Skewed Perspectives: Examining the Influence of Engagement Maximization on Content Diversity in Social Media Feeds

Paul Bouchaud

Skewed Perspectives: Examining the Influence of Engagement Maximization on Content Diversity in Social Media Feeds

Paul Bouchaud[1,2,*]

[1]EHESS/CAMS, Paris, France
[2]CNRS/ISCPIF, Paris, France
[*]paul.bouchaud@iscpif.fr

ABSTRACT

This article investigates the information landscape shaped by curation algorithms that seek to maximize user engagement. Leveraging unique behavioral data, we trained machine learning models to predict user engagement with tweets. Our study reveals how the pursuit of engagement maximization skews content visibility, favoring posts similar to previously engaged content while downplaying alternative perspectives. The empirical grounding of our work contributes to the understanding of human-machine interactions and provides a basis for evidence-based policies aimed at promoting responsible social media platforms.

1 Introduction

Social media, such as Facebook or Twitter, have transformed the way information is disseminated, interactions are formed, and public discourse takes shape. With an evergrowing number of users, the large scale deployment of algorithms curating information have become a cornerstone of such plateforms[1]. Aiming to generate advertising revenue, social media companies employ sophisticated deep-learning algorithms[2,3] to maximize user engagement. However, emerging concerns surrounding these platforms revolve around the algorithmic mechanisms dictating the curation of content.

While engagement maximization, showing content that may elicits users engagement, might seem harmless, the implications of such optimization are far-reaching and pose significant threat to the diversity and objectivity of information presented to social media users[4]. These concerns are fuelled by their offline consequences. For example, seeking to maximize engagement, Facebook's algorithms has compelled political parties in the EU to adopt a negative communication approach, resulting in an inclination towards more extreme policy stances[5,6]. More generally, while for autocracies and emerging democracies, the use of digital media can be associated with an increase in information consumption and political participation[7], the effect are more detrimental for established democracies where an increased political polarization is observed[8,9].

In recent years, there has been a growing interest in studying the influence of the algorithmic layer that mediates online human interactions. These researches draws on a extensive literature on opinion dynamics and agent-based models[10] to examine, for example, the effects of recommendations on online polarization[11,12], how could recommender systems break down echo chambers[13], or the coupling between human cognitive bias and algorithms[14,15]. However, these studies face challenges in terms of empirical evidence due to limited access to large-scale and non-public data; the lack of transparency from corporations hampers the ability of researchers to provide convincing demonstrations of the effects of widely deployed recommenders. Consequently, the scientific community benefit from studies conducted internally by the corportion themselves. For instance, a study by Huszár et al[16] uncovered that Twitter's recommender systems amplify the reach of mainstream political right content over left content, in six out of seven examined countries. Furthermore, internal leaked documents, such as the revelations made by Frances Haugen[17], who whistleblew Facebook prioritizing profit maximization over the public good[6], also serve as valuable sources of information for researchers.

In this article, we aim to investigate the informational landscape depicted when information feeds are curated such as maximizing engagement. After having harvested unique behavioral data, we implement a recommender system, relying on engagement predictive models. We describe how the quest of engagement maximization tend to bias the visibility of content shared by one's friends, favoring content similar to what has been previously engaged with, while downplaying or suppressing

alternative perspectives.

As highlighted by the comment "Humans and algorithms work together — so study them together"[18] there is a pressing need for new researches investigating the collective behavior of humans and algorithms, the algorithmic stewardship of collective behaviors[19]. Our research tends to respond to this methodological imperative and contributes to the understanding of the human-machine interactions. The empirical grounding of our work enhances its practical relevance, allowing the development of evidence-based policies aimed at fostering more responsible social media platforms.

Methods

In this section, we outline the development of engagement predictive models and discuss how we leveraged such models to simulate timelines. Subsequently, we introduce the metrics used to analyse them.

1.1 Predicting engagement

The cornerstone task of predicting engagement on social media platforms, like Facebook or Twitter, relies on extensive behavioral data gathered on a global scale. Such harvested data are proprietary assets of the companies and not publicly available. Nonetheless, to encourage research in this area, Twitter partnered with ACM RecSys and made available large datasets —made of public information— for engagement prediction tasks in 2020[20] and 2021[21]. These datasets have provided valuable opportunities for academic researchers to explore the design and optimization of recommender systems. However, the sole focus of these challenges has been on improving prediction accuracy, rather than addressing the broader ethical implications of recommender systems and their large scale deployment.

1.1.1 Dataset

For our research, the datasets used in the ACM RecSys Challenges were no longuer available. Therefore, we compiled a training dataset through a data-donation program. Specifically, we developed a browser add-on that captured —among other things— Twitter timelines and potential engagement from users. With the participation of 1.6k volunteers, we collected over the first quarter of 2023 more than 2.5 million tweet impressions. Among these impressions, we recorded 43k likes and 8k retweets. Although our training dataset is smaller than the one used in the challenges, we benefited from a lengthy collection period, a couple of months instead of a few weeks, and from having "true" negative samples, i.e. tweets being read but not engaged with. Indeed during the challenges, to release only public information, Twitter published so-called "pseudo-negative" samples, tweets authored by a user's friends, not engaged by the user, without restriction on whether or not the users saw it.

Furthermore, we leveraged additional information not provided during the challenges, such as historical features associated with each Twitter account, e.g. the number of tweets liked since account creation, and a large follow graph. Twitter leverage such follow graph, enriched with interaction networks, in its recommendation pipelines[15], we then seek to improve the prediction of our models by integrating such social informations. The follow graph was constructed through a broad snowball process, starting with seed accounts from participant friends, popular and random accounts, and users selected for this study. After multiple iterations, we obtained the last 5k followers and followees of over half a million accounts, resulting in a large graph with 260 million nodes and 700 million edges. After pruning the graph by removing poorly connected nodes (degree less than 50), we embed the follow graph through node2vec[22]. This algorithm assigns low-dimensional embeddings to nodes that maximize the likelihood of preserving neighborhood relationships, through biased random walks. We further emphasized homophily in the graph exploration by favoring a breadth-first sampling approach (p=1, q=0.25). Homophily should be understood in the sense of common friends, that can to some extent be associated to common interests. We corroborate this point in Supplementary Information, computing the Jaccard index between pairs of accounts' friends and the euclidean distance between them in the latent space, as well as a clustering and semantic analysis. Additionally, for visualization and computational purposes, we performed a dimensional reduction from the 64-dimensional node2vec embeddings to 2 dimensions using PaCMAP[23], a technique that preserves both local and global structure. The resulting latent follow space is depicted in Figure 2 (an annotated version is depicted in Figure SI.17).

1.1.2 Training

To predict user engagement with tweets, we use the collected behavioral data and trained our own models. It is important to note that our objective was not to propose a new model claiming superiority over the current state-of-the-art, but rather to develop a model that predicts engagement well enough for the present demonstration. We provide further details regarding the training process and a comprehensive list of features used by the models in Supplementary Information.

Following insights from the ACM RecSys Challenges[24], we employed gradient-boosting machines, namely LightGBM[25], and trained it on a parsimonious set of features. Among the features, several emerged as particularly important during the

training, these include: the time delay between tweet publication and its impression, the average word length in the tweet, the distance in the follow space —and the PageRank ratio— between the author of the tweet to be recommended and the last authors liked and retweeted by the user and finally the Jaccard index between the user's friends and the tweet's author's friends.

To facilitate scalability and enable extensive simulations, we did not incorporate semantic aspects of the tweets. Experimental results indicate that considering semantic information only marginally improved the model's accuracy while significantly increasing the computational costs. Overall, our model achieved an average precision (summarizing the precision-recall curve) of 71.3% for predicting likes and 88.4% for predicting retweets. Compared to a naive baseline, our model demonstrated a cross-entropy gain of 30.5% for likes and 63.5% for retweets. For reference, the winning implementation[26] of 2021 ACM RecSys Challenge, outperformed the naive baseline by 23.6% for likes and 29.5% for retweets. However, a straightforward comparison is not meaningful in our case, as our implementation leverages crucial features that were not accessible during the challenge.

The set of features used for training the models aimed to be both parsimonious and generalizable, considering for example the distance between authors in the lattent space instead of their absolute positions. To evaluate the models' generalization ability, we drew upon the findings of Barbiero et al.[27] and computed the convex hull of the training dataset in the feature space. Notably, 88.9% of the testing dataset samples fell outside the convex hull of the training set, necessitating the model to extrapolate, rather than interpolate, in order to make predictions[27]. Subsequently, we calculated the $F1$ score for the testing points based on whether they fell within ($F1_{in}$) or outside ($F1_{out}$) the training convex hull. The results revealed that the models predict with very high accuracy engagement for points falling inside the convex hulls ($F1_{in} = 0.80$ (likes) and $F1_{in} = 1.0$ (retweets)) and are reasonably effective extrapolating to points outside of it ($F1_{in} = 0.65$ (likes) and $F1_{in} = 0.90$ (retweets)), indicating successful generalization in predicting engagement even for samples distant from the training data. Furthermore, we computed the cosine similarity in the feature space between the samples in the simulated timelines (described bellow) and the closest vectors in the training set, yielding an average cosine similarity of $0.95_{[0.87,0.99]}$. In summary, our models successfully learned to generalize from the training samples, and the sample from the simulate timelines are similar to training samples.

The models' ability to generalize, coupled with our approach of utilizing non-public information during training while exclusively deriving input features from publicly available data, allows us to estimate the likelihood of engagement (likes and retweets) for a broad set of Twitter account —without being limited to participants who installed our data collection browser add-on— enabling us to thoroughly evaluate the impact of the recommender system.

1.2 Timelines simulation

We opted to simulate timelines for a set $\mathscr{A}$ composed of $|\mathscr{A}| = 2539$ public Twitter accounts. In order to do so, we randomly picked accounts that had either retweeted or had been retweeted in political-related tweets at least five times in the year 2022. The sampling process ensured an equal distribution across different political orientations, and we will explain in the following subsection how these orientations were determined. For each selected account, we gathered its likes and retweets history, as well as the list of accounts it followed.

To compile the corpus of tweets on which the simulation will rely on, we fetched the tweets published or retweeted by the 44k Twitter (public) accounts followed by at least 15 users in $\mathscr{A}$. Also, we enrich our corpus by 51k additional accounts, randomly selected among users' friends. Overall, we collected the tweets published or retweeted by 96k accounts during a 30-day period $\mathscr{P}$: February 18, 2023, to March 19, 2023. This 96k accounts results in coverage of 71% of users' friends.

For each user from $\mathscr{A}$, we simulated the timelines as follows: the user would log in into Twitter at a randomly selected hour, following the empirical distribution (see SI Figure 6), on seven randomly selected days picked in $\mathscr{P}$, and view a set of L tweets recommended by the system. The recommender system initially created a pool of messages composed of all the tweets published or retweeted by the user's friends within the last T hours. Then, for each tweet in the pool, the probability of engagement, defined as the maximum between the probability of being liked and of being retweeted, is computed. Future work would explore the effects of weightings the different engagement signals[28].

We do not consider in this article the long-term feedback loops; irrespective of what is being impressed on the simulation, we always refers to the "true" engagement made by the users. The simulated timeline of one day does not affect the next one.

To gain a deeper understanding of the singular nature of the tweets deemed to optimize engagement, we evaluated a range of recommender systems. Each recommender system generates timelines consisting of a fraction ϕ of tweets maximizing

engagement and a fraction $1-\phi$ of the most recent tweets. This design allow us to investigate the emergence of potential issues and biases associated with the increasing maximization of engagement. While reverse-chronological timelines are not a definitive solution and present significant challenges, as later discussed, we used them as a baseline for engagement-based ranking due to their simplicity and widespread use. Additionally, as an alternative benchmark, we will diminish the accuracy of the models' predictions. Specifically, starting from the raw prediction made by the model we will add a Gaussian noise of zero mean and of increasing standard deviation. We will then analyze the variation of the undermentioned metrics based on the average precision, computed on the test data set, at each noise level. This setup will enable us to observe the emergence of distortion effects as the model's accuracy improves.

Finally, we implemented two simple heuristics to mitigate potential trivial distorting effects. The first heuristic ensured that a single account did not contribute to more than 10% of the timelines during a specific session. The second heuristic aimed to remove duplicated tweets, such as a tweet being retweeted by multiple friends. These heuristics, implemented on Twitter[29], prevent spamming a user from a single account or tweet.

The number of tweets a user would view during each session and the temporal window over which friends' tweets are aggregated to build the pool of available messages are the remaining parameters influencing timeline generation. We set the session length to $L = 30$ tweets, which corresponded to the average length observed through our browser add-on (see the empirical distribution on Figure 9). The aggregation window was set to $T = 18$ hours i.e. the recommender system picks from all tweets published and relayed by a user's friends within the 18-hour period prior to the user's session. This choice aimed to balance the size of the pool and the associated computational costs. The partial open-sourcing of Twitter's recommender system revealed a temporal attenuation mechanism favoring fresh content, with an half time of of 6 hours[30]. A sensitivity analysis of these two parameters has been performed see 1.4.1 and Supplementary Information .

1.3 Political Orientation

The estimation of political orientations leveraged the Politoscope database. This comprehensive database encompasses 705 million tweets related to French politics since 2016, making it a valuable resource for capturing the dynamics of political discourse.

To determine the political leanings of the Twitter accounts, we leveraged the 2022 retweet graph associated with tweets either sent by French political figures or containing French political keywords, see[31] for the details of the data collection procedures. We considered retweets network, as research has shown them to be a reliable signal of ideological alignment[31]. Then, we leverage node2vec[22] algorithm to produce nodes embedding capturing the graphs underlying structure. Subsequently, we compute the angular similarity in the 64-dimensional latent space between the three main French political figures: Jean-Luc Mélenchon (far-left), Emmanuel Macron (center), and Marine Le Pen (far-right), and each of the 1.2 million Twitter accounts that had repeatedly published or retweeted political content during the 2022 period (during which the French presidential and legislative elections occurred). Based on the resulting similarities, we assigned a numerical political leaning to each account, ranging from -1 to +1, the details of the computation are presented in Supplementary Information.

It is important to note that the French political landscape exhibits a circular nature, where anti-system activists bridge far-left and far-right militants[14] (see a 2D spatialized graph of retweets Figure SI.10). To account for this circularity, we implemented a periodic boundary condition at ± 1 in our numerical opinion estimates, and this will be considered in the bellow defined metrics. This approach ensures consistency and interpretability, where negative values indicate left-leaning accounts, positive values indicate right-leaning accounts, and supporters of the current French President Emmanuel Macron around zero. The numerical scale matches with both members of parliament political group, see Figure 11, and a clusters analysis[31].

1.4 Metrics

We introduce several metrics aiming to evaluate whether, and how, the recommender systems may skew the content published by users' friends.

Gini coefficient

First of all, we measured the Gini coefficient, which provides insights into the distribution of publication and impression occurrences among users' friends. For a given set of friends, we calculated the Gini, subject-wise, coefficients ($G(P)$ and $G(I)$) associated with the number of tweets published (P_i) and impressed in the timelines (I_i) by each friend i during the specified time frame. The Gini coefficient formula is defined as follows:

$$G(X) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|X_i - X_j|}{2n\sum_{i=1}^{n}X_i}$$

where X represents either the publication (P) or impression (I) counts. The Gini coefficient allows us to assess whether the recommender system amplifies inequality by prominently displaying tweets from a small subset of friends while neglecting others.

Convex Hull Volume ratio

While the Gini coefficients provide insights into the selection of authors by the recommender system, they treat all accounts as equally different. However, it is important to consider that some accounts may produce similar content. To evaluate the diversity of authors chosen for impression more finely, we examined their positions in the follow space. Specifically, we compared the volume of the convex hull, which represents the region encompassing the selected accounts, responsible for either 75% of the impressions or having authored 75% of the pool of messages. The analysis of the volumes ratio allows us to identify whether the recommender system tends to show to a user tweets from a limited region of the follow space, despite the user having subscribed to a potentially diverse set of accounts. By focusing on the top 75% of accounts, we filter out potential outliers and highlight the specific areas where the recommender system appears to concentrate its "attention".

The ratio of convex hull volumes depends of the dimension of the space in which they are computed into. The intrinsic value of the ratio is not of particular interest in this article, its evolution as ϕ evolves is; we will then provide the evolution of the ratio in multiple dimensions.

Latent Unexpectedness

Subsequently, we wonder if the recommender is selecting content authored by accounts situated in the region of past engagement. To assess this, we determine the convex hull C of the set of accounts that users from $\mathscr{A}$ engaged with, (indifferently through liking or retweeting) during the two weeks prior to the analyzed session (we are referring to real data, see section 1.2). Subsequently, we introduce the ratio ρ, which compares the proportion of impressed accounts within the convex hull of past engagement to the proportion of publishing friends within this convex hull. Additionally, we compute δ as the ratio between, the average distances between the impressed accounts and the convex hull boundaries ∂C, and the average distances between the publishing accounts and ∂C; restricting the computation to accounts falling outside of the convex hull C. The average distance to the boundary ∂C reflects the level of "unexpectedness" of a recommendation, as introduced by Li et al.[32].

$$\rho = \frac{\langle \alpha \in C \rangle_{\alpha \in timelines}}{\langle \alpha \in C \rangle_{\alpha \in publication}} \quad \delta = \frac{\langle d(\alpha, \partial C) \rangle_{\alpha \in timelines \setminus C}}{\langle d(\alpha, \partial C) \rangle_{\alpha \in publication \setminus C}}$$

While Li et al.[32] determined the convex hull of past engagement in the latent space associated with a heterogeneous information network, our approach focuses on the latent space of follow relationships. See Figure 2, a 2D representation of the follow space, as well as the convex hulls associated to a user.

Political landscape

Focusing on politics, we aim to evaluate the potential distortion of the political landscape between the content users have chosen to subscribe to and the content that is being impressed upon their timelines. To accomplish this, we employ the Wasserstein metric $W(P,I)$, which provides a measure of the discrepancy between the distributions of political opinions held by the authors of tweets published or relayed by users' friends and the distribution of impressed content. The Wasserstein distances quantifies the minimum cost of transforming one distribution into another, where the cost is measured by the amount of mass that needs to be moved and the distance traveled, and does not require the two distribution to have the same support, contrary to the Kullback–Leibler divergence for example. Additionally, we seek to grasp the directionality of the distortion through the metric $D(F,o_i)$, which quantifies the amount and intensity of content diverging ideologically from users' own opinion. The calculations for $W(P,I)$ and $D(F,o_i)$ are as follows:

$$W(P,I) = \int_{-1}^{+1} |P(x) - I(x)| dx \quad ; \quad D(F,o_i) = \int_{-1}^{+1} F(x)|x - o_i| dx$$

Here, o_i represents the opinion of user i. In practise, we bin the opinion space in quantiles, determined on the whole Politoscope dataset. Also, the periodicity at ± 1 associated with the opinion metric space is taken into account when computing both $W(P,I)$ and $D(F,o_i)$.

By comparing $D(P,o_i)$ and $D(I,o_i)$, we can assess whether the recommender reduces the exposure to divergent content thereby, confirming users' existing opinions. The computed metrics $W(P,I)$ and $D(F,o_i)$ are illustrated in Figure 1 for two sets of distributions.

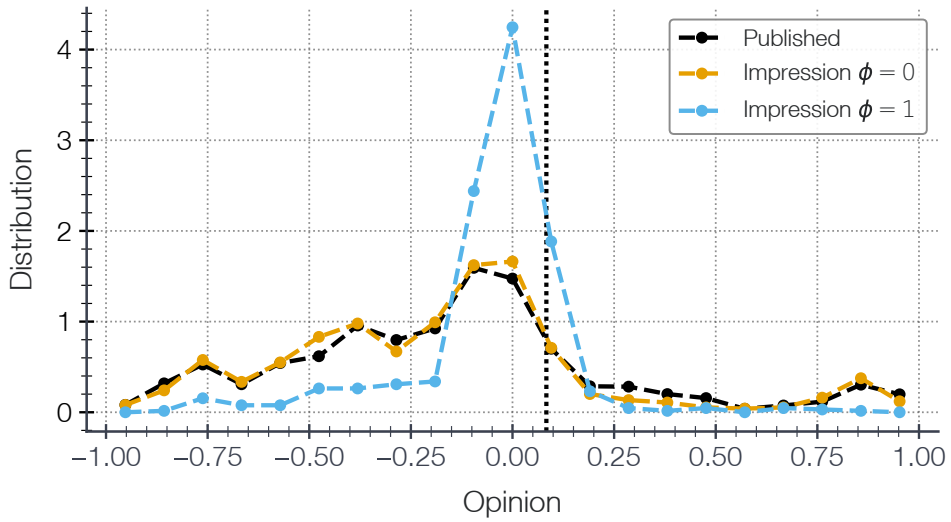


Figure 1. Example of deformation arising from the maximization of engagement in the timelines. The user has an opinion $o_i = 0.09$ (black dashed line), its friends published or relayed an amount of diverging content $D(P, o_i) = 0.38$. For reverse chronological timelines, $\phi = 0$, one has: $W(I,P) = 0.02$, $D(I, o_i) = 0.38$ whereas for maximizing-engagement timelines, $\phi = 1$, one has: $W(I,P) = 0.23$, $D(I, o_i) = 0.16$

Cost of fairness

Finally, we investigate the cost associated with maximizing engagement while ensuring, session-wise, that the distribution of political opinion remains identical, bin-wise, between the published content and the content displayed to users. We will also consider the distribution of distances to the previously engaged authors' convex hull boundaries. To evaluate the cost of these constraints, we compute the ratio between the average engagement probability of the selected tweets under the constraint and the optimal engagement probability achieved without any constraints. This provides a measure of the impact that these constraints have on the overall engagement of the recommended content.

Result

Overall, each of our users followed on average 1650 [193,4998] accounts. For each simulated session, the recommender had to pick messages within a pool large of, on average, 2826 [1586,5500] tweets. Among this pool, only 5.0 [1.1,13.1]% of them were deem likeable and 0.69 [.12,2.34]% retweetable by the model we trained (i.e. above the threshold maximizing F1 score).

Gini coefficient

On average, 465 [58,1195] authors account for 75% of the pool of message produced by a user set of friends, but only 85 [34,122] authors account for 75% of the impressions when timelines maximize engagement ($\phi = 1$); the Gini coefficient $G(I)$ is 28.3 % [27.9,28.7] higher than $G(P)$ (95% interval determined via bootstrapping over users). As displayed on Figure SI.12, the ratio $G(I)/G(P)$ increases as the engagement is maximized, and when the predictions are getting more accurate (see Figure SI.15a). For reverse-chronological timelines $G(I)$ is 23.3 % [22.9,23.6] higher than $G(P)$, similarly when the level of noise is large enough to randomize the ranking completely, $G(I)$ is "only" 20.8 % [20.5,21.1] higher than $G(P)$.

Consistent with empirical findings on Twitter[33], we observed a negative correlation between the increase in inequality and the Gini coefficient $G(P)$ of friends' publication activity (Spearman's rank $\rho = -0.78, p < 10^{-9}$).

Convex Hull Volume ratio

The convex hull of the accounts responsible for 75% of the impression in the timelines has a volume 82.4 [81.4,83.3]% smaller than the volume of the convex hull of the accounts responsible for 75% of the production of tweets, in 2D. The recommender focus on a small subsection of the graph of follow, see an example on Figure 2. While the ratio of volume depends of the dimension of the latent space, the decay remains the same as ϕ increases, in the various dimension (two, three and eight dimensional space) we tested, see Figure SI.3. Similarly, as the the predictions get more accurate as the volumes ratio decreases, see Figure SI.15b.

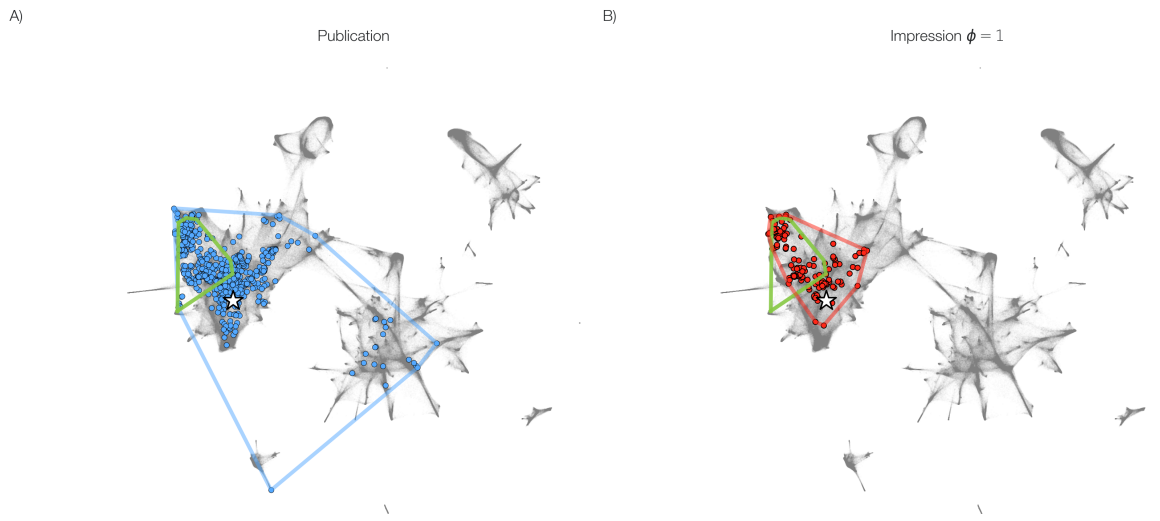


Figure 2. In grey: latent follow space (1.3M nodes). For an arbitrary user (displayed by the white star), we display (A) in blue the accounts having authored more than 75% of the pool of messages and (B) in red the accounts representing more than 75% of the impressions in the timelines, as well as the boundaries of the associated convex hulls. The boundary of the convex hull of the authors engaged by the user in the last two weeks, is displayed in green.

Latent Unexpectedness

As the timelines are made of tweets maximizing engagement (increasing ϕ), as the impressed authors either belong to, or are closer to, the convex hull of previously engaged authors, see Figure 3. In particular, among the authors impressed in the timelines maximizing engagement, the ones belonging to the convex hull of previously engaged authors are $\rho(\phi = 1) = 43.3\ \%_{\ [40.9,45.4]}$ more numerous, with respect to the friends having published (for reverse chronological timelines $\rho(\phi = 0) = 8.9\ \%_{\ [8.4,9.6]}$). Also, the average distance to the convex hull, for the author not belonging to it, decreases as the fraction ϕ of tweets maximizing engagement in the timelines increases. The impressed authors, not belonging to the hull, are $\delta(\phi = 1) = 40.0\ \%_{\ [38.5,41.5]}$ closer to its boundary than the average publishing accounts (for reverse chronological timelines $\delta(\phi = 0) = 8.2\ \%_{\ [7.0,9.5]}$). One obtains similar trends when considering the convex hulls in other dimensional space, see Figure SI.14.

Overall, maximizing engagement lead timelines with a reduced diversity of accounts; preferentially showing tweets from accounts belonging the region of the follow space within which the subject engaged in the past.

If one want to maximize the engagement under the constraint of faithfully depicting the diversity of account the user decided to follow, in terms of (signed) distance to the boundary of the convex hulls of past engagement, the average probability of engagement of the tweets impressed in the timelines would decrease by $-38.0\ \%_{\ [-38.9,-37.1]}$.

Political landscape

As displayed on Figure 13, as the timelines are made of tweets maximizing engagement instead of recent ones, as the distribution of political views from users friends is distorted between what is being published and what is being impressed. The Wasserstein metric $W(I,P)$ increases with ϕ; for $\phi = 0$ (reverse-chronological timelines) the Wasserstein metric is basically null, $W(I,P) = .048$ [.048,.049], the political landscape is faithfully depicted, for $\phi = 1$ one has $W(I,P) = .146$ [.143,.149]. The political landscape impressed on users timelines is such that the proportion of diverging content decreases with ϕ. While for reverse-chronological timelines, the proportion $D(I)$ of diverging content being impressed equals the one $D(P)$ being produced by a user's set of friends; when the timelines are made of tweets maximizing engagement, $D(I)$ decreases up to -29.7 [−30.6,−28.9]% for $\phi = 1$. Similarly, when adding noise, as the the predictions are accurate as both the distortion $W(I,P)$ increases and the amount of diverging content decreases, see Figure SI.15c.

Pairwise permutation tests on the average distortion $W(I,P)$ and confirmation $D(I)/D(P)$ fails to reject the hypothesis according to which these metrics are equals across subject's political affiliation.

If one want to maximize the engagement under the constraint to faithfully depict the political landscape the user decided to

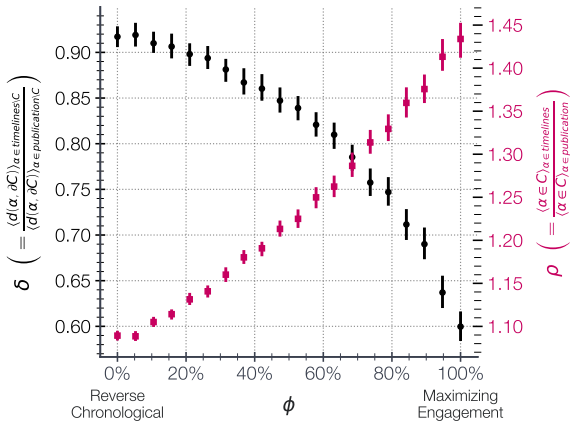


Figure 3. Latent unexpectedness δ (in black dots) and ρ (in red squares) with respect to ϕ (error bars represent 95% confidence intervals, determined via bootstrapping over users)

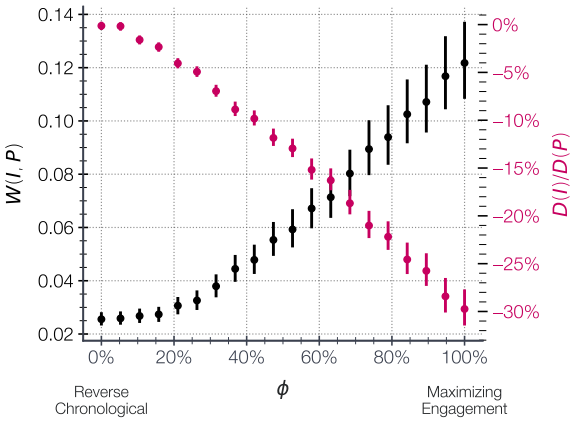


Figure 4. $W(I,P)$ (in black dots), and $D(I)/D(P)$ (in red squares) with respect to ϕ (error bars represent 95% confidence intervals, determined via bootstrapping over users).

follow i.e. enforcing the session-wise publication distribution of political opinion, the average probability of engagement of impressed tweets would decrease by $-10.7\ \%_{[-10.4,-11.0]}$.

1.4.1 Sensitivity Analysis

As displayed on Figure SI.7, as the aggregation window T increases, the timelines become more skewed. Conversely, increasing the session length L results in a reduction of the distortions, see Figure SI.8. These findings are in line with expectations. As the temporal window expands, the pool of available messages grows, affording the recommender system a larger latitude to choose engaging tweets, leading to the important distortion previously discussed. Conversely, as session length increases, more mundane tweets, with lower probability of engagement, are displayed; resulting in timelines where engaging tweets become diluted amidst a larger number of bland ones, ultimately attenuating the overall skewness.

1.5 Discussion

In this study, using unique behavioral data, we trained machine learning models to predict user engagement with tweets. Through simulation experiments, we examined the content that a recommender system seeking to maximize engagement would show to users. The predictive models uncovered the homophilic nature of human behavior. When these models are employed as recommender systems, they shape the reality perceived by users, acting as a distorted lenses that alter the perceived political landscape, limiting exposure to diverse content.

The investigation into the negative consequences of maximizing engagement was performed on Twitter-like timelines.

This choice was justified by the harvested behavior data and Twitter's recent openness to academic researches, which allowed for the utilization of additional features in the predictive models, such as the follow relationship graph. The findings may be nuanced for other platforms, considering their unique designs. For instance, Facebook, historically emphasized offline relationships like family and work, fostering cross-cutting ties[34]. In contrast, Twitter connections are primarily based on shared interests[35]. Although engagement maximization plays a crucial role, other heuristics are taken into account when generating Twitter timelines. The partial open-sourcing of Twitter recommenders has revealed the implementation of collaborative filtering strategies that group users by interest to proposed community-based recommendations[2,36]. However, regardless of the platform, when the content presented to users is curated by recommender systems trained on behavior data, the pursuit of engagement maximization would tends to reinforce human homophilic preference.

While analysing the effects of engagement maximisation through diversity measures, we do not express a normative stance on the value of cross-cutting exposure. A rich literature is painting a complex and nuanced picture of both online and offline effect of cross-cutting exposure[37–40], discussing this issue is well beyond the scope of the present article. Also, despite considering a range of recommender spanning from reverse chonological feeds to feeds maximizing engagement, we do not consider the former as the panacea for healthier online environment. Indeed, chronological feeds are recency-based ranking, predominantly rewarding individuals who post the most frequently. Such ranking fails to address the potential issue of overexposure to content from the same sources, as illustrated by the Gini coefficient associated to the authors impressions being 20% higher than the one associated to content production.

Consequently, a balanced and nuanced approach is necessary to foster healthier online environments, to bridge the gap between engagement signals and desired notions of value worth of optimization[41,42]. For example, Ovadya and Thorburn[43] introduced bridging-based ranking aiming to reward content that bridges divides, encouraging positive interactions across diverse audiences, even on contentious topics. The framework used in the present paper can be applied to implement and assess a variety of alternative recommender systems allowing evidence-based policies.

Future research will explore the consequences of long-term engagement maximization. We may hypothesize that in order to enhance long-term engagement, recommender systems may steer users towards regions where highly engaging content can be recommended in the future. This investigation may shed light on the mechanisms behind radicalization[44,45], and so-called "rabbit holes"[46], which are not captured in the present description, the maximization of engagement being session-wise.

Our research highlights that the algorithmic confirmation bias and the presence of echo chambers are not mere harmful byproducts of social media designs that one could patch. Rather, they emerge as direct outcomes of deploying increasingly accurate predictive models, trained on human behavior, with the goal of maximizing user engagement. By examining the objectives being optimized, our researches show actionable ways in which the evaluation of algorithms mediating social platforms could be improved.

Acknowledgments

The author deeply thanks Pedro Ramaciotti Morales and David Chavalarias for their precious insights and careful proofread. The author extends their sincere acknowledgments to Mazyiar Panahi, the head of Politoscope and Multivac platforms, for enabling the collection of the large-scale retweet network data that was instrumental in this research. Finally, the author acknowledges the Jean-Pierre Aguilar fellowship from the CFM Foundation for Research, the support and resources provided by the Complex Systems Institute of Paris Île-de-France and the Region Île-de-France.

Author contributions statement

The author confirms being the sole contributor of this work and has approved it for publication.

References

1. Hao, K. The facebook whistleblower says its algorithms are dangerous. here's why. (2022).
2. Saturluri, V. *et al.* SimClusters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, DOI: 10.1145/3394486.3403370 (ACM, 2020).
3. Zhao, Z. *et al.* Recommending what video to watch next. In *Proceedings of the 13th ACM Conference on Recommender Systems*, DOI: 10.1145/3298689.3346997 (ACM, 2019).
4. Mac, R. Engagement ranking boost, m.s.i., and more. (2021).

5. Morris, L. In poland's politics, a "social civil war" brewed as facebook rewarded online anger (2021).

6. Hagey, K. & Horwitz, J. Facebook tried to make its platform a healthier place. it got angrier instead. (2021).

7. Lorenz-Spreen, P., Oswald, L., Lewandowsky, S. & Hertwig, R. A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nat Hum Behav* **7**, 74–101, DOI: 10.1038/s41562-022-01460-1 (2022).

8. SCHAUB, M. & MORISI, D. Voter mobilisation in the echo chamber: Broadband internet and the rise of populism in europe. *Eur. J. Polit. Res.* **59**, 752–773, DOI: 10.1111/1475-6765.12373 (2020).

9. Allcott, H., Braghieri, L., Eichmeyer, S. & Gentzkow, M. The welfare effects of social media. *Am. Econ. Rev.* **110**, 629–676, DOI: 10.1257/aer.20190658 (2020).

10. Noorazar, H., Vixie, K. R., Talebanpour, A. & Hu, Y. From classical to modern opinion dynamics. *Int. J. Mod. Phys. C* **31**, 2050101, DOI: 10.1142/s0129183120501016 (2020).

11. Morales, P. R. & Cointet, J.-P. Auditing the effect of social network recommendations on polarization in geometrical ideological spaces. In *Fifteenth ACM Conference on Recommender Systems*, DOI: 10.1145/3460231.3478851 (ACM, 2021).

12. Donkers, T. & Ziegler, J. De-sounding echo chambers: Simulation-based analysis of polarization dynamics in social networks. DOI: 10.2139/ssrn.4437898 (2023).

13. Vendeville, A., Giovanidis, A., Papanastasiou, E. & Guedj, B. Opening up echo chambers via optimal content recommendation. In *Complex Networks and Their Applications XI*, 74–85, DOI: 10.1007/978-3-031-21127-0_7 (Springer International Publishing, 2023).

14. Chavalarias, D., Bouchaud, P. & Panahi, M. Can few lines of code change society ? beyond fack-checking and moderation : how recommender systems toxifies social networking sites (2023). 2303.15035.

15. Rossi, W. S., Polderman, J. W. & Frasca, P. The closed loop between opinion formation and personalized recommendations. *IEEE Trans. Control. Netw. Syst.* **9**, 1092–1103, DOI: 10.1109/tcns.2021.3105616 (2022).

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

17. Journal, T. W. S. The facebook files (2021).

18. Matias, J. N. Humans and algorithms work together — so study them together. *Nature* **617**, 248–251, DOI: 10.1038/d41586-023-01521-z (2023).

19. Bak-Coleman, J. B. *et al.* Stewardship of global collective behavior. *Proc. Natl. Acad. Sci. U.S.A.* **118**, DOI: 10.1073/pnas.2025764118 (2021).

20. Belli, L. *et al.* Privacy-aware recommender systems challenge on twitter's home timeline (2020). 2004.13715.

21. Belli, L. *et al.* The 2021 RecSys challenge dataset: Fairness is not optional. In *RecSysChallenge '21: Proceedings of the Recommender Systems Challenge 2021*, DOI: 10.1145/3487572.3487573 (ACM, 2021).

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

23. Wang, Y., Huang, H., Rudin, C. & Shaposhnik, Y. Understanding how dimension reduction tools work: An empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization. *J. Mach. Learn. Res.* **22**, 1–73 (2021).

24. Twitter. What twitter learned from the recsys 2020 challenge.

25. Ke, G. *et al.* Lightgbm: A highly efficient gradient boosting decision tree. *Adv. neural information processing systems* **30**, 3146–3154 (2017).

26. Deotte, C., Liu, B., Schifferer, B. & Titericz, G. GPU accelerated boosted trees and deep neural networks for better recommender systems. In *RecSysChallenge '21: Proceedings of the Recommender Systems Challenge 2021*, DOI: 10.1145/3487572.3487605 (ACM, 2021).

27. Barbiero, P., Squillero, G. & Tonda, A. Modeling generalization in machine learning: A methodological and computational study (2020). 2006.15680.

28. Milli, S., Pierson, E. & Garg, N. Choosing the right weights: Balancing value, strategy, and noise in recommender systems (2023). 2305.17428.

29. Twitter. Twitter's recommendation algorithm.

30. Twitter. The-algorithm/ranking.thrift.

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

32. Li, P. & Tuzhilin, A. Latent unexpected recommendations. *ACM Trans. Intell. Syst. Technol. Transactions on Intell. Syst. Technol.* **11**, 1–25, DOI: 10.1145/3404855 (2020).

33. Bouchaud, P., Chavalarias, D. & Panahi, M. Is Twitter's recommender biased ? An audit (2023). Preprint.

34. Mutz, D. C. & Mondak, J. J. The workplace as a context for cross-cutting political discourse. *The J. Polit.* **68**, 140–155, DOI: 10.1111/j.1468-2508.2006.00376.x (2006).

35. Conover, M. *et al.* Political polarization on twitter. *ICWSM* **5**, 89–96, DOI: 10.1609/icwsm.v5i1.14126 (2011).

36. Twitter. Twitter/the-algorithm: Source code for twitter's recommendation algorithm.

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

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

39. Min, S. J. & Wohn, D. Y. All the news that you don't like: Cross-cutting exposure and political participation in the age of social media. *Comput. Hum. Behav.* **83**, 24–31, DOI: 10.1016/j.chb.2018.01.015 (2018).

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

41. Ekstrand, M. D. & Willemsen, M. C. Behaviorism is not enough. In *Proceedings of the 10th ACM Conference on Recommender Systems*, DOI: 10.1145/2959100.2959179 (ACM, 2016).

42. Milli, S., Belli, L. & Hardt, M. From optimizing engagement to measuring value. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, DOI: 10.1145/3442188.3445933 (ACM, 2021).

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

44. Haroon, M. *et al.* Youtube, the great radicalizer? auditing and mitigating ideological biases in youtube recommendations (2022). 2203.10666.

45. Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A. F. & Meira, W. Auditing radicalization pathways on YouTube. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, DOI: 10.1145/3351095.3372879 (ACM, 2020).

46. Govers, J., Feldman, P., Dant, A. & Patros, P. Down the rabbit hole: Detecting online extremism, radicalisation, and politicised hate speech. *ACM Comput. Surv Comput. Surv.* DOI: 10.1145/3583067 (2023).

47. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, DOI: 10.1145/3292500.3330701 (ACM, 2019).

48. Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PLoS ONEONE* **9**, e98679, DOI: 10.1371/journal.pone.0098679 (2014).

49. Zhang, X. *et al.* Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations (2022). 2209.07562.

2 Supplementary Information

2.1 Training

To accelerate the training process, we implemented an undersampling of negative records, which refers to tweets that were viewed but not engaged with by a specific participant. We took careful measures to ensure the inclusion of negative records from the same session as positive ones. Additionally, we aimed to balance the representation of tweets from friends and non-friends, as well as from authors of engaged tweets. To enhance the diversity of our training dataset, we incorporated all impressions collected during Twitter searches, as these searches are expected to provide a more varied set of tweets compared to home timelines.

After an extensive exploratory data analysis, we designed sets of convoluted features and performed a systematic hyper-parameter tuning procedure. The hyperparameter tuning was conducted using equal time budgets and the Optuna "stepwise algorithm"[47], which efficiently explores the hyperparameter space. In particular, we explored the possibility of incorporating a wider history by considering the distance in the follower space between the authors of the tweet to be recommended and the last five authors that the users liked or retweeted. We also explored the integration of text embeddings, such as the cosine similarity between the tweet's text embedding and previously liked or retweeted tweets. However, we found that these additional features did not significantly improve the predictive power of the models compared to the computational cost involved. Future work focused on explainability rather than solely maximizing accuracy may consider incorporating these features for the additional insights they may provide. Finally, we settled for the following set of features as a trade-off between parsimony and predictive power

Table 1. Set of features

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

We provide in the Figure 5, the importance of each features in the models predicting the like and the retweet, in terms of number of times the features are used in the model.

2.2 Login hour

The analysis of Twitter usage patterns has been performed leveraging our data donation browser add-on. Our simulation incorporates the specific distribution for each day, see Figure 6. While we acknowledge that Twitter usage extends beyond browser platforms and includes mobile devices, leveraging this distribution provides the most precise and comprehensive representation available to us.

2.3 Political leaning assignation

As explained in the main text, the political orientations were estimated using the Politoscope database, which includes 705 million French political tweets since 2016. To enhance the reliability of the analysis, we only retained edges with a minimum

weight of 2. In other words, an edge was included between two accounts only if at least two political-related retweets had occurred between them in 2022.

Firstly, we set the opinion values of the far-left and far-right leaders to ± 0.75, chosen arbitrarily. Next, the opinion of the leader with a centrist leaning is determined by averaging the opinions of the two extreme leaders weighted by the angular similarities between the nodes' embeddings; determined via node2vec[22]. Interestingly, this calculation yields an opinion value close to zero (-0.02). For each Twitter account, we calculate the angular similarity between the account's embedding and the embeddings of the three leaders. The political leaning of the account is then determined as the average opinion of the two closest leaders, weighted by their angular similarities. In cases where the two closest leaders are the extreme ones, we account for the periodicity of the opinion space, ensuring that the assigned opinion spans the entire range from -1 to +1. Also, we only assign a political leaning if the angular similarity with the closest leader is at least 10% higher than the similarity with the farther away leader. Accounts without a clear political leaning, make up less than 10% of the accounts in our database.

We conducted experiments to ensure the stability of the resulting opinion scale when different anchors were chosen, including both the selected leaders and the arbitrary assigned opinions. Moreover, we observed a correspondence between the opinion scale and a cluster analysis of the retweet graph[31]. Additionally, the political groups declared by French members of Parliament align with the opinion scale, as shown in Figure 11. To visually represent the political landscape, Figure 10 presents a spatialized representation of the retweet graph, where nodes are color-coded according to their assigned numerical opinion, facilitating a clear interpretation of the political landscape.

2.4 Noisy predictions

To explore a more diverse range of recommender systems, we introduced noise into the predictions generated by our predictive models. Specifically, the recommender system would present the user with the top L tweets based on their engagement score, which corresponds to the probability of receiving a "like". Gaussian noise was then added to this engagement score. We focused on "likes" as they represent the most common form of engagement, allowing us to analyze the evolution of the metrics introduced based on average precision (evaluated on the test dataset), rather than the level of noise.

As illustrated in Figures 15a, 15b, and 15c, as the average precision increases, the distortion metrics also increase.

2.5 Graph of follow

In order to gain further insight into the structure captured by node2vec embeddings, we conducted an analysis involving half a million randomly selected accounts from the graph. We computed the pairwise Euclidean distance within three different dimensional spaces: the original 64-dimensional space produced by node2vec, as well as the reduced 8-dimensional and 2-dimensional spaces obtained through PaCMAP[23] reduction. The results, depicted in Figure 16, reveal that the pairwise distance of embeddings for accounts sharing a connection (i.e. at least one account follows the other) is significantly lower compared to pairs of accounts without such a connection. This behavior holds true across all three investigated dimensions.

Furthermore, we calculated the Jaccard index for each pair of accounts, measuring the similarity of the accounts they follow. Through the computation of the Spearman's rank correlation coefficient, we observed that the distance between two accounts in the embedding space is negatively correlated with the Jaccard index, the similarity of account they follow. This pattern persisted in the original node2vec space, with a Spearman's $\rho = -0.79$, as well as in the reduced dimensions with values of $\rho = -0.70$ (in 2D) and $\rho = -0.72$ (in 3D) (all $p-\text{values} < 10^{-16}$). These findings confirm that node2vec embeddings effectively capture the similarity between account neighborhoods.

The observed neighborhood similarity can be interpreted as homophily of interests, where accounts that produce similar content form communities in the graph of follow[2]. Figure 17 illustrates the latent space of followers in two dimensions, clustered using HDBSCAN[23], with the most popular accounts within the largest clusters labeled. Notably, we observe various interest-based communities related to sports, US politics, cryptocurrencies, gaming, etc. Also, Twitter accounts within each cluster tend to share the same language. The predominant language in account descriptions represents over 77.3% of the accounts description within a cluster, the second most popular language accounts for 17.9%; over the whole dataset the two most popular language accounts for 47.6% and 37.4% respectively of the accounts descriptions.

Lastly, we computed the cosine similarity between tweet text embeddings (computed by TwHIN-BERT[49], a language model for multilingual tweet representations) authored either by accounts within or outside a cluster. As depicted in Figure 18, the tweets authored by accounts within the same cluster are more similar to each other than to those authored by accounts outside the cluster.

Overall, the embedding of the graph of follow effectively captures the similarities and dissimilarities between accounts neighborhoods, embedding closely accounts that follow similar sets of accounts. To some extent, these homophily of

neighborhood can be linked to homophily of interest. As presented in[2], the user-user graph of follow is the foundation of Twitter SimClusters community-based recommendation algorithms.

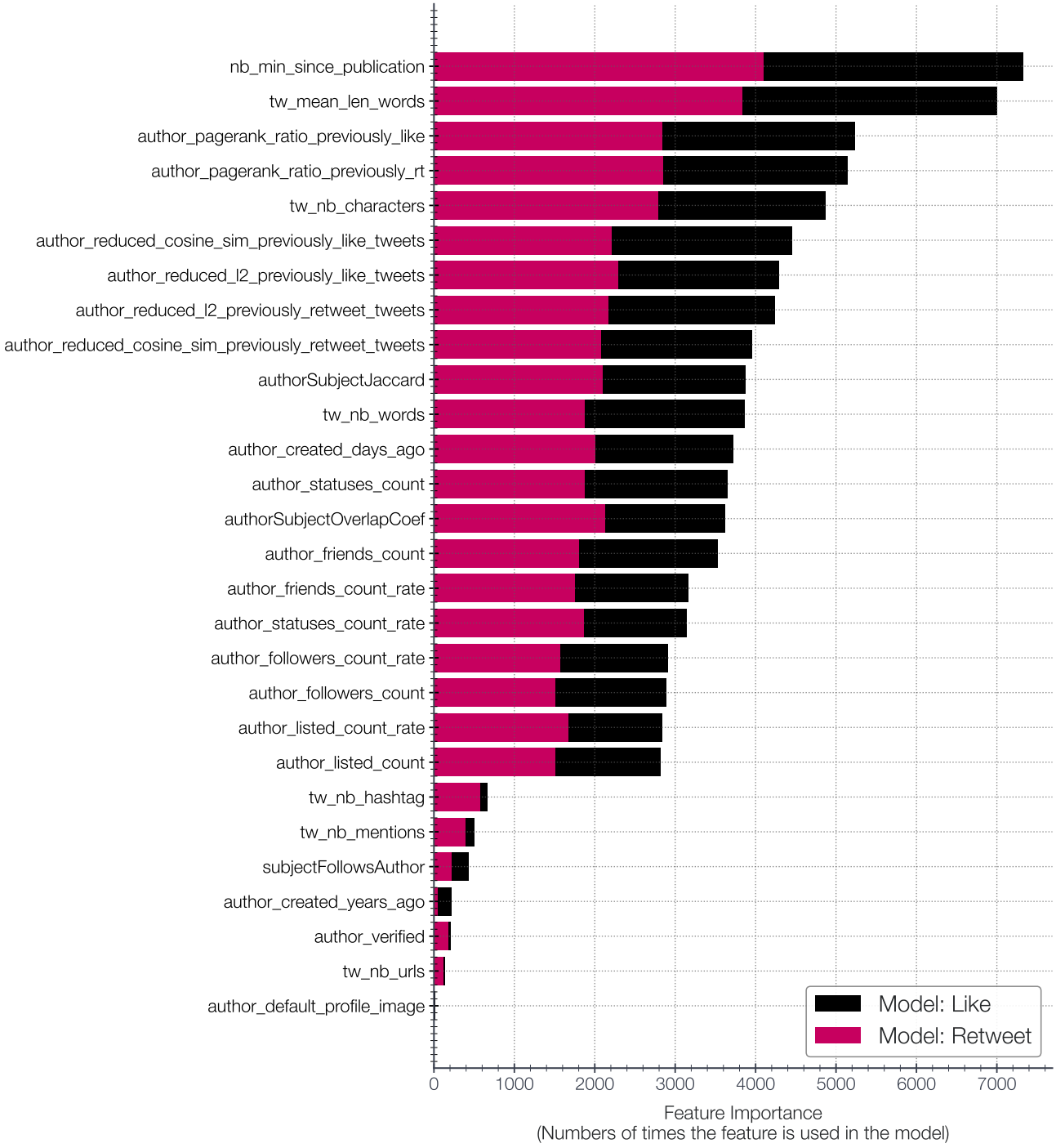


Figure 5. Features importance

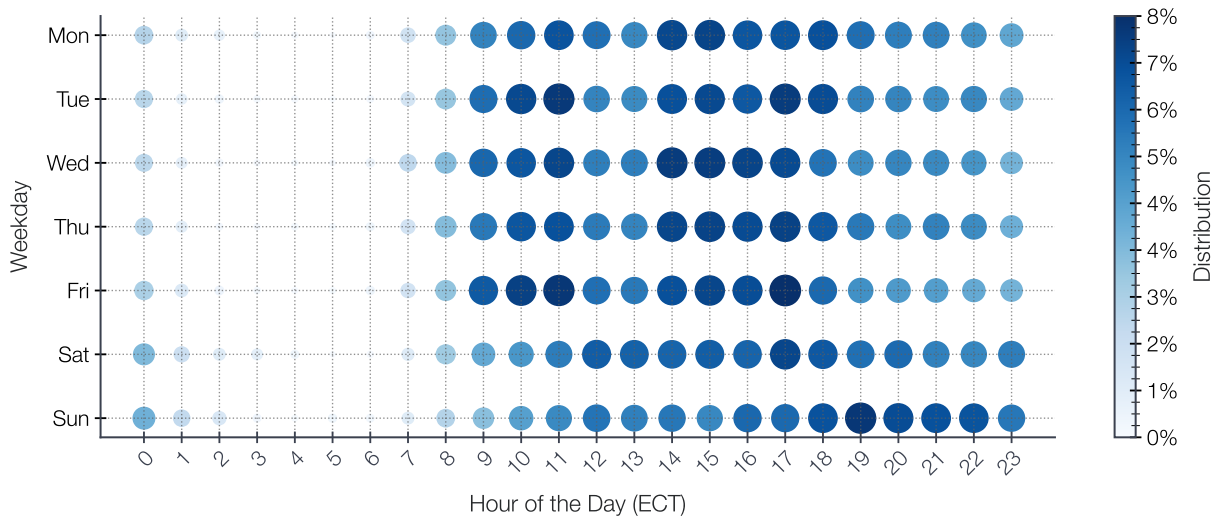
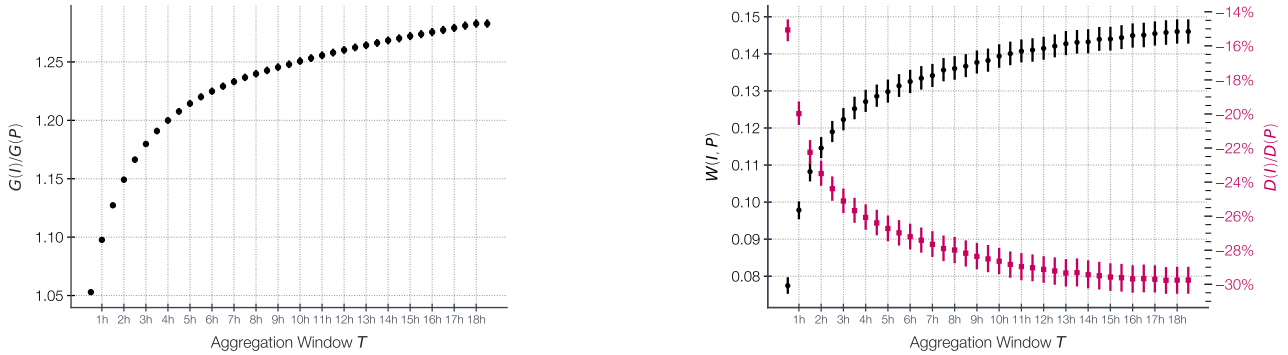


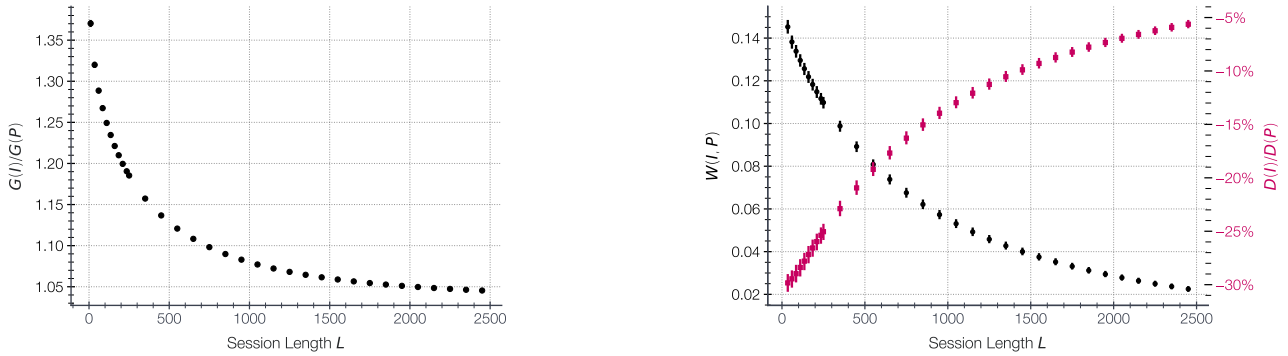
Figure 6. Distribution of hour of the day at which users log-in Twitter, as captured by our data donation browser add-on.



(a) Ratio $G(I)/G(P)$

(b) $W(I,P)$ and $D(I)/D(P)$

Figure 7. Metrics as a function of the aggregation window T, (error bars represent 95% confidence intervals, determined via bootstrapping over users)



(a) Ratio $G(I)/G(P)$

(b) $W(I,P)$ and $D(I)/D(P)$

Figure 8. Metrics as a function of the session length L, (error bars represent 95% confidence intervals, determined via bootstrapping over users).

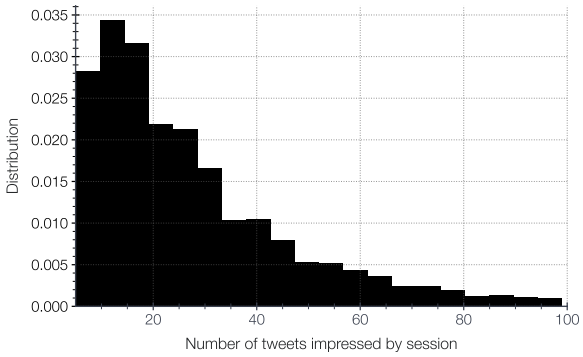


Figure 9. Distribution of session length (number of impressed tweets) as measured by our browser add-on. We truncated the distribution at 100 tweets, 2.6% of the session are in the truncated tails.

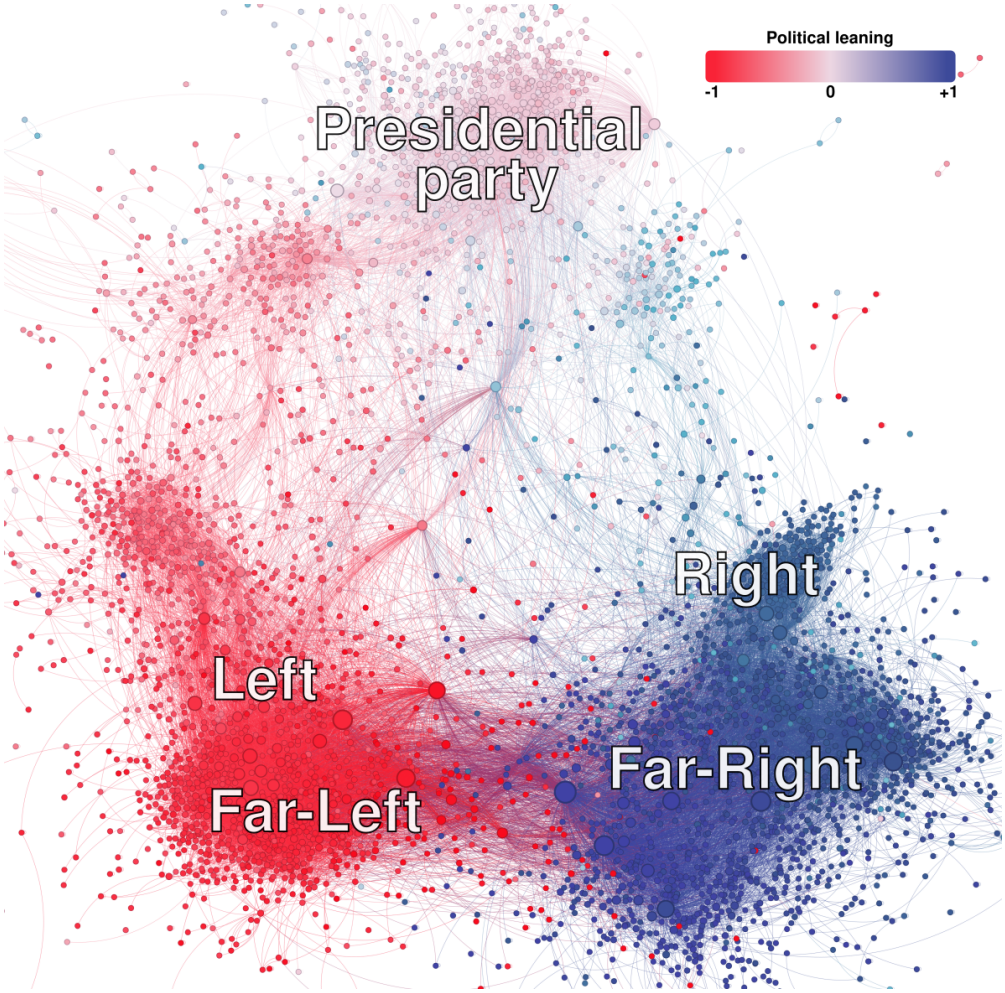


Figure 10. Graph of retweets associated to political messages published during the first semester of 2023, filtering edge with a weight less than 5. Spatialized via ForceAtlas2[48]. Nodes are colored by the assigned numerical opinion.

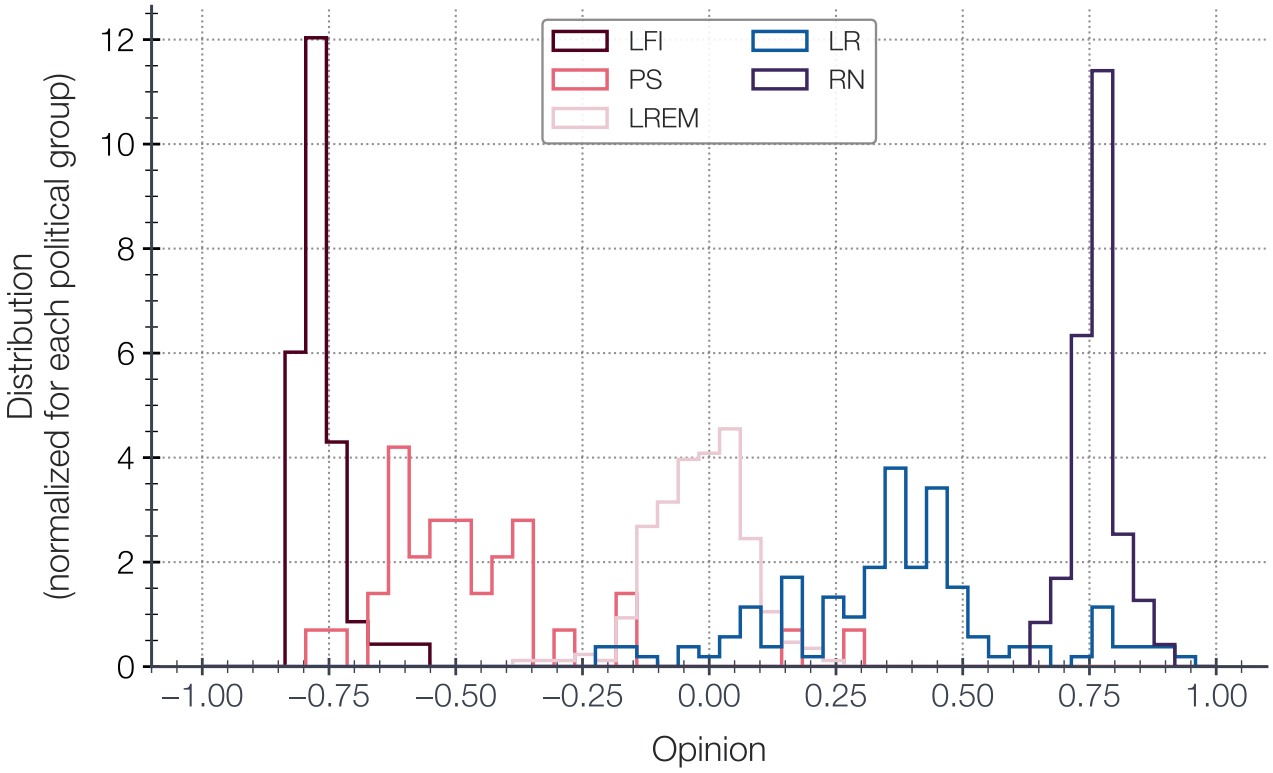


Figure 11. Distribution of the assigned numerical opinion for the member of French Parliament depending of their declared political group. Main french political groups from left to right: "LFI", "PS","LREM", "LR", "RN"

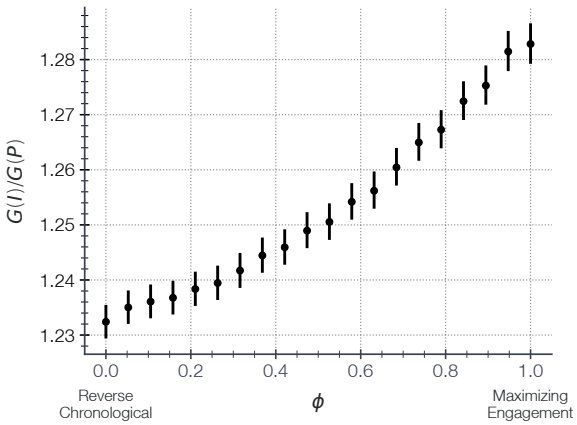


Figure 12. Ratio $G(I)/G(P)$ between the Gini coefficient associated to friends impressions $G(I)$ and publication $G(P)$ with respect to ϕ (error bars represent 95% confidence intervals, determined via bootstrapping over users).

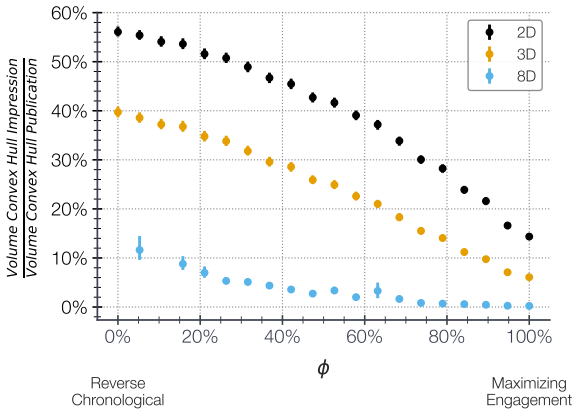


Figure 13. Ratio between the volume of the convex hull of the accounts responsible for 75% of the impression in the timelines and the volume of the convex hull of the accounts responsible for 75% of the production of tweets, in 2, 3 and 8 dimensional spaces (error bars represent 95% confidence intervals, determined via bootstrapping over users).

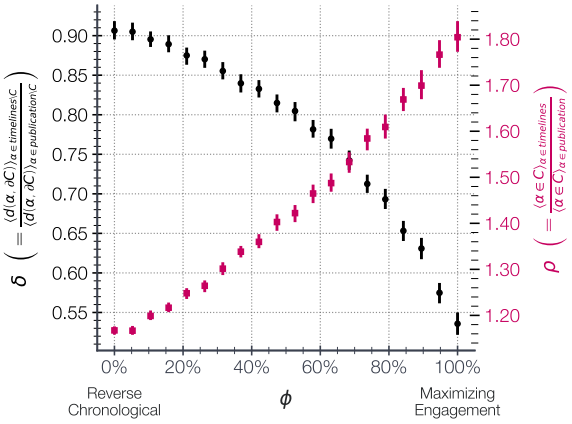
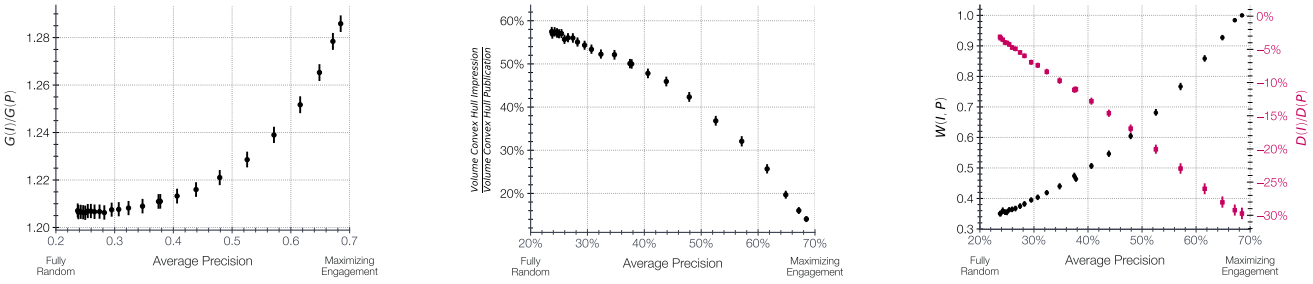


Figure 14. Latent unexpectedness δ (in black dots) and ρ (in red squares) with respect to ϕ, in 3D, (error bars represent 95% confidence intervals, determined via bootstrapping over users)



(a) Ratio $G(I)/G(P)$

(b) Convex hull volumes ratio

(c) Distortion of the political landscape

Figure 15. Metrics as a function of the average precision of the "like" predictive model, (error bars represent 95% confidence intervals, determined via bootstrapping over users)

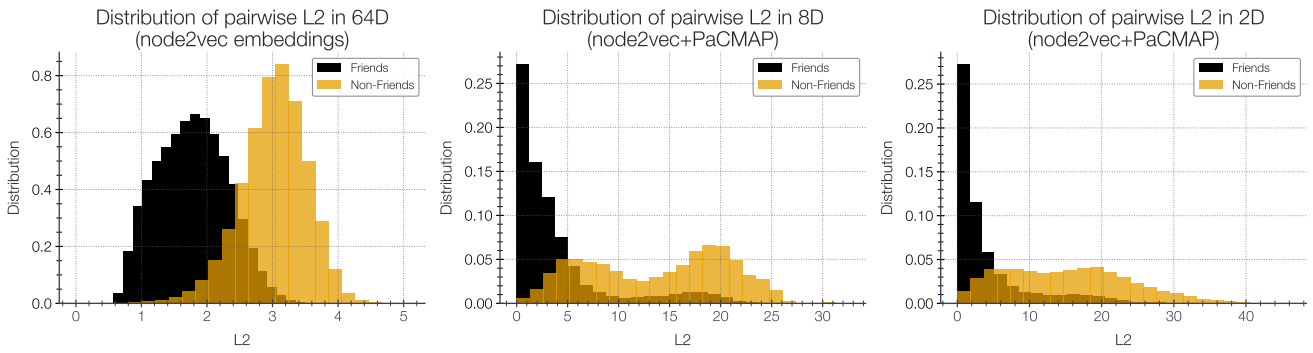


Figure 16. Distribution of the euclidean distance between random pairs of accounts, segmenting pairs in which at least one account follow the other.

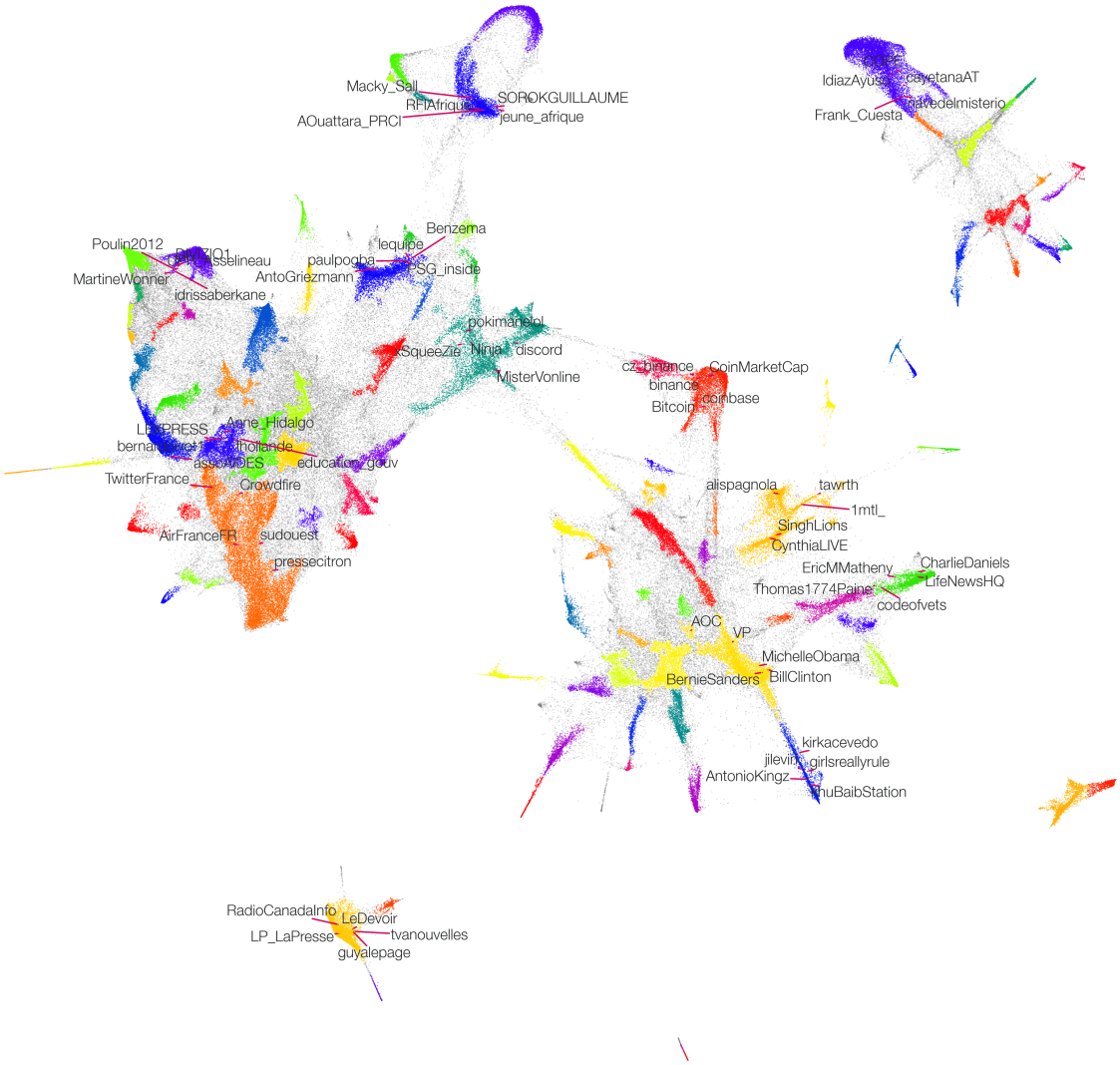


Figure 17. Latent follow space, clustered using HDSBSCAN. In grey are displayed outliers, belonging to no identified clusters. For each of the 16 larger cluster we label the most popular accounts.

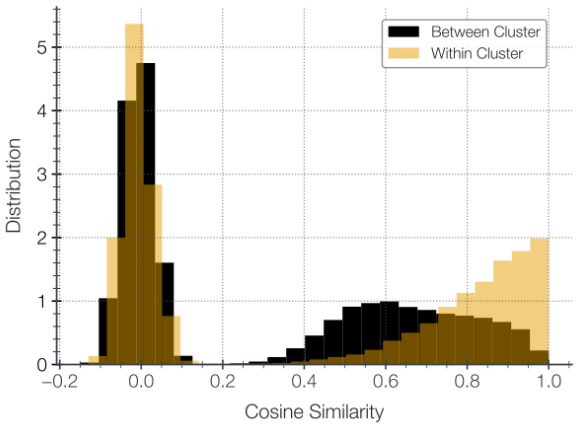


Figure 18. Distribution of the cosine similiarty between the TwHIN-BERT embeddings of tweets published by members of either the same cluster or different clusters.