the role of human cognitive biases and of the characteristics of new digital environments in the self-reinforcement processes that fragment opinion spaces and distort to a large extent Internet users' perception of reality.

In particular, we demonstrate that, as soon as users have a slight negativity bias, recommender systems that seek to solely maximize users' engagement lead to an overexposure to negativity, a phenomenon called *algorithmic negativity bias* (Chavalarias, 2022) and to a stronger ideological fragmentation of on-line landscapes compared to situations where information circulates in a neutral way.

Framework Description

Let's model the generic properties of online social networks in order to study the stylized phenomena associated with some of their key features. The detailed characteristics of SNSs varies from one platform to another but they all have some core features in common :

1. **[Publication]** At anytime t, a user i can publish a message m_i^t,
2. **[Networking]** Each user j can subscribe to i's information diffusion network $\mathcal{N}_i^d(t)$ (on some social networking sites subscriptions are open, on others they should be agreed by i).
3. **[Information]** Each user i can read the messages produced by the set $\mathcal{N}_i^r(t)$ of accounts they have subscribed to and eventually share them with their own subscribers $\mathcal{N}_i^d(t)$.

Subscription networks between users of a social networking site can be represented by an evolving directed network $\mathcal{N}(t) = \cup_i\{\mathcal{N}_i^d(t) \cup \mathcal{N}_i^r(t)\} = \{s_{ij}\}_{i,j}$ in which an edge s_{ij} exists when the user j has suscribed to i's account (information flows from i to j). $\mathcal{N}$ is the backbone of information circulation on such platforms. Its evolution is generally influenced by a *social recommender system* that suggests new "friends" to users.

The average number of subscriptions per user being quite high (*e.g.* > 300 on Facebook, > 700 on Twitter), most social networking sites implement a *content recommender systems* that helps any user i to find the "most relevant" messages among those produced by their social neighborhood $\mathcal{N}_i^r(t)$. On platforms such as Facebook, Twitter, YouTube, LinkedIn, Instagram or TikTok, these content recommender systems take the form of a personalized news feed $\mathcal{F}_i$ that aggregates "relevant" messages in a stack. They constitute the main source of information for the users of these platforms (on Youtube the recommender is responsible for 70% of watch time for exemple [2]). Social recommendation and content systems shape users' opinions through the constraints they place on the global flow of information as well as on the processes of social ties formation. Although most of them are black boxes, we know that these recommender systems learn from the actions of their users according to some very generic objective function.

This being said, in order to model the interaction between human cognition, recommender systems, user's opinions and social network evolution, we have to model three things :

2. https ://www.cnet.com/tech/services-and-software/youtube-ces-2018-neal-mohan

— **[Recommender system]** The process of message sorting of news feed by the content recommender system,
— **[User's attention, cognitive bias and opinion dynamics]** The user's motivations to read and share a message, their potential cognitive bias, their opinion and its evolution after exposure to a message,
— **[Network evolution]** The way users decide to subscribe and unsubscribe to other accounts and the role of the social recommender system in this process.

We will include all these elements in a model in discrete time where each time step will correspond to one day of interaction between users. Each of these elements is the object of a research field in its own right, so that it is not a question here of proposing advances on each of these dimensions. We will rather consider the state-of-the-art models for each of them in order to calibrate them on empirical data and study their interactions.

The content recommender system

A content recommender systems has access to a set of users' characteristics, as for example the number of subscribers per user, the list of the accounts they have subscribed to, the number of shares per message, etc., and a set of messages' features, as for example their number of shares or their sentiment, and produces at each time step t and for each user i and ordered list of all messages produced by accounts from $\mathcal{N}_i^r(t)$ to be displayed for reading.

The Users

Users are described as entities with an internal state (their "opinion"), some interface with their environments (*e.g.* read some messages) and a repertoire of actions on the environment (publish a message, share a message, subscribe to a user's account, etc.). For simplicity, we assume that the opinion dynamics is solely driven by the interactions among users. We will call this stylized representation of the users "agent".

Users' opinions

We built on the literature of non Bayesian opinion dynamics modeling (see (Noorazar *et al.*, 2020) for a review) and assign to each agent i at t an opinion o_i^t in a metric space $\mathcal{O}$, and an opinion update function $\mu_i : \mathcal{O}^2 \to \mathcal{O}$ that defines its propensity to change its opinion o_i after reading a message that conveys opinion o_j of agent j. $\mathcal{O}$ and μ will be estimated empirically.

Agents' Rule 1 (Opinions' update) *: after sharing agent j's message at time t, agent i's opinion is updated according to $o_i^{t+1} \leftarrow \mu_i(o_i^t, o_j^t)$.*

Users' online activity

At each time step t, each agent i publishes $n_i^p(t)$ new messages, assumed to perfectly reflect they view, and shares $n_i^s(t)$ read messages authored by other agents. $n_i^p(t)$ and $n_i^s(t)$ will be estimated empirically.

Agents' Rule 2 (reading a message) *: at each time step, agent i will "scroll" in their feed and randomly stop to read carefully some messages.*

Once read, the user may engage with the message :

Agents' Rule 3 (engagement with a message) *: the probability that an agent i shares a message from an agent j (i.e. republish the message identically at the next time step) depends on the difference of opinion $|o_i^t - o_j^t|$.*

In the literature, different functional forms for the probability of engagement have been proposed such that the exact from should be estimated empirically according to the kind of opinion space that is modeled.

Users' cognitive bias

Many cognitive bias are worth to be studied in the perspective of the analysis of recommender systems' impacts. As an illustration, we will focus on two famous bias in psychology : the previously mentioned *confirmation bias* and the *negativity bias* (Epstein, 2018; Knobloch-Westerwick *et al.*, 2017; Rozin & Royzman, 2001) —the propensity to give more importance to negative piece of information. Our goal in this example is to evaluate the strength of the *algorithmic negativity bias* (Chavalarias, 2022) : the large scale over-exposition to negative contents due to the algorithmic machinery.

To quantify the *algorithmic negativity bias* effect, we attribute a valence to messages published by the agents, that can be either "negative" or "positive/neutral". We thus assign to each agent i a proportion ν_i^t of negative messages published at t and a propensity Bn_i to interact in a privileged way with negative messages (*negativity bias*). ν_i^t and Bn_i will be estimated empirically. The negativity bias of our agents is then implemented as a variation of rule 2 :

Agents' Rule 4 (reading a message with valence) *: at each time step, agent i will "scroll" in their feed and randomly stop to read carefully some messages. The probability of stopping and read a negative message is Bn_i times higher than for a non-negative message.*

The above defined set of rules allows us to study the feedback loops between the aforementioned cognitive biases and a learning recommender. On the one hand the recommender seeks to maximize the user engagement, on the other hand, the user is more likely to engage with content aligned with their existing belief and/or of negative nature. As a consequence, we can expect the recommender to be more and more biased as it learns users' bias over time.

Network evolution

Opinions co-evolve with interaction networks in a feedback loop. The homophilic nature of human interactions indicates that users tend to interact and form relationships with people who are similar to them (McPherson *et al.*, 2001), and cut social ties with people who happen to share content that is not aligned with their views. Besides this, SNSs usually suggest new connexions to users via social recommender systems that are most of the time based on structural similarities (*e.g.* mutual friends) (Tokita *et al.*, 2021).

We will take into account these factors in a parsimonious yet realistic model of link formation and pruning. The network specifications at initialization of our simulations (connectivity, types of agents, etc.) will be determined empirically.

Links suppression (Rewiring rule 1)

Agents score their subscriptions to monitor the interest they have in maintaining them. For every subscription s_{ji} of i to j, the disagreement $\delta_{ij}(t) \geq 0$ of i with the content received through s_{ji} is initialized at 0 and updated at each time step according to $\delta_{ij}(t+1) = \gamma \times \left(\delta_{ij}(t) + n_{ij}^t |o_i^t - o_j^t|\right)$, with $\gamma < 1$ being a daily discount factor and n_{ij}^t the number of messages read by i during time step t that have been authored or relayed by j. If $s_{ji}(t) = 1$ and the disagreement $\delta_{ij}(t) > \tau$, i will unsubscribe from j, *i.e.* $s_{ji}(t+1) = 0$.

$1/\gamma$ is a characteristic time of agents' evolution that is difficult to estimate empirically. It will be set arbitrary to a reasonable value. So will be the τ which determination would depend on the knowledge of γ. We have verified that our results do not depend on the precise knowledge of these two parameters.

Links formation (Rewiring rule 2)

To maintain the connectivity measured empirically, we assume that when an agent breaks an edge with an unaligned user, it starts following a randomly chosen second neighbors (a rewiring mechanism often observed in SNSs (Tokita *et al.*, 2021)).

Instantiation of a recommender systems : the example of Twitter

In order to understand the complex relation between the specific choice of a recommender systems and its systemic effects on opinion dynamics and social networks evolution, we apply thereafter the above described framework to the modeling of political opinion dynamics on Twitter. Passing, we find realistic parameter values that could be used to model the impact of other SNSs recommender systems.

At the time of the study, Twitter's data availability, its widespread use —more than 300 millions of monthly active users worldwide— and its predominant role in political communication justify our choice to use it as our experimental field for testing the proposed framework. Moreover, as measured empirically, Twitter is also a digital media where negative contents are more viral than others (see Fig. S18) and where the users themselves are biased towards the production of negative contents (see see Fig. S11) This raises the important question, both for public debate and for the well-being of users, of the extent to which this overflow of negativity is due to Twitter's algorithmic architecture.

Briefly, Twitter is an online social network launched in 2006 allowing its users to exchange publicly 280 characters-long messages that are broadcasted to theirs "followers", users who subscribed the author's account. Content is displayed to the users on a feed called *Home timeline* according to personalized recommendations. The messages are ranked by a machine learning algorithm predicting the likelihood the user will engage with the tweet. In the following, we will focus on the two main forms of engagement on Twitter (Twitter, 2020) : (1) the careful read of a tweet –which often requires a click to expand

the content– (2) the retweet, *i.e.* the fact of republishing the message identically with the mention of its author, without any comment nor modification.

Despite that social influence extents well beyond retweet, empirical studies observed that retweets are more relevant to characterize people's opinion and monitor its evolution, at least in a political context, than, for example, Twitter mentions (Conover *et al.*, 2011; Garimella *et al.*, 2018). It is indeed possible to predict with high accuracy the political orientations of political activists from their retweet data only (Gaumont *et al.*, 2018). In what follows, the empirical applications of our framework will focus on retweets networks.

1.1 Choice of a recommender system

Several leaks as well as official announcements suggest that many social networking sites use the users' engagement maximization as the objective function for their recommender systems. We will thereafter analyze the consequences of such objective functions on the social fabric.

Recommender's Rule 1 *: at each time step, the recommender will rank and display for each agent i a subset of messages from $\mathcal{N}_i^r(t-1)$ according to their probability of being shared, as predicted by the recommender.*

Due its flexibility and efficiency, we implemented this optimization through XGBoost algorithm (Chen & Guestrin, 2016).

Recommender systems fulfill their objectives by relying on certain inputs. Modeling such algorithm should thus define some type of data it has access to. The variety of input data used by commercial recommender systems is part of the domain of business secrecy such that little is known about which input data are really used. We will here select two broad categories of data that are likely used by commercial recommenders (cf. (Xu & Yang, 2012), (Huszár *et al.*, 2022)) :

— *Sentiment analysis* : the negative or neutral nature of a tweets, as well as the proportion of negative content retweeted by the user in the past.

— *Popularity assessment* : the popularity of the tweet's author i.e. average number of retweets to its messages, the number of time the message has been retweeted and the frequency at which the user retweets the author.

In order to investigate the consequences of the different input features, we will compare three different implementations of the recommender :

— *Neg* : use only input data from sentiment analysis,

— *Pop* : use only input data from popularity scores,

— *PopNeg* : use the combined features of the *Neg* and *Pop* algorithms.

To assess the effect of these three implementations of recommender systems on the social fabric, we will compare them to a neutral recommender systems, the reverse-chronological presentation of content, thereafter call *Chrono*. *Chrono* is often referred as non-algorithmic recommendation due to its simplicity. It was briefly implemented by Twitter until the takeover by Elon Musk, from which it is no longer possible to disable the recommendation algorithm.

Empirical Calibration

In this section, we fully calibrate our model using empirical data regarding French politics, collected on Twitter in autumn 2021 within the *Politoscope* project (Gaumont *et al.*, 2018), a social macroscope for collective dynamics on Twitter. The *Politoscope* continuously collects since 2016 political tweets about French politics and makes it possible to select subsets of the most active users over any given period.

Network of users' interactions

While accessing Twitter's graph of followees-followers is possible through Twitter API, such a graph would be misleading if used in our model. Indeed, the content recommender effectively used on Twitter is already well trained, content from someone followed may never be shown to the user, distorting our simulation. To circumvent this limitation, we instead consider the empirical network $\tilde{\mathcal{N}}$ of retweets and quotes combined. Such a network seems indeed to be a reasonable proxy to what is actually shown to the user by the platform. Considering quotes, and not only retweets, allows to include ideologically unaligned content as discussed below. Each of our simulations was initialized over the empirical network $\tilde{\mathcal{N}}$ of interactions over the selected period.

Calibration of opinion space

We will henceforth understand the term "opinion" as an ideological positioning within the political arena, excluding de-facto political agnostics. Not all candidates having the same digital communication strategy, we will include in what follows only leaders having a significant presence on Twitter during the considered period.

The reconstruction of opinion spaces from SNSs data has been a very active field of research these last several years, with reconstructions in one (Barberá, 2015; Briatte & Gallic, 2015), two dimensional spaces (Chomel *et al.*, 2022; Gaumont *et al.*, 2018) or even in spaces with variable dimensions (Reyero *et al.*, 2021). As for retweet networks, retweeting someone on a recurring basis has been demonstrated to be an indicator to some ideological alignment (Conover *et al.*, 2011; Garimella *et al.*, 2018; Gaumont *et al.*, 2018).

With a clustering analysis of political retweet graphs, Gaumont et al. (Gaumont *et al.*, 2018) achieved 95% accuracy over opinion's classification, validating the use of the retweet graph for such a study [3]. The spatialization of the Politoscope retweet graph of autumn 2021 depicts a multi-polar circular political arena (cf. Fig. 1) where the relative positions of the political leaders are in adequacy with the publicly depicted political scene. As discussed in SI, we used this spatialization to model the opinion space $\mathcal{O}$ as a circular one dimensional metric space with $o_i \in]-1, +1]$, making it possible to initialize the opinion of our agents in $\tilde{\mathcal{N}}$ with their empirical estimation, compute the impact of the recommender's suggestion on user's opinion, and determine the global impact of different recommender systems on the distribution of the users' opinions in $\mathcal{O}$.

3. We made the same verification on our own dataset and found similar performances.

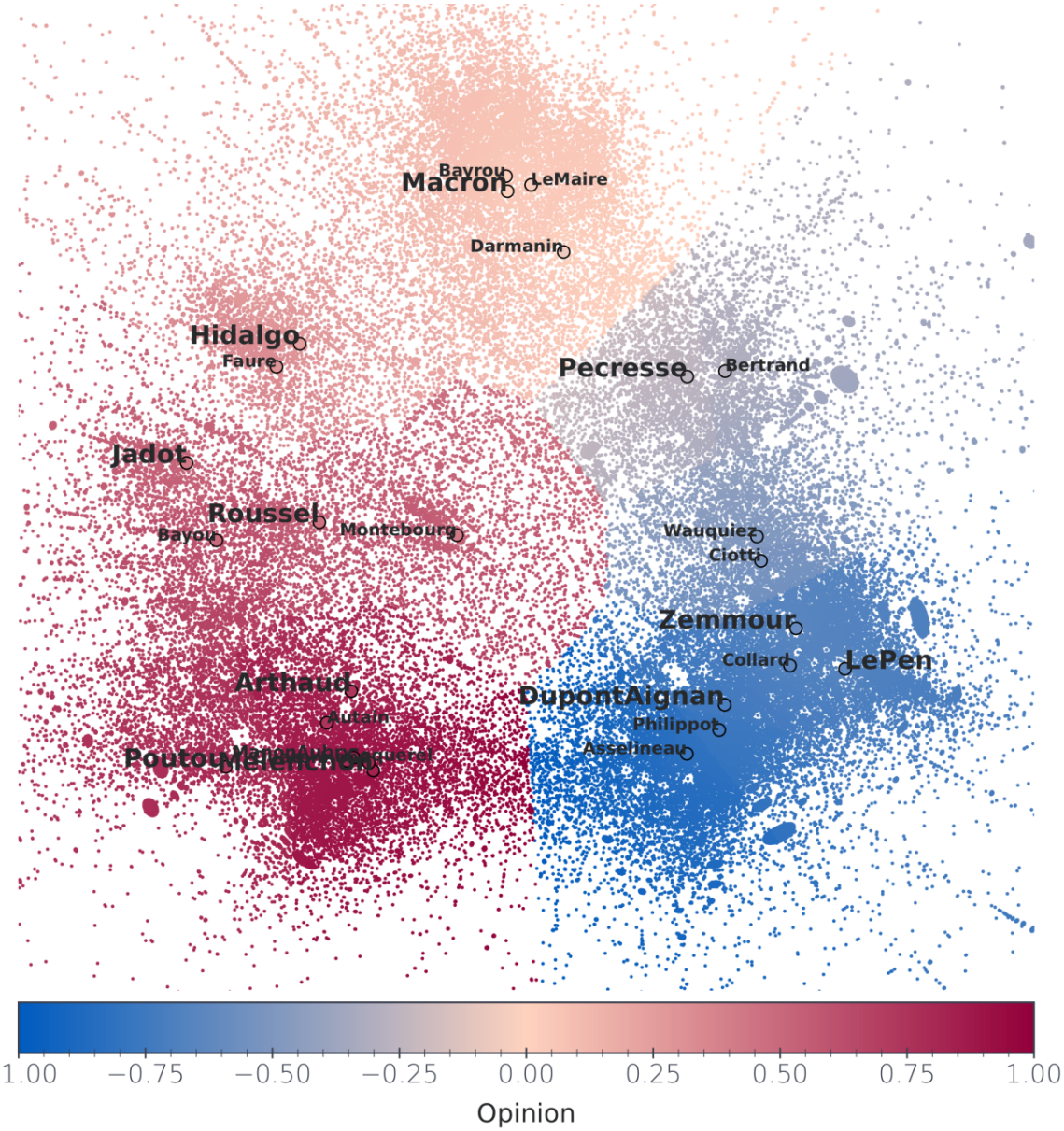

FIGURE 1 – Multi-polar graph of the French pre-electoral political Twittersphere calculated during September 2021. Each node corresponds to a user, colored according to the opinion assigned by the described method. Political leaders are highlighted, in particular the candidates for the 2022 French presidential election.

Calibration of agents' opinion update

Having a metric space for the opinion space, we can build on the sizable literature on opinion dynamics (Deffuant *et al.*, 2000; Jager & Amblard, 2005; Noorazar *et al.*, 2020). Thanks to the full history of user's interaction from our Twitter dataset, and assuming for the sake of simplicity that the functional form of μ, the opinion update function, is the same for all agents, we determined the most likely μ using symbolic regression. The regression was performed using genetic algorithms (Fortin *et al.*, 2012) over the set of arithmic and trigonometric functions as well as an implemention of the difference in the

periodical opinion space.

This empirical calibration allowed us to identify a linear function of opinion updating $o_i^{t+1} \leftarrow o_i^t + \lambda_i(o_j^t - o_j^t)$ with $\lambda_i \in \mathbb{R}$. Note that this function was already a widely used in the opinion dynamics literature (Deffuant *et al.*, 2000; Jager & Amblard, 2005).

Because of our lack of information on tweets' impression and given our opinion attribution method, we have decided to simplify the model by assuming that agents change their their opinion, *i.e.* ideological positioning, only when they retweet a message.

Then, we fitted for each agent the opinion update parameter λ_i, which the absolute value reflects the influenceability of the agent, *i.e.* to what extent will they change opinion when retweeting someone else, using the list of daily messages effectively retweeted by the user (see SI **??**).

Such a fitting leads to a relatively high accuracy, with more than 75% of our final fitted opinions off by less than 0.05 after 30 iterations (corresponding to end of October, cf. Fig. **??**). This is less than intra-communities opinion diversity. We should emphasize that the goal of the present work is not to accurately predict the opinion of online social media users, but only to provide a faithful simulation of online users' behavior to study the consequences of algorithmic recommendation. In particular, users' opinion are used within the simulation to determine the probability of retweeting a content, thus being off by 0.05 in opinion does not alter the behavior of the simulation. The only significant changes of opinion are the larger ones ($\Delta_{op} > 0.05$), for which the fitted updates rules leads to a relative error less than 25% for more than 60% of the prediction, and even more accurate for particularly large displacements $\Delta_{op} \in [0.5, 1]$ (cf. Fig. **??**). To confirm the sanity of the used method, we considered other time periods, other graph spatialization settings and forecast the opinions one month (November) after the fitting, obtaining similar accuracy, as discussed in supplementary text.

Calibration of agents' activities

In absence of information specifying which messages are displayed on users' screens, we hypothesize that users read messages until they reach their daily number of retweets or when they read all the messages from $\mathcal{N}_i^r(t-1)$. We identified the $110k$ most active users over the period of autumn 2021, get their political tweets and estimated their publication behaviors. The number of daily posted tweets ($\tilde{n_i^p}(t)$, original publications) and retweets ($\tilde{n_i^s}(t)$, shared publications) were exponentially distributed at the individual level (as already observed in (Baumann *et al.*, 2020; Perra *et al.*, 2012)). At the population level, the empirical exponential scales $\tilde{\theta_i^p}$ and $\tilde{\theta_i^s}$ for the different users were distributed according the distribution displayed on Fig. S1. We build on these empirical observations to set the number of tweets and retweets of agent i in $\tilde{\mathcal{N}}$ as independently drawn from two exponential distributions of empirically determined rates $\tilde{\theta_i^p}$ and $\tilde{\theta_i^s}$ respectively.

Latitude of acceptance

Once the opinions assigned, we determined the distribution of difference of opinion Δ_{op} between a user and the authors of retweeted messages. In order to cancel the bias in the representation made by the platform (Huszár *et al.*, 2022), as well as taking into

account the different sizes and positions of the communities, we had to renormalize the distribution of difference of opinion as observed from the retweets with the patterns of publication on quoted tweets (cf. S1.4).

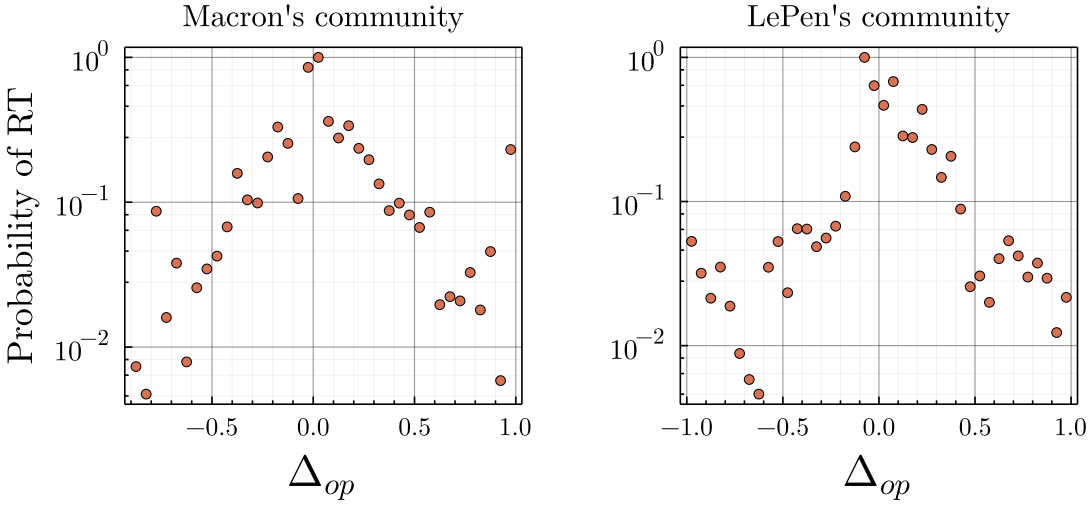

FIGURE 2 – Estimated probability for a user, in the ideological neighborhood of Emmanuel Macron or Marine Le Pen, to retweet a read message according to the different of opinion with its author, $\Delta_{op} = o_{reader} - o_{sender}$, considering periodic boundary conditions. We renormalize such that a perfectly aligned message is retweeted with certainty.

One notices on Fig. 2, that the probability of retweeting a message decays roughly exponentially as the difference of opinion increases, with some refinement revealing political strategies. The asymmetry of the distributions illustrates how Emmanuel Macron community (center) tends to retweet content even further right than further left, while the opposite effect is noticed for left-wing leader Jean-Luc Melenchon community (cf. SI S3) After having determined such distributions for the whole range of opinions, we assigned to our simulated agent the distributions associated to their initial opinions.

Negativity

To calibrate negativity-related properties of our model, we performed a sentiment analysis on $190k$ French political tweets exchanged by 500 unique users during October 2021. This analysis has been performed using the French version of the Bi-directional Encoders for Transformers, CamemBERT (Martin *et al.*, 2020), fine-tuned on French Tweets. We then assigned to our agents an intrinsic negativity ν_i, the proportion of negative content published, drawn from the empirical distribution in function of their initial opinion, as well as a negativity bias, as discussed in supplementary materials.

Evaluation of recommenders' effects

In order to characterize the behavior of our agent-based model we hereby introduced metrics of particular interest :

Algorithmic negativity bias Γ : this is the negativity over-exposure generated by the recommender system defined as the ratio between the negativity in the perceived environment —the content of the timeline— and the negativity in the "real environment" *i.e.* in one's in-neighborhood $\mathcal{N}_i^r$.

To further explore the model, we perform a community detection on the resulting retweets graph using Leiden algorithm (Traag *et al.*, 2019), an improvement guaranteeing connected communities over the usual Louvain method. Once performed we examine :

Newman's modularity $\mathcal{Q}$ (Leicht & Newman, 2008) : assessing the density of connections within a community.

Diversity within a community σ_X^{intra} : the standard deviation of an observable, such as the opinion, the intrinsic or perceived negativity, within a given cluster, normalized by the standard deviation of the observable within the overall population, averaged over the clusters.

Diversity between communities : σ_X^{inter} : the standard deviation of clusters' observable –such as average opinion, intrinsic or perceived negativity– normalized by the diversity among the whole population, assessing the diversity of the different communities with respect to the overall population.

Results

Assessing the impact of recommenders on negativity and opinion polarization with empirical networks

The model was initialized on real data with *all* parameters but two, which we know have no impact on the results, being empirically calibrated. We have simulated one month of interactions to estimate the activity and opinion evolution of each account in the real dataset, and analyzed the aforepresented metrics. As for the account's activity prediction, our framework being stochastic, none of the four recommender systems where able to predict low intensity interactions ≤ 10 retweets/month), overestimating small weights with respect to the real distribution (see Fig. S1).Nevertheless, for larger weights (> 10 retweets/month), *PopNeg* faithfully matches the empirical distribution, while *Pop* overestimates large weights, as one may expect, and *Chrono* underestimates them.

As displayed on figure 3, the overexposure to negativity Γ is non-existent in chronological mode, as expected, while the three algorithmic recommenders lead users to be overexposed to negative content. The *Neg* recommender, solely based on negativity, leads to the highest overexposure, users are shown on average 26% more negative content that what they would have in the neutral *Chrono* mode.

Within the population, the overexposure to negativity is extremely diverse, as depicted in Fig. 4, with some users experiencing an algorithmic negativity bias corresponding to an overexposure of more than 300%. This happens even to users with a large neighborhood and to users without any negative bias. For user with a small number of friends (less than 10), we notice a small ($r = 0.02$) but significant ($p < 10^{-7}$) correlation between the number of in-neighbors and the negativity overexposure. Indeed, as the number of friends increases, so does the size of the pool of message from which the recommender is picking

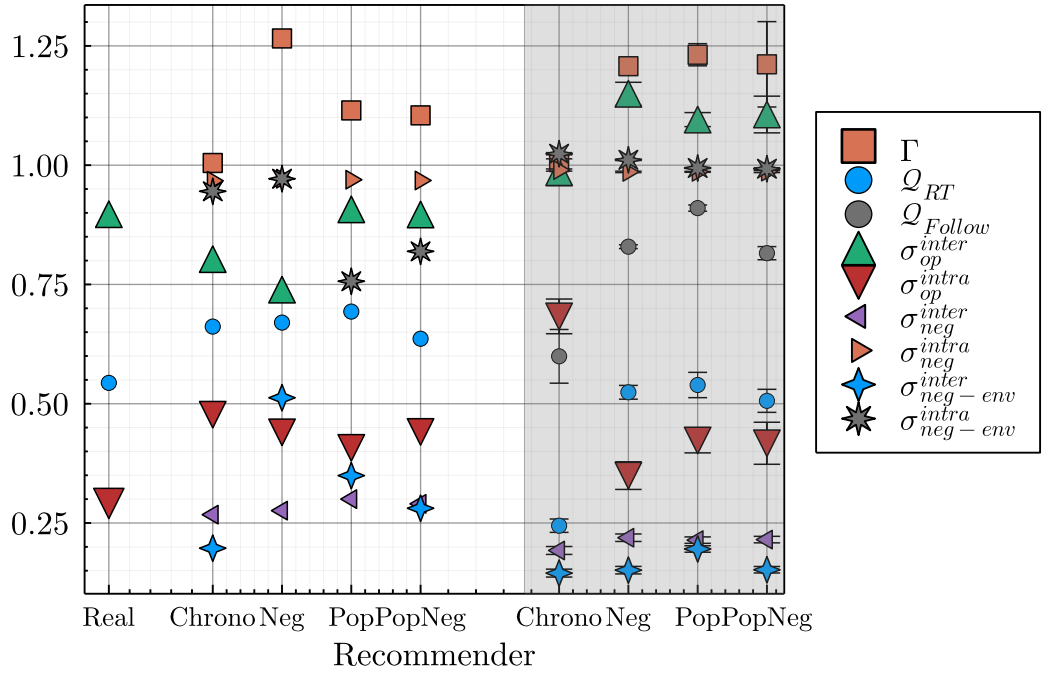

FIGURE 3 – Metrics comparison between the four recommenders and the empirically observed data. Simulation initialized on real graph (white area) and on synthetic graph (grey area), the error bars in the latter case correspond to the standard deviations over 10 repetitions, starting from the same synthetic inputs

from, allowing it to select the most engaging messages (that are most of the time the most negatives ones), leading to a higher negativity overexposure. Such results are a direct consequence of the feedback loop between human negativity bias and the engagement maximization goal hard-coded within the recommender.

The large variations in the level of negativity overexposure at the individual level are important to note since from an individual perspective, it can plunge users into toxic environments, disconnected from reality, which can potentially have detrimental consequences on their mental health and social relationships, as documented, for example, during the COVID-19 pandemic (Levinsson *et al.*, 2021).

The diversity of opinion depends of the recommender system, pointing to another harmful consequences for online sanity. Indeed, while *Chrono* and *Neg* lead to the same $\sigma_{op}^{inter/intra}$, the two social modes, namely *Pop* & *PopNeg*, result into a higher fragmentation of the social fabric. The average diversity of opinion within clusters, σ_{op}^{intra}, is poorer —but not as poor as empirically observed—, and the different clusters are centered around different opinions —higher σ_{op}^{inter}, close to the empirically observed one.

In contrast, the modularity $\mathcal{Q}_{RT}$ of the retweet graph revealed to be independent of the recommender, as well as the diversity of negativity within and between clusters $\sigma_{neg}^{inter/intra}$, the graph structure being strongly constrained from the initialization

By looking at recommender features importance, we notice that the frequency of past interactions between the user and the author, is by far the most informative feature, another illustration of human confirmation bias, reinforced by the recommender. Similarly, the dif-

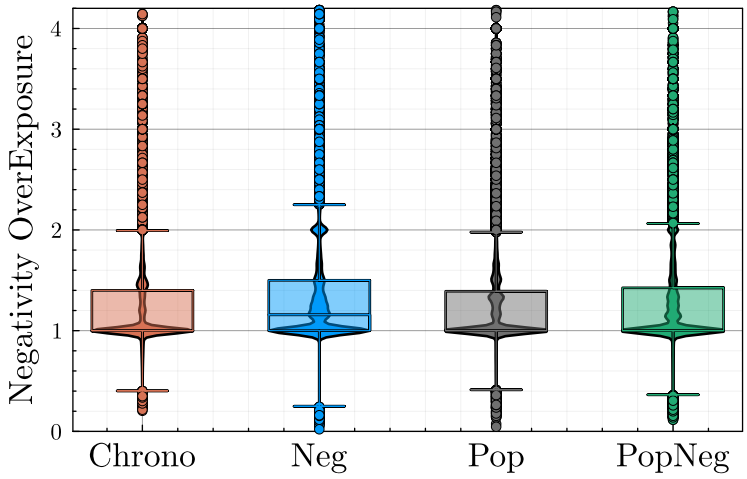

FIGURE 4 – Distribution of the overexposure to negativity within the population for the four recommenders. For clarity, the distribution is truncated at a overexposure of 3.5, the truncated tails represents 3.1%, 2.6%, 2.6% and 3.6% of the total distribution, for *Chrono,Neg,Pop,PopNeg* respectively

ferent clusters are, in these social modes, less diverse in perceived negativity $\sigma^{intra}_{neg-env}$. The unequal perceived negativity, may partially justify the difference of acceptance latitude for the different opinion, but further experiment considering impressions information would be needed to assess the relation between perceived negativity and confirmation bias.

Assessing the impact of recommenders on the formation of social groups with synthetic networks

The previously considered empirical data are the product of years of evolution, shaped by the platform's recommenders. Thus by initializing the network of interactions with a real network, we miss most of the impact of the different recommender systems' on social networks formation. In order to further investigate the consequences of the recommender on the social fabric, we hereby consider randomly initialized networks and analyze their evolution [4]. We drawn for $25k$ agents the properties from the empirical distributions and considered an initial network of follow generated through the Barabási Albert model. Such networks do not aim to realistically mimics all real social networks features but only to provide a zero-th order starting point to illustrate the different consequences of the recommender.

The probability of retweeting a read message is set to decays exponentially with the difference of opinion with a mean of 0.2, to roughly match the empirical one, without specifying it too strongly to French political strategies. The empirical determination of τ being impossible, without having access to what messages is shown to the users on a long time period, we arbitrarily fixed it 0.5 with a time discount factor of 0.9, corresponding to a time-scale of 10 days — by considering alternatives values, the qualitative results discussed below remains.

4. Source code available on (Bouchaud *et al.*, 2023)

A sensitivity analysis of the agents' negativity bias in synthetic networks also demonstrates that the algorithmic negativity bias phenomena ($\Gamma > 1$) appears as soon as agents have some negativity bias ; and its intensity is almost independent of the strength of agents' own negativity biases (see Fig. S16). As long as the users favor negative content over neutral ones, recommender systems based on engagement will lead with certainty to an overexposure to negativity.

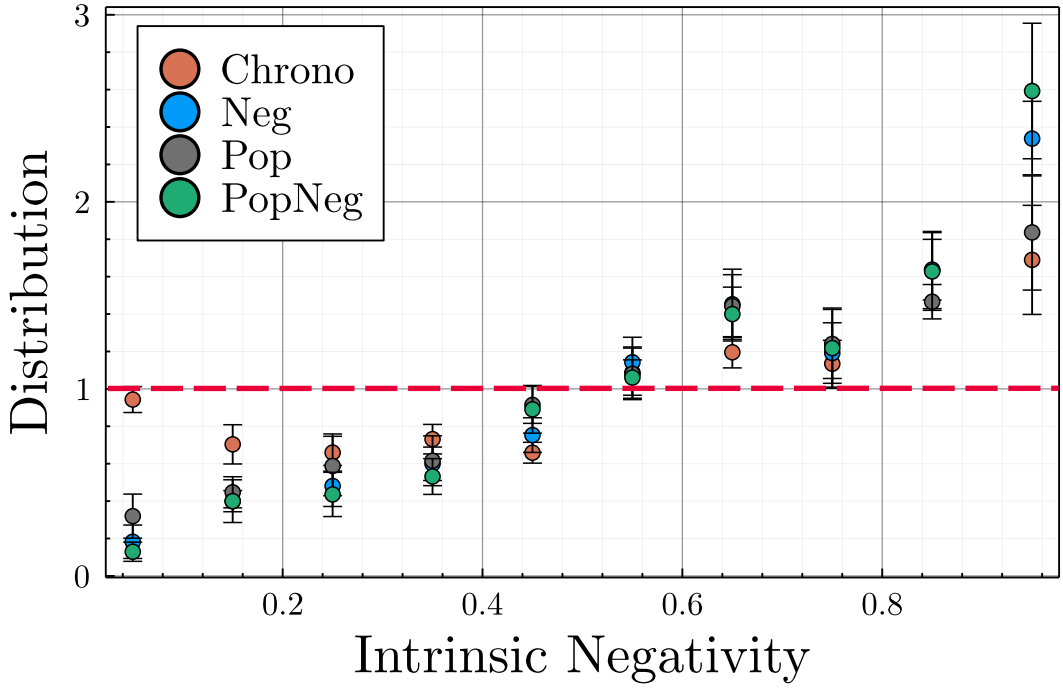

FIGURE 5 – Over-representation of negative agent among the 1% most popular agents compared to the overall population. Analysis performed after two months of simulation, the error bars correspond to the standard deviations over 10 repetitions, starting from the same synthetic inputs.

Starting with an unconstrained random network $\mathcal{N}$ allows the full expression of recommender actions and makes it possible to check that the proposed model for network evolution is compatible with what is observed empirically (see Fig. S12 for an example). As depicted on Fig. 3, after two months of simulated evolution, the modularity of the retweet and follow networks significantly increases with algorithmic recommendation in respect with a neutral presentation of content, in *Chrono*, as well as the ideological fragmentation or the overexposure to negativity.

The algorithmic negativity bias does not only impacts the information environment of the agents toward more toxic environements, is also impacts the structure of *social power* in the population, defined as the capacity of an agent to influence the public debate (Jia *et al.*, 2015). Fig. S13 displays the intrinsic negativity unbalance between the overall population and the top 1% agents receiving the most retweet by tweet while Fig. S14 shows the proportion of negative agents in function of the most popular quantile for *Neg* algorithm. This analysis clearly demonstrates that the amplification of individual negativity bias by engagement-optimizing recommendation algorithms leads to a concentration of online social power in the hands of the most negative users.

For example, while agents publishing negative content half of the time are faithfully represented among the most popular, the users publishing no negative content are absent from the most popular ones for the three algorithmic recommenders. Frighteningly, agents publishing systematically negative content are more than twice as numerous among the most popular than in the overall population ; the two recommenders considering the negativity of the message, namely *Neg* and *PopNeg*, leading to the highest over-representation.

It is also noteworthy that despite being neutral in its selection, *Chrono* nevertheless leads to a significant unfair representation as a consequence of individual negativity bias. Even in a neutral mode, the users will more likely read negative content and hence retweet it, increasing its author popularity.

Discussion

On January 6, 2021, a crowd convinced that the election was stolen stormed the Capitol in Washington, D.C. Whatever the extent to which this event can be attributed to misinformation about the electoral process, it is clear that it was not a fad : one year after Jan. 6 "52% of Trump voters, as well as 41% of Joe Biden voters, somewhat agree or strongly agree that it is time to cut the country in half" (Politics, 2021) while a late 2020 survey concluded that "Americans have rarely been as polarized as they are today" (Dimock & Wike, 2020). In order to remedy this situation of extreme polarization of public opinion, which tends to be reproduced in other countries such as Brazil, the United Kingdom or Italy, we must go beyond the reflex of "fack-checking" and the praise of better moderation of harmful content in online social networks.

As pointed out by other studies using complementary approaches to ours (Ceylan *et al.*, 2023; Törnberg, 2022), we must acknowledge the impact of SNSs on social structures and in particular in the amplification of polarization and hostility among groups. It is not only a phenomenon that affects the general public, the entire information ecosystem is at risk. After Facebook changed its algorithm in 2018 to favor "meaningful social interactions", "company researchers discovered that publishers and political parties were reorienting their posts toward outrage and sensationalism" and internal memo mentionned that "misinformation, toxicity, and violent content are inordinately prevalent among reshares" (Hagey & Horwitz, 2021).

At a time when states are thinking about regulating large social networking sites (SNSs), it is more necessary than ever to have models to quantify their effects on society. In this article, thanks to the modeling of social networks as complex systems and the calibration of the models using big data, we could give hints about what is really going on under the hood.

Using a large scale longitudinal database of tweets from political activists (Gaumont *et al.*, 2018), we have built and calibrated an agent-based model able to simulate the behavior of users on Twitter, some of their cognitive biases and the evolution of their political opinion under the influence of recommender systems. Among other things, we have empirically estimated parameters common to many models of opinion dynamics that were previously arbitrarily defined –like the widespreadly used opinion update rule Agents Rule 1. We also went beyond commonly adopted assumptions, such as a fixed threshold of ideological disagreement for engaging in an interaction, by considering interaction propabilities and estimating their law.

Thanks to this calibrated model, we could compare the consequences of various recommendation algorithms on the social fabric and to quantify their interaction with some

major cognitive bias. In particular, we demonstrated that the recommender systems that seek to solely maximize users' engagement *necessarily* lead to an overexposure of users to negative content, a phenomenon called *algorithmic negativity bias* (Chavalarias, 2022) and to an ideological fragmentation and polarization of the online opinion landscape.

The important point is that these consequences of the way recommender systems are currently implemented, which are harmful to individuals and society, are not necessarily intentional, they only results from the positive feedback between human flawed cognition and SNSs' economic goals. As most of these platforms have become systemic due to their size, their unregulated pursuit of profitability poses systemic societal risks both to their users and to the sanity of our democracies.

Policy makers are increasingly aware of the threats posed to our democracies by the current implementation of SNSs but lack the keys to regulate this sector. Modeling SNSs and their effect on individuals an social groups with an interdisciplinary approach can give some of those keys.

For example, we have shown that when a self-learning algorithm is used to recommend content, feeding it with measures of user or content popularity leads to an increase in the overall toxicity of the platform for individuals and the collectives they form. It also lead to a concentration of the social power in the hand of the most toxic accounts. This means that regulators should focus on the types of data that feed into recommender systems and potentially outlaw some. These kind of findings could help to identify when and where business secrecy, which is so often brandished when platforms are asked to cooperate with public bodies, must be relaxed in the context of their regulation.

On the other hand, platforms can take steps to mitigate negativity bias at the user level and prevent it from becoming algorithmic negativity bias and spreading to the collective level.

Studies of the effect of recommender algorithms are an emerging field in academia that should be supported by the relevant authorities in order to identify, in all independence, the right regulatory levers and implement an evidence-based policy. Needless to say, this will require greater openness of SNSs data towards academia [5]. Some of the empirical calibration made on Twitter in this study, like the opinion update rule or the reshare behavior, could be usedful to model other platforms like Facebook, but nothing compares to an empirical calibration on the native data of a platform

In conclusion, it is not enough to point to malicious users who produce toxic content and call for better moderation. We need to further study the effects, at the individual and collective levels, of large-scale deployment of recommender systems by major technology companies and assess their potential harm. Science shall contribute to evidence-based policymaking by modeling the impact of these platforms on the social fabric. Democracy is at stake.

Aknowledgments

We are thankful to Victor Chomel for the fruitful discussions, as well as Jeanne Bruneau–Bongard for her careful proofread of the first draft of the present article. This work was supported by a grant from the "Fondation CFM pour la Recherche", as well

5. Let us remind that Twitter has been one of the sole large SNSs to make some of its data available and was considering ending this data openness at the time this article was written.

as the support of the Complex Systems Institute of Paris Île-de-France and the Region Île-de-France.

Références

Allcott, Hunt, Braghieri, Luca, Eichmeyer, Sarah, & Gentzkow, Matthew. 2020. The welfare effects of social media. *American Economic Review*, **110**(3), 629–76.

Barberá, Pablo. 2015. Birds of the Same Feather Tweet Together : Bayesian Ideal Point Estimation Using Twitter Data. *Political Analysis*, **23**(1), 76–91.

Bastian, Mathieu, Heymann, Sebastien, & Jacomy, Mathieu. 2009. *Gephi : An Open Source Software for Exploring and Manipulating Networks*.

Baumann, Fabian, Lorenz-Spreen, Philipp, Sokolov, Igor M., & Starnini, Michele. 2020. Modeling echo chambers and polarization dynamics in social networks. *Physical Review Letters*, **124**(4), 048301. arXiv :1906.12325 [physics].

Bouchaud, Paul, Chavalarias, David, & Maziyar, Panahi. 2023. *Replication Data for : Chavalarias, Bouchaud, Panahi (2023) Can few lines of code change Society ? Beyond fack-checking and moderation : how recommender systems toxifies social networking sites*.

Briatte, François, & Gallic, Ewen. 2015 (May). Recovering the French Party Space from Twitter Data. *In : Science Po Quanti*. 00000.

Ceylan, Gizem, Anderson, Ian A., & Wood, Wendy. 2023. Sharing of misinformation is habitual, not just lazy or biased. *Proceedings of the National Academy of Sciences*, **120**(4), e2216614120. Publisher : Proceedings of the National Academy of Sciences.

Chavalarias, David. 2022. *TOXIC DATA - Comment les réseaux manipulent nos opinions*. Flammarion edn.

Chen, Tianqi, & Guestrin, Carlos. 2016. XGBoost : A Scalable Tree Boosting System. *In : Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM.

Chomel, Victor, Cuvelle-Magar, Nathanaël, Panahi, Maziyar, & Chavalarias, David. 2022 (Nov.). Polarization identification on multiple timescale using representation learning on temporal graphs in Eulerian description.

Conover, Michael D., Goncalves, Bruno, Ratkiewicz, Jacob, Flammini, Alessandro, & Menczer, Filippo. 2011. Predicting the Political Alignment of Twitter Users. *In : 2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing*. IEEE.

Deffuant, Guillaume, Neau, David, Amblard, Frederic, & Weisbuch, Gérard. 2000. Mixing beliefs among interacting agents. *Advs. Complex Syst.*, **03**(01n04), 87–98.

Dimock, Michael, & Wike, Richard. 2020 (Nov.). *America is exceptional in the nature of its political divide*.

Doherty, Carroll, Kiley, Jocelyn, & Jameson, Bridget. 2016 (June). *Partisanship and Political Animosity in 2016*. Tech. rept. Pew Research Center.

Epstein, Robert. 2018 (Apr.). The Search Suggestion Effect (SSE) : How Search Suggestions Can Be Used to Shift Opinions and Voting Preferences Dramatically and Without People's Awareness. *In : 98 th annual meeting of the Western Psychological Association*.

Fortin, Félix-Antoine, De Rainville, François-Michel, Gardner, Marc-André, Parizeau, Marc, & Gagné, Christian. 2012. DEAP : Evolutionary Algorithms Made Easy. *Journal of Machine Learning Research*, **13**(jul), 2171–2175.

Garimella, Kiran, Morales, Gianmarco De Francisci, Gionis, Aristides, & Mathioudakis, Michael. 2018. Quantifying Controversy on Social Media. *ACM Transactions on Social Computing*, **1**(1), 1–27.

Gaumont, Noé, Panahi, Maziyar, & Chavalarias, David. 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), e0201879.

Hagey, Keach, & Horwitz, Jeff. 2021. Facebook Tried to Make Its Platform a Healthier Place. It Got Angrier Instead. *Wall Street Journal*, Sept.

Huszár, Ferenc, Ktena, Sofia Ira, O'Brien, Conor, Belli, Luca, Schlaikjer, Andrew, & Hardt, Moritz. 2022. Algorithmic amplification of politics on Twitter. *Proceedings of the National Academy of Sciences*, **119**(1), e2025334119.

Jacomy, Mathieu, Venturini, Tommaso, Heymann, Sebastien, & Bastian, Mathieu. 2014. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLoS ONE ONE*, **9**(6), e98679.

Jager, Wander, & Amblard, Frédéric. 2005. Uniformity, Bipolarization and Pluriformity Captured as Generic Stylized Behavior with an Agent-Based Simulation Model of Attitude Change. *Computational Mathematical Organization Theory*, jan.

Jia, Peng, MirTabatabaei, Anahita, Friedkin, Noah E., & Bullo, Francesco. 2015. Opinion Dynamics and the Evolution of Social Power in Influence Networks. *SIAM Review*, **57**(3), 367–397. Publisher : Society for Industrial and Applied Mathematics.

Knobloch-Westerwick, Silvia, Mothes, Cornelia, & Polavin, Nick. 2017. Confirmation Bias, Ingroup Bias, and Negativity Bias in Selective Exposure to Political Information. *Communication Research*, **47**(1), 104–124.

Leicht, E. A., & Newman, M. E. J. 2008. Community Structure in Directed Networks. *Physical Review Letters*, **100**(11).

Levinsson, Anna, Miconi, Diana, Li, Zhiyin, Frounfelker, Rochelle L., & Rousseau, Cécile. 2021. Conspiracy Theories, Psychological Distress, and Sympathy for Violent Radicalization in Young Adults during the COVID-19 Pandemic : A Cross-Sectional Study. *International Journal of Environmental Research and Public Health*, **18**(15), 7846. Number : 15 Publisher : Multidisciplinary Digital Publishing Institute.

Martin, Louis, Muller, Benjamin, Suárez, Pedro Javier Ortiz, Dupont, Yoann, Romary, Laurent, de la Clergerie, Éric, Seddah, Djamé, & Sagot, Benoît. 2020. CamemBERT : a Tasty French Language Model. *In : Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

McPherson, Miller, Smith-Lovin, Lynn, & Cook, James M. 2001. Birds of a Feather : Homophily in Social Networks. *Annual Review of Sociology*, **27**(1), 415–444.

Mestre, Abel. 2022. Eric Zemmour, nouveau président de la fachosphère ? *Le Monde*, Mar.

Noorazar, Hossein, Vixie, Kevin R., Talebanpour, Arghavan, & Hu, Yunfeng. 2020. From classical to modern opinion dynamics. *International Journal of Modern Physics C*, **31**(07), 2050101. Publisher : World Scientific Publishing Co.

Perra, Nicola, Gonçalves, Bruno, Pastor-Satorras, Romualdo, & Vespignani, Alessandro. 2012. Activity driven modeling of time varying networks. *Scientific reports*, **2**(1), 1–7. Publisher : Nature Publishing Group.

Politics, UVA Center for. 2021 (Sept.). *New Initiative Explores Deep, Persistent Divides Between Biden and Trump Voters – Sabato's Crystal Ball*.

Ramaciotti Morales, Pedro, & Cointet, Jean-Philippe. 2021 (Sept.). Auditing the Effect of Social Network Recommendations on Polarization in Geometrical Ideological Spaces. *In : RecSys '21 : 15th ACM Conference on Recommender Systems*.

Reyero, Tomás Mussi, Beiró, Mariano G., Alvarez-Hamelin, J. Ignacio, Hernández, Laura, & Kotzinos, Dimitris. 2021. Evolution of the political opinion landscape during electoral periods. *EPJ Data Science*, **10**(1), 31. Number : 1 Publisher : Springer Berlin Heidelberg.

Rozin, Paul, & Royzman, Edward B. 2001. Negativity Bias, Negativity Dominance, and Contagion. *Personality and Social Psychology Review*, **5**(4), 296–320. Publisher : SAGE Publications Inc.

Santos, Fernando P., Lelkes, Yphtach, & Levin, Simon A. 2021. Link recommendation algorithms and dynamics of polarization in online social networks. *Proceedings of the National Academy of Sciences*, **118**(50), e2102141118. Publisher : Proceedings of the National Academy of Sciences.

Tokita, Christopher K., Guess, Andrew M., & Tarnita, Corina E. 2021. Polarized information ecosystems can reorganize social networks via information cascades. *Proceedings of the National Academy of Sciences*, **118**(50), e2102147118. Company : National Academy of Sciences Distributor : National Academy of Sciences ISBN : 9782102147111 Institution : National Academy of Sciences Label : National Academy of Sciences Publisher : Proceedings of the National Academy of Sciences.

Traag, V. A., Waltman, L., & van Eck, N. J. 2019. From Louvain to Leiden : guaranteeing well-connected communities. *Sci Rep*, **9**(1).

Twitter. 2020. *Using Twitter*.

Törnberg, Petter. 2022. How digital media drive affective polarization through partisan sorting. *Proceedings of the National Academy of Sciences*, **119**(42), e2207159119.

Vosoughi, Soroush, Roy, Deb, & Aral, Sinan. 2018. The spread of true and false news online. *Science*, **359**(6380), 1146–1151.

Xu, Zhiheng, & Yang, Qing. 2012. Analyzing User Retweet Behavior on Twitter. *In : 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE.

Zubrow, Keith. 2021. Facebook whistleblower says company incentivizes "angry, polarizing, divisive content". *CBS NEWS*, Oct.