Supporting information

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 harmfull implementations of recommender systems to individuals and society in order better regulate their deployment on systemic SNSs.

1 Opinion

Wishing opinions spanning continuously the range $[-1, +1]$ and exhibiting more refinement than a simple dichotomy or clustering, we decided to assign to each user an opinion corresponding to leaders' opinion weighted by the inverse distance to the leader; euclidean distance measured in the projected space obtained through the force-directed layout algorithm ForceAtlas2 (Jacomy *et al.*, 2014) (with default settings). Within this layout, nodes —in total disregard to theirs attributes— repulse each others while (undirected) edges attract their source/target nodes —proportional to the associated weight, if any. The resulting position of a node cannot be interpreted on its own, but only compared to others. On the above retweet graph, the higher the number of retweet between two users, the closest those two nodes.

We are left with the task of assigning a numerical opinion to the political leaders, this crucial task will once again be carried out using the graph structure. We considered these opinions as the angular difference between the vector (barycenter-leader) with respect to the reference vector (barycenter-Macron) in the projected space, here the barycenter is calculated among all leaders. The reference direction has been chosen for two reasons : first the community around Emmanuel Macron is quite stable over time, especially compared to the far-right/far-left communities, avoiding having unstable anchored points. Secondly, when expressing their views on a given issue other political leaders, use de-facto Emmanuel Macron —the sitting president— as a reference. The spatialization and thus leaders' opinion, rely on the activity of their community, evolving with time.

Far from being a drawback, the dynamical nature of opinions allocation conveys the continuous adaptation, reshaping of the political landscape caused both by endogenous and exogenous events. For example the intensification of Eric Zemmour political ambitions is reflected by a relative inversion between Marine Le Pen and Eric Zemmour, two far-right figures, between September and October 2021, the former appearing "less extreme" than the latter in October, as displayed in table 1, —the political Twittersphere of October 2021 is depicted in figure 5. To someone initiated to French politics, the relative opinions of the different leader reflects quite accurately the different political current, Hidalgo, Jadot, Roussel, Poutou and Melenchon at the left of Macron, Pecresse, Zemmour, LePen, Dupont Aignan, Philipot, Asselineau at his right.

The circularity of the arena motivates us to consider periodic opinions, with a transition suited between Melenchon and Asselineau, corresponding to "conspiracy views". The opinion assignment process leads to a distribution depicted in figure 6.

2 Opinion update parameter

The result of the symbolic regression on the functional forms of μ was the very expression, $o_i \leftarrow o_i + \mu_i(o_j - o_i)$, used in the opinion dynamics literature as in the Deffuant model (Deffuant *et al.*, 2000) or in the Jager and Amblard model (Jager & Amblard, 2005).

Then, we fitted for each agent the opinion update parameter λ_i. In order for the calibration to be computationally efficient, we fix during the fitting procedure the opinion of the other agents and only update the considered user's one. Such an approximation is reasonable considering the distribution of monthly change of opinion, the vast majority of users only changing slightly their opinion — the average change between September and October 2021 equals -0.03.

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. 8). This is less than intra-communities opinion diversity.

To verify the sanity of the opinion update parameter fitting procedure, we considered another time period, Spring 2020 in addition to Autumn 2021, as well as other spatialization settings. Gephi (Bastian *et al.*, 2009) Force Atlas2 setting used by default were modified —stronger gravity coupled to gravity sets to 0.001 and a scale sets to 5— leading to the Twittersphere depicted in figure 7. While we notice a relative inversion between far-right leaders such as Nicolas Dupont Aignan and Eric Zemmour, the overall arena is similar. All in all, the general accuracy is equivalent between the different variants, 85% of the predictions off by less than 0.1 (figure 8), more than 60% of the prediction offs by

less than 25% for large displacement $\Delta_{op} > 0.05$ (figure 9). Also, using the fitted opinion update parameter, we forecast the change of opinion, using the list of retweets effectively exchanged during the month following the fitting. The accuracy is poorer but yet acceptable considering that the goal of the present work is not to predict the opinion of users, but only to faithfully simulate their online retweet behavior. As displayed on figure 10, the relative error for monthly opinion changes larger than $\Delta_{op} > 0.5$ is less than 25% for more than 75% of the predictions.

3 Calibration of agents' activities

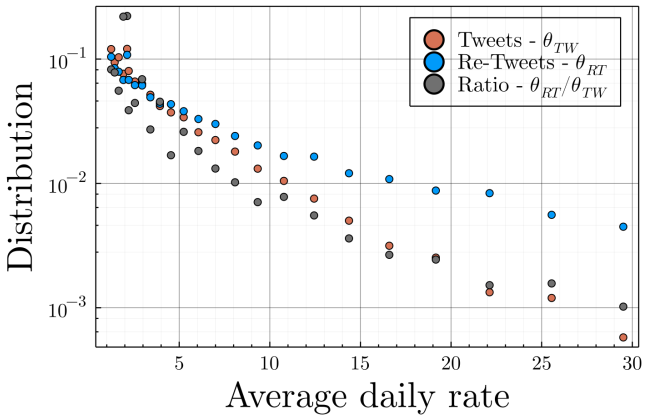

FIGURE S1 – Distribution of accounts activities in the Politoscope over the period October 01-30 2021.

Latitude of acceptance

While retweeting a message is generally a sign of agreement, quoting one may express a variety of intermediate positions, including total disagreement. The study of the distribution associated with quotes allow us to verify that the one associated with retweets is not a mere consequence of the process of assigning opinions to users, the latter being solely based on the retweets graph and not the quotes' one. In order to estimate the probability that a given user will retweet a read tweet which diverges from its own opinion by Δ_{op}, we renormalize the distribution associated with retweets by the one associated with quotes, binning readers' opinion. Indeed, in order for a user to quote a message, the recommender should have shown it to the user —under the reasonable assumption that the large majority of users' actions on Twitter is ruled by the *Home timeline* and not by manual searches— the renormalization allows us to cancel the bias in the representation made by the platform (Huszár *et al.*, 2022), as well as taking account the different sizes and positions of the communities. However, by renormalizing we de-facto neglect potential political strategy such as quoting massively the opposite site to attack the leader or to gain in visibility. Further work, would once again greatly benefit from having access to impressions information.

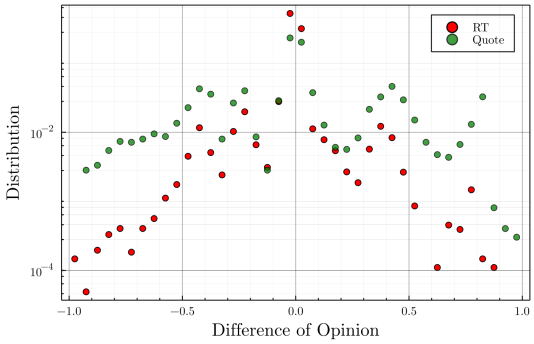

FIGURE S2 – Empirical distribution of the differences of opinions between members of Macron (center) community and the opinions of accounts that are retweeted or quoted, $\Delta_{op} = o_{reader} - o_{sender}$, considering periodic boundary conditions. The quote distribution is used to renormalize the RT ditribution to get the normalized distribution of Fig. . We renormalize such that a perfectly aligned message is retweeted with certainty.

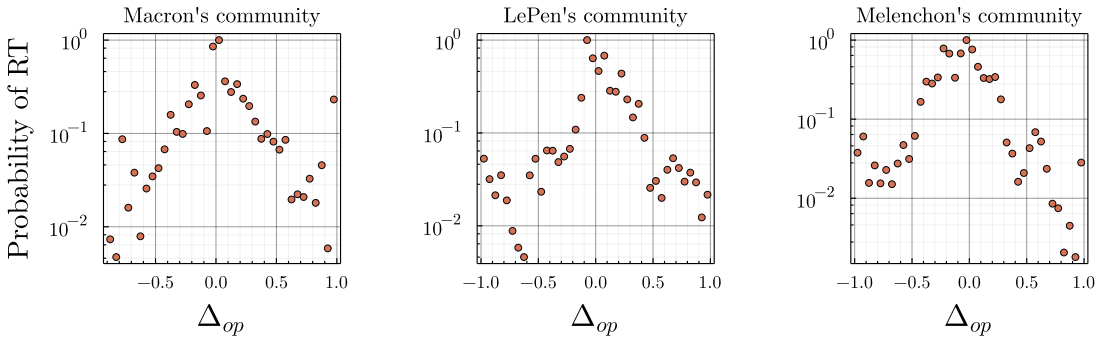

FIGURE S3 – Estimated probability for a user, in the ideological neighborhood of Emmanuel Macron (center), Marine Le Pen (extreme right) or Jean-Luc Melenchon (radical left) to retweet a read message according to the different of opinion with its author, $\Delta_{op} = o_{reader} - o_{sender}$, considering periodic boundary conditions. We renormalize such that a perfectly aligned message is retweeted with certainty.

4 Negativity

The negativity considered within the sentiment analysis, is understood in the psychological sense; a message is labeled as negative if it is unpleasant, offending, harmful, inciting revolt etc., in total disregard of the societal or political implication. By manually labelling a thousand of tweets, we estimated the overall accuracy of the CamemBERT classification around 73%. We refined the accuracy estimation by distinguishing clearly negative tweets such as :

> Eric Ciotti [@ECiotti] Un adolescent interpellé à #Marseille avec plusieurs centaines de grammes de cannabis/cocaïne et avec un fusil à pompe. Le matériel du parfait écolier, tout va très bien madame la marquise !"["A teenager arrested in #Marseille with several hundred grams of cannabis/cocaine and with a shotgun. The equipment of the perfect schoolboy, everything is fine

madam the marquise !"] Twitter, 1 Oct 2021, https://twitter.com/ECiotti/status/1443944007143407624

leading to an accuracy close to 89%, from less negatively blunt messages, such as :

Gérald DARMANIN [@GDarmanin] "C'est le devoir de chaque Français que de se souvenir des visages innocents de Nadine, Simone et Vincent, brutalement arrachés à la vie par une idéologie mortifère"[It is the duty of every French person to remember the innocent faces of Nadine, Simone and Vincent, brutally torn from life by a deadly ideology", (the three victims of 2020 stabbing attack at Notre-Dame de Nice)] Twitter, 29 Oct 2021, https://twitter.com/GDarmanin/status/1454130435039109122

leading to an accuracy of 78%. The accuracy related to tweets for which the determination of negativity —in the above-defined sense— is even fuzzy for a human speaker is close to 60% :

Clémentine Autain [@Clem_Autain] "@Anne_Hidalgo a dit qu'elle n'utiliserait pas les mots crime contre l'humanité pour parler de la colonisation. On aurait pu imaginer que ces propos fassent un tollé, mais non. [...] On choisit ce qui est monté en épingle."[@Anne_Hidalgo said that she would not use the words "crime against humanity" to refer to colonization. One could have imagined that these words would cause an outcry, but no. [...] We choose what we want to make a fuss about] Twitter, 16 Oct 2021, https://twitter.com/Clem_Autain/status/1449377126709399554

Finally the accuracy for neutral/positives tweets is close to 72% :

Bruno Le Maire [@BrunoLeMaire] "C'est par le travail que nous créons le pouvoir d'achat pour les Français. Depuis 2017, un million d'emplois ont été créés par les entreprises."[It is through work that we create purchasing power for the French. Since 2017, one million jobs have been created by businesses.] Twitter, 20 Oct 2021, https://twitter.com/brunolemaire/status/1450904607111139338?lang=fr

Philippe Poutou [@PhilippePoutou] "En soutien aux étudiants et étudiantes sans-fac à Nanterre, qui ne demandent rien d'autre que d'avoir le droit d'étudier, et dans de bonnes conditions." [In support of students without a university in Nanterre, who ask for nothing more than to have the right to study, and in good conditions.] Twitter, 13 Oct 2021, https://twitter.com/PhilippePoutou/status/1448287937238638593

As displayed on Figure 11, the users having exchanged political content are heavily negative, half of them published more than 60% of negative messages, and a quarter more than 75%. Figure 15 exhibits the average negativity as a function of the opinion, one notices a correlation between the negativity and the opinion extremity. Opinion extremity hereby considered as the absolute value of the opinion, with Emmanuel Macron as a reference at 0, the historical moderate parties represented by their leaders Anne Hidalgo and Valerie Pecresse around 0.3 and more extreme candidates at more than 0.6 like Jean-Luc Melenchon or Marine Le Pen. The correlation between the average negativity and the absolute value of opinion equals 0.3 ($p < 10^{-9}$). Determining the average negativity of every user present in the above depicted Twittersphere is unrealistic considering the computational cost of the sentiment analysis and the need of sufficient tweets for each

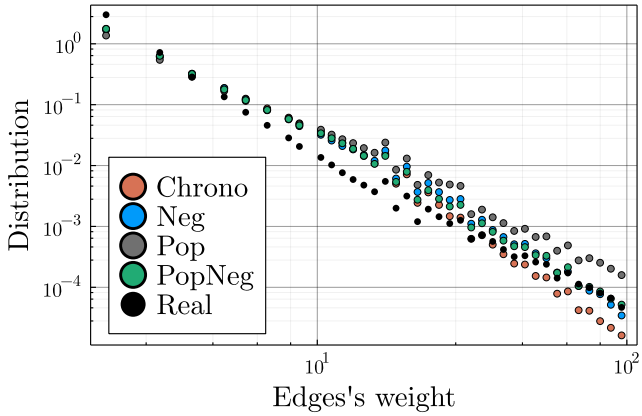

FIGURE S4 – Distribution of edges' weights for the four implemented recommenders as well as the empirical distributions in October 2021

user to obtain a significant statistics. We will then assign to our user an intrinsic negativity drawn from the empirical distribution in function of their initial opinion.

The sentiment analysis performed on the tweets allow us to estimate the negativity bias of our users. Such an estimation is arduous due to the mere use of Twitter recommender and by our impossibility to know what message is actually presented by the platform to the users. As displayed on figure 17, more than half of the messages are not retweeted a single time, only Twitter knows if its because these messages are bland or just have not been shown to others. The average negativity of messages decreases for an increasing number of retweets bellow few hundreds —as display on figure 18— retweet that we suppose to be associated to the author identity instead of the mere content, hypothesis to be verified in further works. The average negativity then increase significantly for a large number of retweets : highly popular messages are heavily negative.

In absence of impressions information, every estimation of the negativity bias is debatable, we nevertheless assumed that the messages published by a given political leader are presented by the algorithmic recommendation in a similar way. Hence we estimate the negativity bias of our agents by comparing for each leader the average number of retweets for the messages labeled as negative and labeled as neutral/positive. The estimated negativity bias is presented for the different political leaders in table 2 ; using this estimation we assigned a negativity bias to our users based on the leaders present in their communities.

TABLE S1 – 2022 French Presidential candidates, having significant online presence, determined opinion in September and October 2021

Candidates	September 2021	October 2021
DupontAignan	-0.72	-0.76
LePen	-0.62	-0.60
Zemmour	-0.60	-0.63
Pecresse	-0.26	-0.28
Macron	0.00	0.00
Hidalgo	0.25	0.28
Jadot	0.42	0.42
Roussel	0.46	0.43
Poutou	0.71	0.68
Arthaud	0.73	0.66
Melenchon	0.82	0.77

TABLE S2 – Negativity bias associated to the community having retweeted political leaders in October 2021

Leaders	NegBias	Leaders	NegBias	Leaders	NegBias	Leaders	NegBias	Leaders	NegBias
Asselineau	1.38	Philippot	1.15	DupontAignan	2.13	Collard	1.64	LePen	1.18
Zemmour	1.34	Ciotti	4.49	Wauquiez	2.56	Pecresse	1.5	Darmanin	2.15
LeMaire	1.06	Macron	1.14	Bayrou	0.94	Hidalgo	0.96	Faure	1.04
Jadot	0.99	Montebourg	1.88	Roussel	1.33	Bayou	1.8	Poutou	6.1
Arthaud	2.05	Autain	4.0	ManonAubry	1.06	Quatennens	1.57	Melenchon	1.37

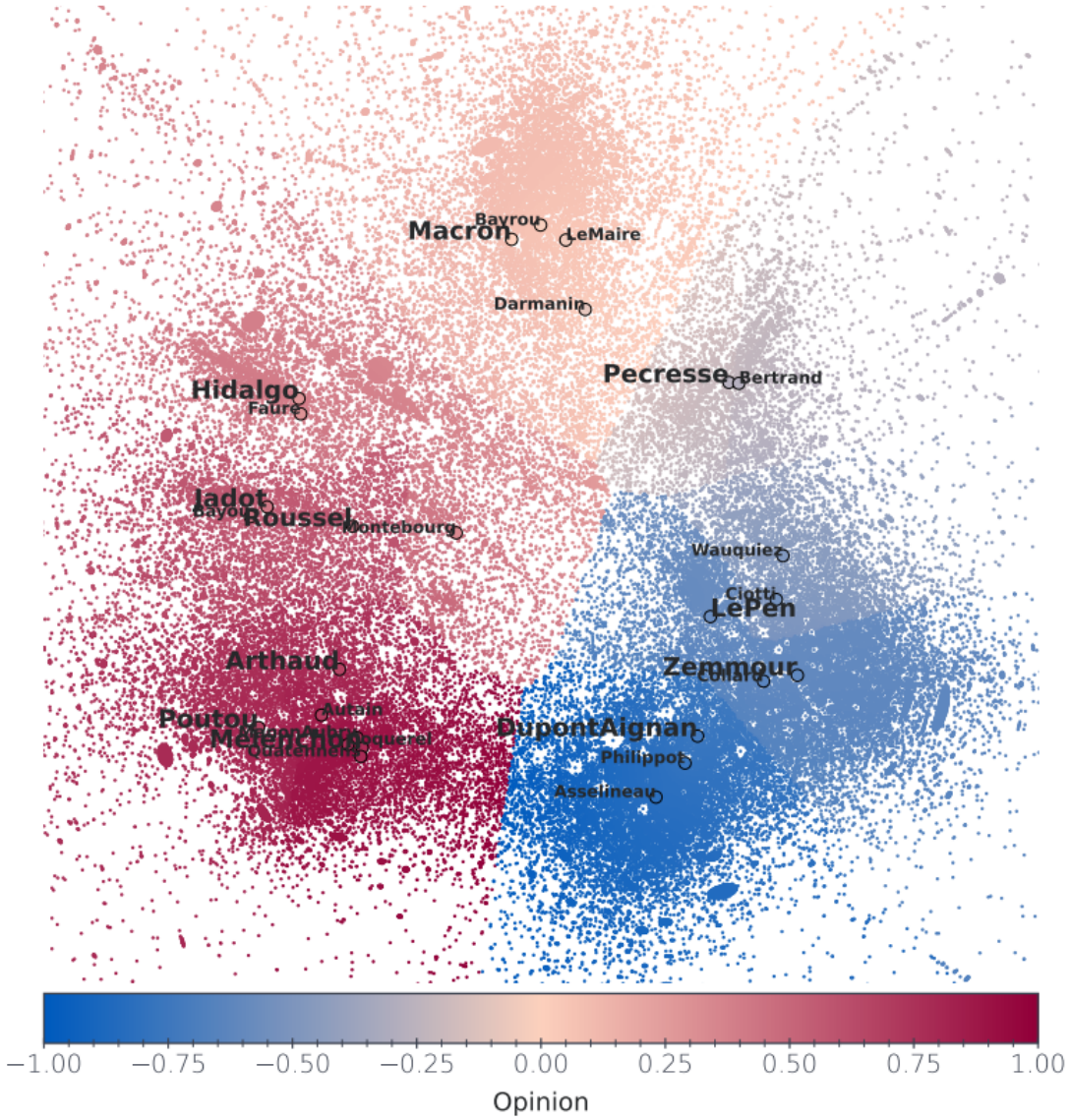

FIGURE S5 – Multi-polar graph of the French pre-electoral political Twittersphere calculated during October 2021. Each node corresponds to a user, colored according to the opinion assigned by the described method. Political leaders are highlighted, in particular the candidates for 2022 French presidential election

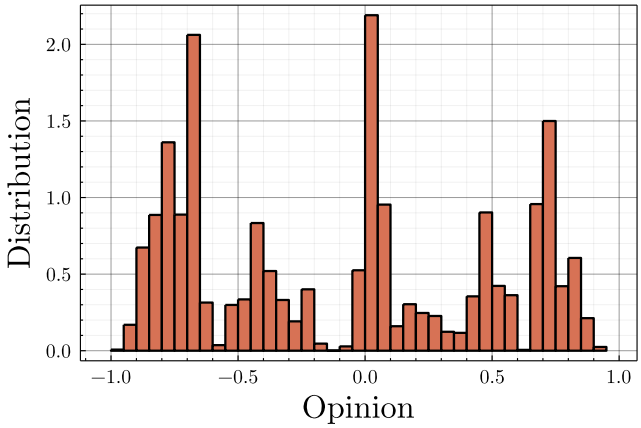

FIGURE S6 – Distribution of users' assigned opinion in September 2021 using the above described method

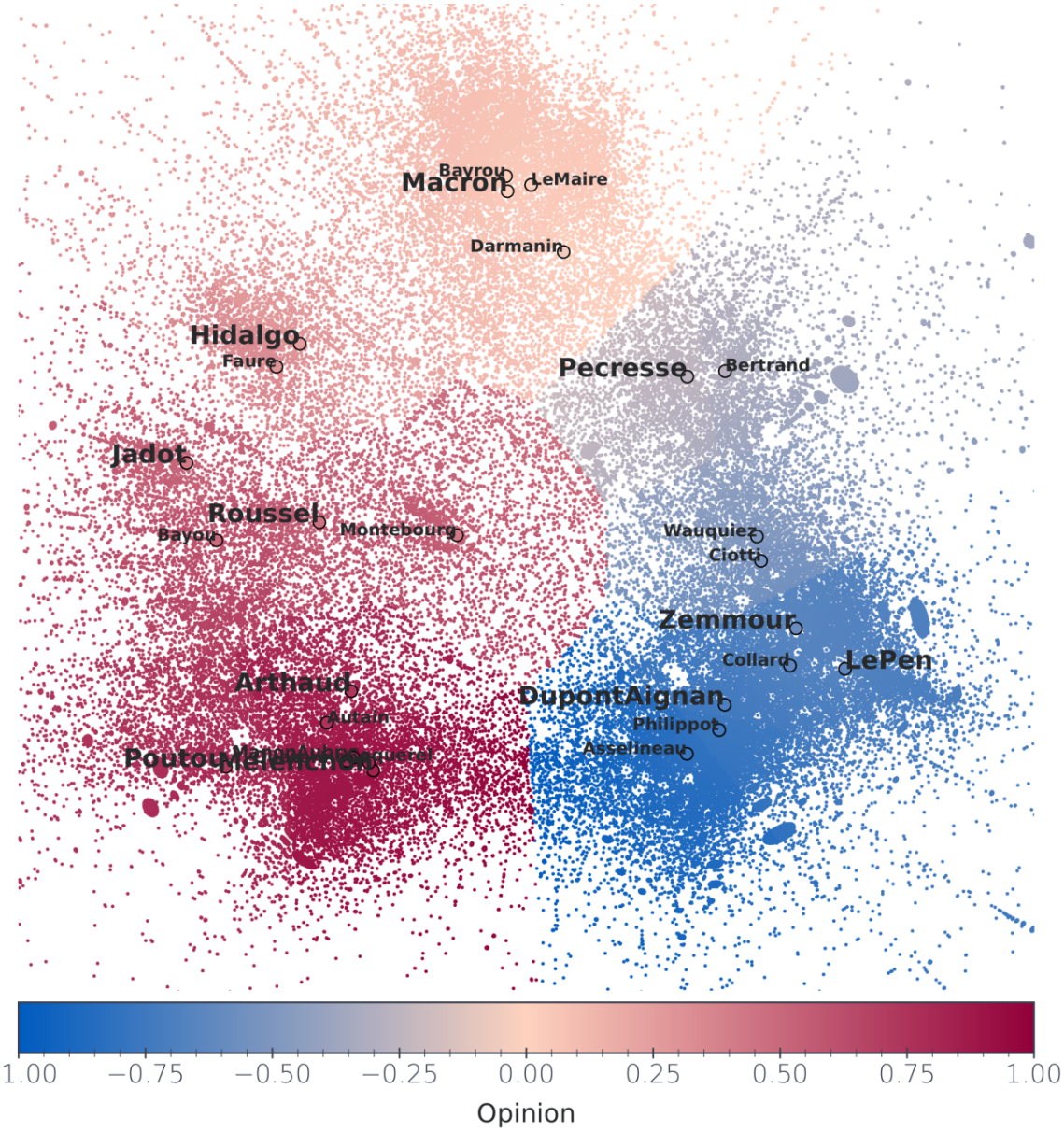

FIGURE S7 – Multi-polar graph of the French pre-electoral political Twittersphere calculated during September 2021, using custom ForceAtlas2 graph spatialization settings : strong gravity, gravity sets to 0.001 and a scale sets to 5. Each node corresponds to a user, colored according to the opinion assigned by the described method. Political leaders are highlighted, in particular the candidates for 2022 French presidential election

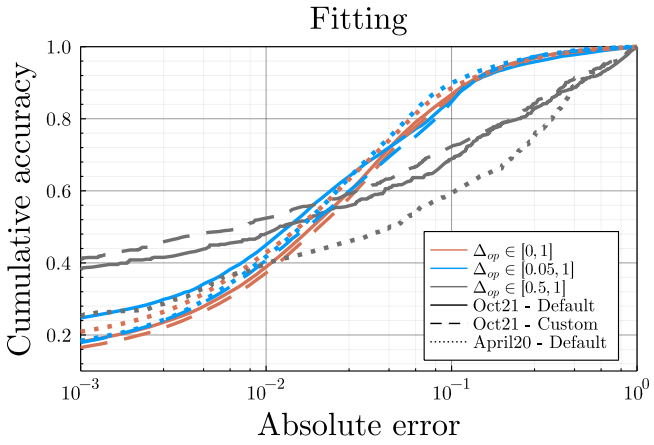

FIGURE S8 – Cumulative accuracy of the opinion update parameter μ fitting procedure in function of the absolute error, for various monthly opinion changes, time periods and graph spatialization settings

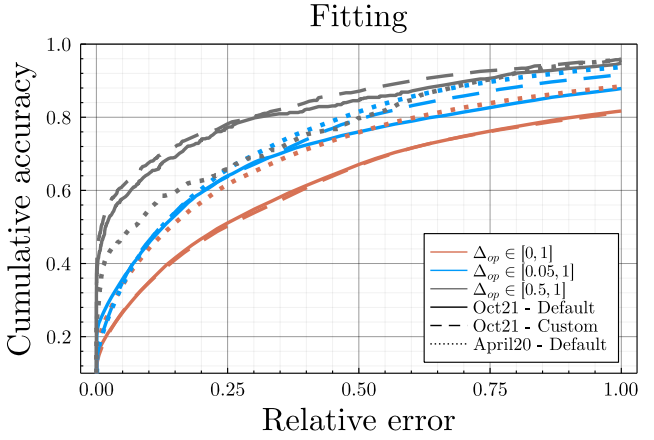

FIGURE S9 – Cumulative accuracy of the opinion update parameter μ fitting procedure in function of the relative error, for various monthly opinion changes, time periods and graph spatialization settings

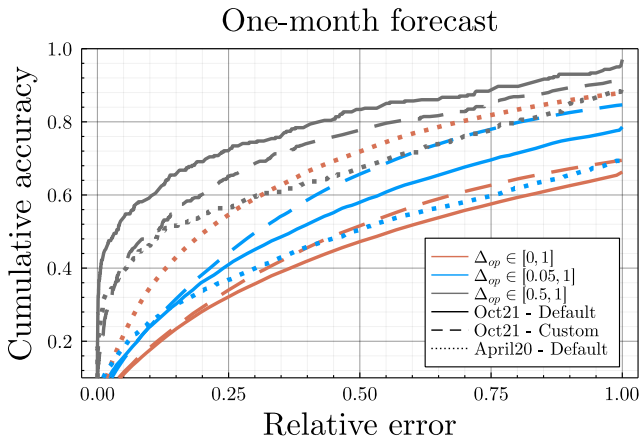

FIGURE S10 – Cumulative accuracy of the forecasted opinion, one month after fitting the opinion update parameter μ, in function of the relative error, for various monthly opinion changes, time periods and graph spatialization settings

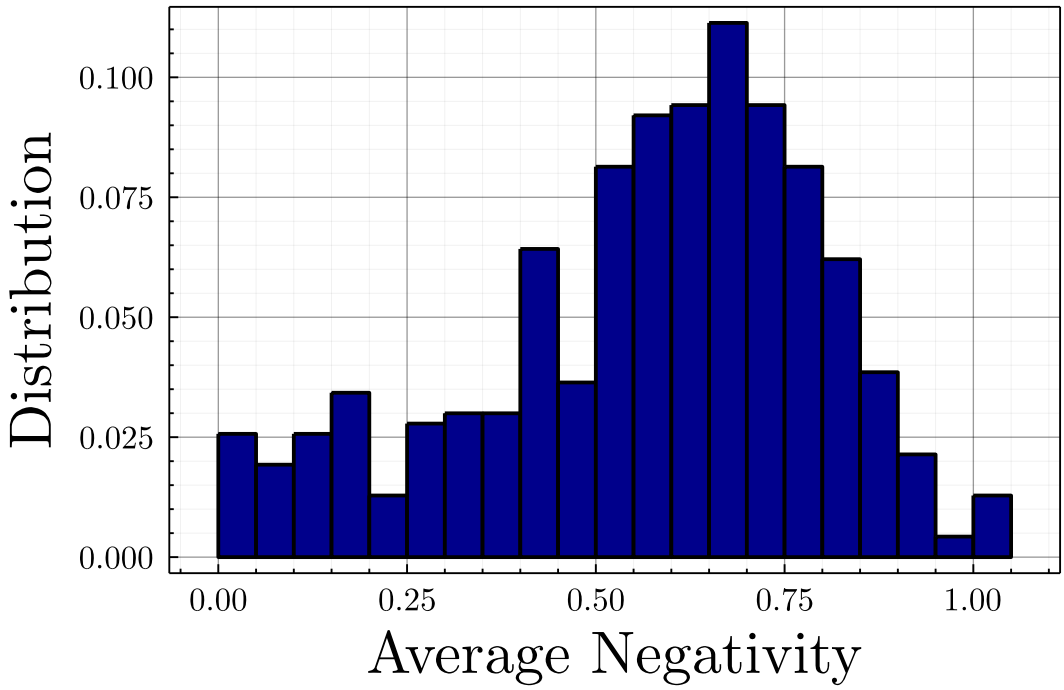

FIGURE S11 – Empirical distribution of users average negativity. Analysis restricted to the 480 users having published more than 5 messages during October 2021

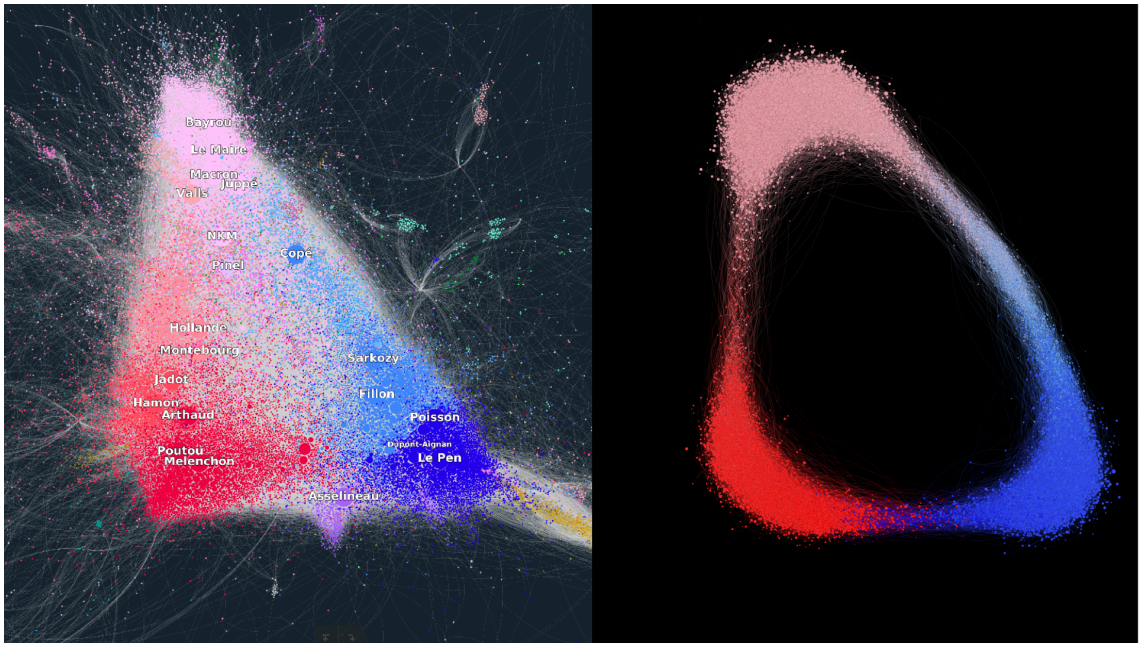

FIGURE S12 – Visual comparison between a real network from the French political twittersphere (left) and a synthetic random graph initilized according to the Barabási Albert model and evolved under *PopNeg* (right). As confirmed in main text, the synthetic graph successfully reproduces the modular structure of online political landscapes.

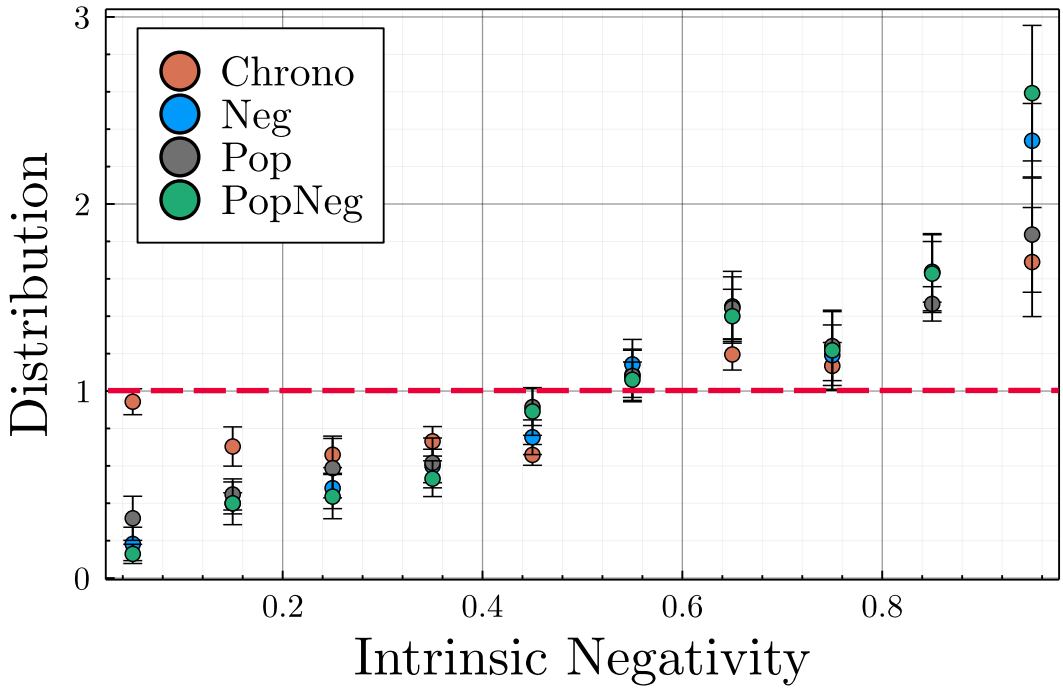

FIGURE S13 – Intrinsic negativity unbalance between the overall population and the top 1% agents receiving the most retweet by tweet in synthetic graphs (proportion of agent in the top 1% normalized by the proportion of agent in the whole population. The dash line represent the balanced ratio. The more negative an agent is, the more likely it is to be in the top 1%. Error bars represent the 95% confidence interval.

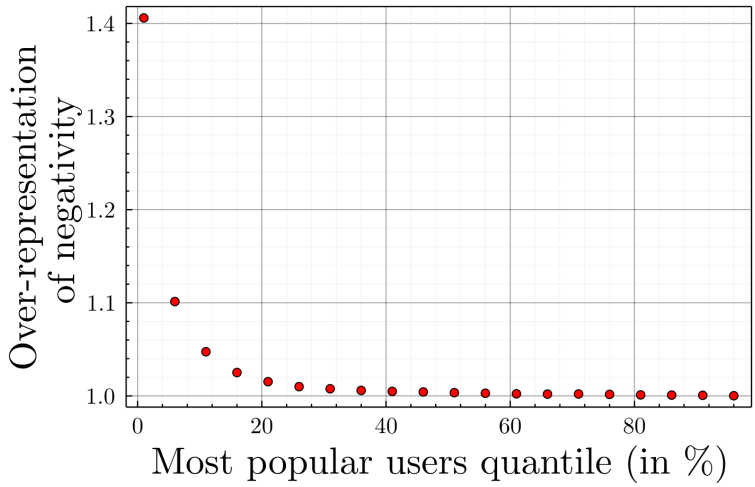

FIGURE S14 – Over-representation of negative agents among the top 1% of popular users for synthetic graphs under *Pop* algorithm.

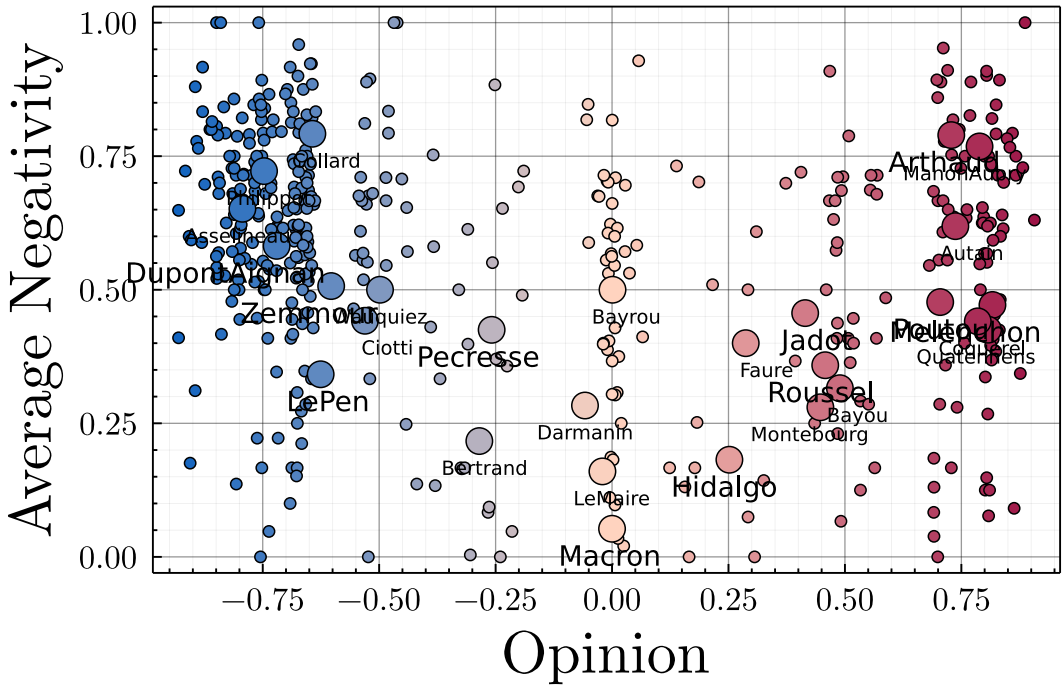

FIGURE S15 – Average negativity by user in function of their opinion, with political leaders highlighted, in particular the candidates for 2022 French presidential election. Analysis restricted to the 480 users having published more than 5 messages during October 2021

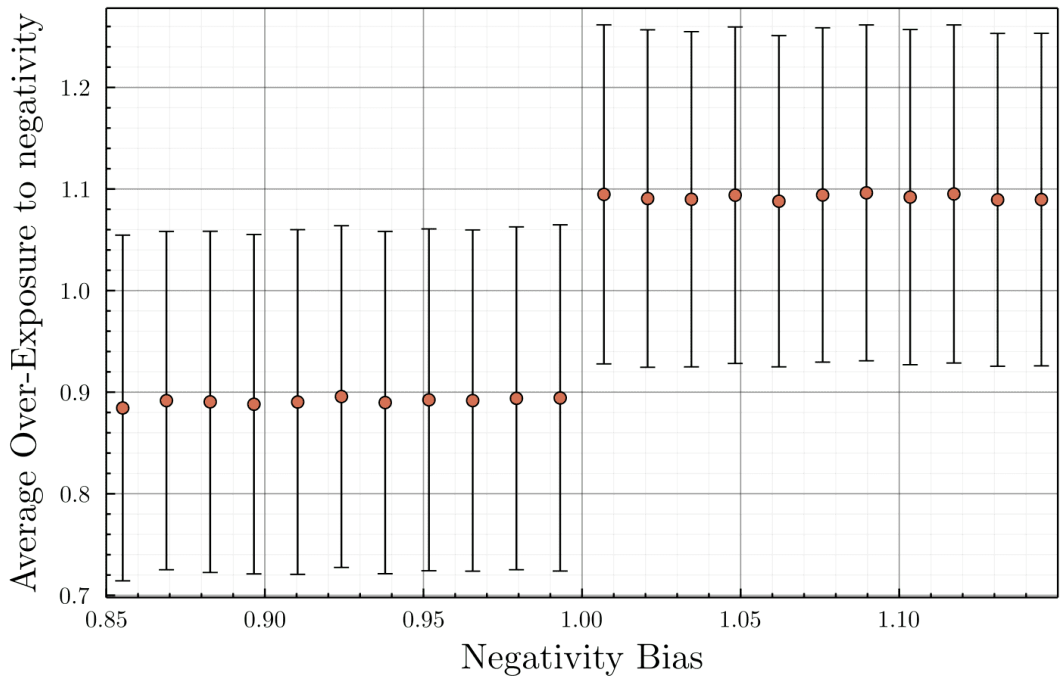

FIGURE S16 – Average overexposure to negativity in function of users negativity bias using *Pop* recommender in synthetic graphs.

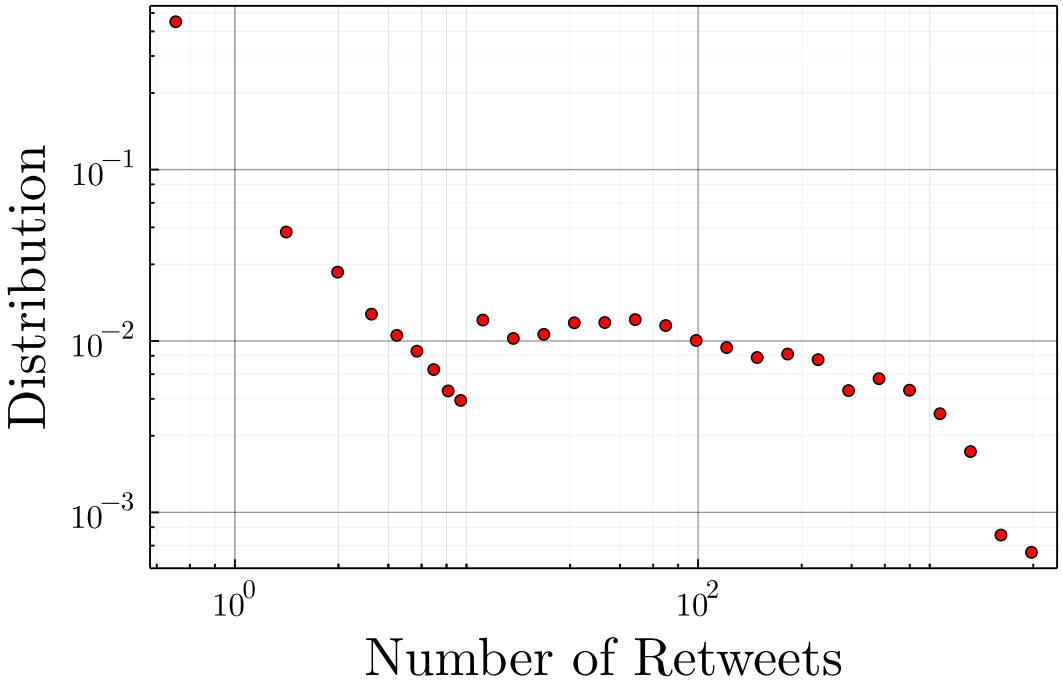

FIGURE S17 – Distribution of number of retweets, analysis performed on French political related tweets published in October 2021

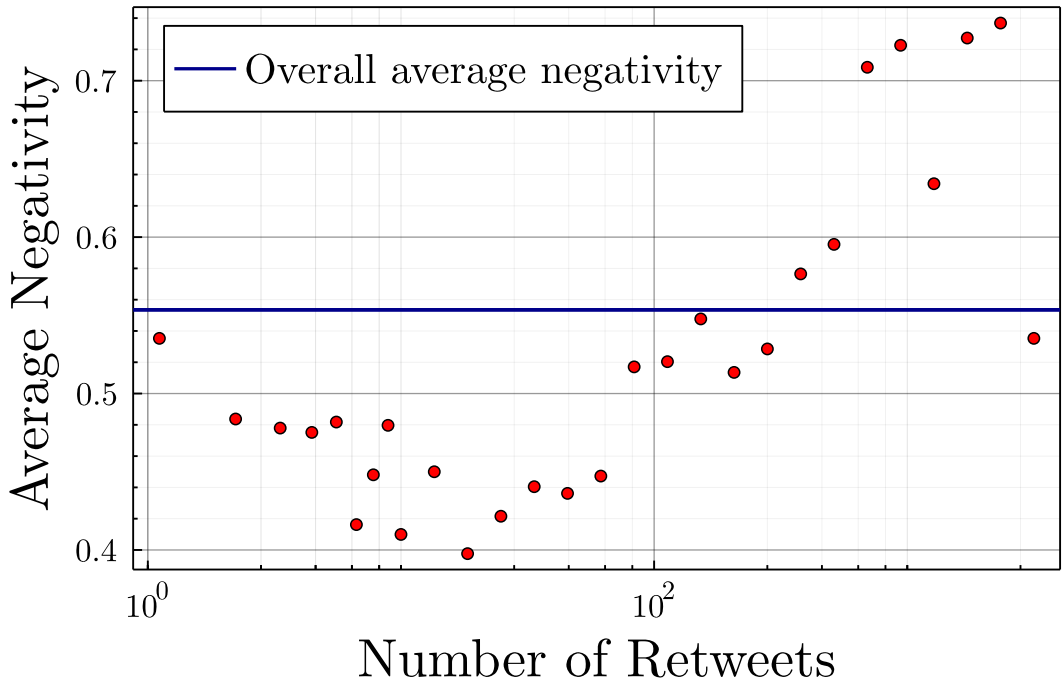

FIGURE S18 – Empirical average negativity of the French political related tweets published in October 2021 in function of the number of retweet

Références

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

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

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.