



Recommender system in X inadvertently profiles ideological positions of users

Paul Bouchaud, Pedro Ramaciotti Morales

► To cite this version:

Paul Bouchaud, Pedro Ramaciotti Morales. Recommender system in X inadvertently profiles ideological positions of users. 2026. <hal-05468493>

HAL Id: hal-05468493

<https://hal.science/hal-05468493v1>

Preprint submitted on 20 Jan 2026

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY-NC-SA 4.0 - Attribution - Non-commercial use - ShareAlike - International License

Recommender system in X inadvertently profiles ideological positions of users

Paul Bouchaud^{1,2,†} & Pedro Ramaciotti^{1,2,3,‡}

¹Complex Systems Institute of Paris Ile-de-France CNRS, Paris, France.

²médialab, Sciences Po, Paris, France.

³Learning Planet Institute, CY Cergy Paris University, Paris, France.

† paul.bouchaud@cnrs.fr

‡ pedro.ramaciotti-morales@cnrs.fr

Abstract

Studies on recommendations in social media have mainly analyzed the quality of recommended items (e.g., their diversity or biases) and the impact of recommendation policies (e.g., in comparison with purely chronological policies). We use a data donation program, collecting more than 2.5 million friend recommendations made to 682 volunteers on X over a year, to study instead how real-world recommenders learn, represent and process political and social attributes of users inside the so-called *black boxes* of AI systems. Using publicly available knowledge on the architecture of the recommender, we inferred the positions of recommended users in its embedding space. Leveraging ideology scaling calibrated with political survey data, we analyzed the political position of users in our study (N=26,509 among volunteers and recommended contacts) among several attributes, including age and gender. Our results show that the platform’s recommender system produces a spatial ordering of users that is highly correlated with their Left-Right positions (Pearson $\rho=0.887$, p-value < 0.0001), and that cannot be explained by socio-demographic attributes. These results open new possibilities for studying the interaction between human and AI systems. They also raise important questions linked to the legal definition of algorithmic profiling in data privacy regulation by blurring the line between active and passive profiling. We explore new constrained recommendation methods enabled by our results, limiting the political information in the recommender as a potential tool for privacy compliance capable of preserving recommendation relevance.

Can recommender systems on social media inadvertently learn and leverage political preference of users? The answer to this question is relevant to data privacy regulation, when internal representations made by recommenders may map to traditional forms of measurement (such as political survey data), and for the study of interactions between individuals and AI systems. Recent studies have shown that machine learning procedures can autonomously learn sensitive attributes (1), and that social media data can be leveraged in estimating individual political preferences (2), ideological positions (3), and stances on issues (4). This body of work has nurtured the expectation that any recommender system capable of serving relevant content on social platforms must, at least to some degree, capture and leverage political preferences among several other attributes on which personal preferences may hinge. At the same time, the massification of social media and its entanglement with politics have motivated a growing body of work focused on political polarization and segregation, not least because of the role and potential consequences of AI systems computing recommendations on these platforms (5). A better understanding of the mechanisms through which AI systems learn, represent, and leverage ideological orientations on online platforms presents significant implications for developing regulatory instruments and compliance tools, and for studying social systems interacting with AI ones.

Most works studying recommender systems and politics investigate recommended items (e.g., content or contacts) either i) analyzing recommendations by measuring properties such as diversity and biases (6–9), or ii) assessing the causal impact of recommendations through controlled experiments with different recommendation policies (10–12). Several studies have addressed the question of machine representations of ideology from data during training, but mainly of contents (as opposed to individual actors), focusing on a few actors (13) (e.g., political elites such as Members of Parliament, comparing machine representations with measures such as ideology scaling of parliamentary voting), or on limited experimental settings, as opposed to large real-world systems with large user bases (14). Fewer studies have addressed how recommender systems in social platforms learn, represent, and process political information on large numbers of regular users and contents when computing recommendations (15). The main challenges in gaining further insights into how these AI systems work are the difficulty in using AI explainability methods in these real-world settings, and in collecting exposure data (6, 7, 10, 11, 16).

In this work, we develop and validate a method to infer an approximation of the embedding that X’s (previously Twitter) AI recommender system leverages in computing friend and content recommendations. X is a relevant platform for such a study because political figures actively use it (17) (making it a pervasive tool in elections (18)), because of its entanglement with the media sphere (19), and because

users cannot create private or semi-private communication spheres, casting a wide scope for recommendation. First, we collected more than 2.5 million real “Who to follow” friend recommendations shown on the platform via a browser plugin installed by 682 users during a data donation campaign, between January 2023 to May 2024 in France (a country with high percentage of usage of X, which is also covered by the EU’s General Data Protection Regulation, GDPR, forbidding the processing of sensible data categories without consent, including data on political opinions), involving more than 26 thousand users. Second, we used publicly available knowledge about the platform’s AI system architecture to infer an approximation of the embedding computed by X’s recommender system pipeline and that drives recommendations on the platform (20). At its core, this AI system embeds users, posts, and advertisements in a high-dimensional space where inner product similarities are ranked to recommend these items to users (21). Using this explicit formulation of inner product and ranking operations, we propose and validate a method to infer the most probable embedding positions for all users in our study.

To examine the political attributes of users in this space, we rely on ideological scaling (22) of the follower network connecting them to political figures (3), which we calibrated using political survey data (4). We compute additional user attributes to inspect the embedding space, including estimated age and gender from users’ profile text and photos (23), and their interests in various domains (e.g., news, sports, or science) based on the contents of their posts. Using these attributes we inspect the embedding space of the recommender in the platform to explain computation of recommendations in terms of these attributes. Specifically, we seek to establish the following facts. 1) That a spatial direction in the embedding of the recommender positions users in a way that is highly and significantly correlated with their Left-Right positions. 2) That this spatial direction is not significantly aligned with other direction in the embedding that position users according to other attributes relevant for recommendation. 3) That the positions of users on the direction along which positions are correlated with Left-Right leanings has an impact in the computation of recommendations.

The existence of political profiles inadvertently learned by AI systems raises important questions for regulation whenever these profiles can be shown to be sufficiently similar to traditional forms of profiling, such as those resulting from self-positioning ideology questions in survey research. Several data privacy regulations forbid the processing of data that reveal political opinion without explicit consent from users. In the context of the expectation that recommenders in digital platforms must learn, represent and leverage political opinions (among other traits) in some form, we provide the first quantifiable measurement of this phenomenon for a platform counting hundreds of millions of users worldwide. We show that X’s recommender system produces a linear internal

representation of recommended contacts that correlates with their Left-Right leanings.

Our findings point to a growing inconsistency in the application of several data privacy regulations. Regulations in South Korea, Switzerland, Brazil, and the EU forbid the processing of political opinions without explicit consent, and newer EU regulations such as the Digital Services Act (DSA) build on previous data privacy ones. Our work suggests that the distinction between active and passive profiling may come into question when AI systems process data produced by large numbers of regular individuals and when AI explainability methods such as ours may make visible (to platforms, regulators, and regular users) how this processing maps to the treatment of uncontroversial definitions of political opinion data. Our work also has implications for research on AI mediation of digital socio-information systems. First, it provides a way to further investigate the dynamics of human-AI systems with regard to political opinion processes and information exposure. Second, it enables new recommendation policies and algorithm design tools resulting from explicitly modifying the machine representations of political attributes. We explore the latter by investigating the potential consequences of complying with data privacy regulation, constraining the ideological leanings in the spatial representation of individuals in the recommender of X, measuring changes in content relevance and diversity.

Reconstructing the embedding space of the recommender in X

The “Who to follow” (WTF) recommendations (24) reportedly drive over one-eighth of all follower connections (25), and uses several algorithms to surface relevant accounts that are then ranked before recommending them (26). See Extended Figure S1 of the Supplementary Information document for an overview of the architecture of the service. X’s recommender system leverages explicit user signals—who they follow, repost (i.e., share), reply to, or click on—resulting in their representation in a multidimensional embedding space. Positions in this embedding space are then leveraged in computing similarities via inner product operations, producing candidates for recommendation which are then ranked and served to users. See Extended Figure S2 for an illustration of the data representation of entities that are embedded, according to X’s scientific publications (21). While the architecture and algorithms involved in the computation of this embedding are publicly known to some extent, the data used during training and the positions of entities in the embedding of the recommender are not. This poses a double problem for explainability: first because of the opacity of the model (i.e., the so-called “black box” problem), but also the lack of access to the model itself, which constitutes a major barrier for most independent scientific research on algorithmic

recommendation.

To collect data traces allowing us to infer their positions in the embedding space of the WTF service, we created a data donation program in which users could install a browser plug-in that relayed to us recommendations of friends and contents serviced to them by the platform. We advertised the browser plug-in on X and on local media outlets in the fall of 2022, offering volunteers who installed it a personalized analysis of their feed. No monetary incentive was offered to volunteers. From January 1st, 2023, to May 1st, 2024, we collected 2,549,008 personalized friend recommendations seen on the platform by 682 fully-compliant volunteers (set $\mathcal{V}$) i.e., individuals who opted into the data collection program, used the browser extension for at least one month, visited X to see at least 30 contact recommendations, and did not opt out of the study by uninstalling the plug-in or by later requesting their data to be deleted from our database. From over 105K accounts recommended during the observation period, we retain 26,338 accounts that were recommended to two or more volunteers (set $\mathcal{U}$) to define the population for our study as $\mathcal{E} = \mathcal{U} \cup \mathcal{V}$ ($|\mathcal{E}| = 26,509$), for which we have data related to recommendations.

Using the platform’s API, we also collected the follower network subtended by population $\mathcal{E}$, as well as profiles and posts produced by them. We cast the problem of approximating the positions of users in $\mathcal{E}$ in the embedding of the recommender as an optimization problem for the prediction of observed recommendations, constrained to the known architecture of the embedding used by the WTF service (21). Following X’s own design, we consider a 256-dimensional embedding space (27) and we solved the optimization problem (namely, estimating the embeddings that best explained observed data) with a gradient descent algorithm. We assess the quality of the inferred positions in the embedding by evaluating the accuracy with which the resulting recommender can predict observed WTF recommendations in our study. Holding out 10% of our data (including both recommended and not recommended users), our trained model yields an AUC-ROC value of 0.700. For comparison, three baseline approaches—recommending random second neighbors, second neighbors with the highest number of followers, or second neighbors followed by the most first neighbors—all perform below chance level (AUC of 0.467, 0.425, and 0.423 respectively). The probability of a random baseline—predicting edges either uniformly or weighted by frequencies observed in the training dataset—to equal or outperform our trained model is less than 10^{-5} .

When evaluating only positive test recommendations (i.e., friend recommendations observed in the study) against a pool of 100 randomly chosen non-recommended second-order friends, the model achieves a Precision@1 of 0.725 and a Precision@3 (the number of recommendations shown on the right-side panel on X) of 0.691.

We tested the robustness and reliability of the inferred po-

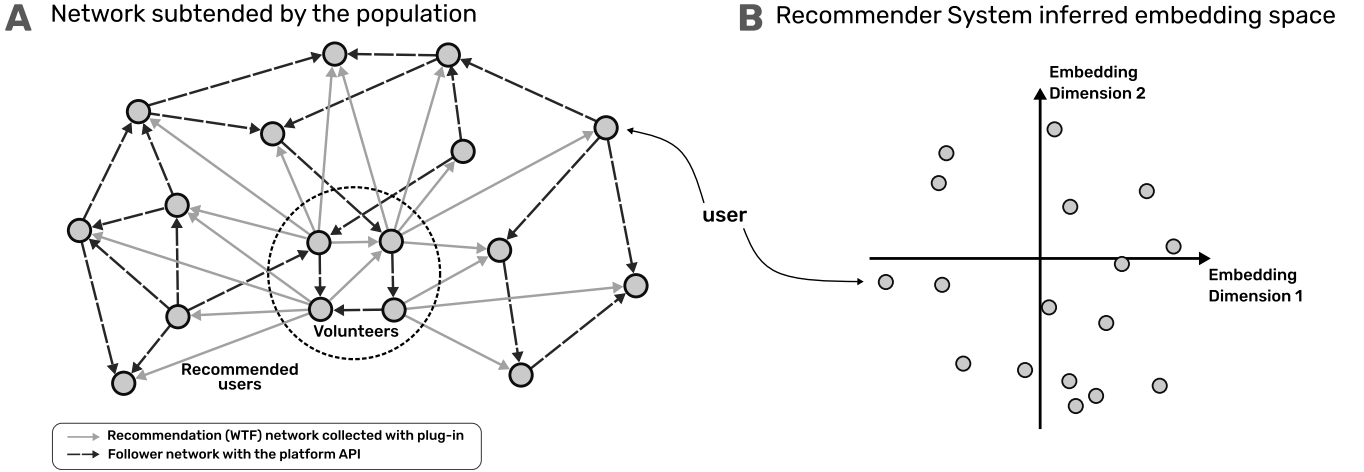


Figure 1: Schematic representation of the population of X users for the study and the inference of their position in the embedding space of X 's recommender system. (A) Population of the study (data volunteers $\mathcal{V}$ and recommended users $\mathcal{U}$) subtending networks for friend recommendation (in orange arrows) and following relations (in purple arrows), collected respectively with a data donation program using a browser plug-in and X 's API. (B) Schematic representation of the embedding space with every user i in the population $\mathcal{E}$ having a position ϕ_i in the 256-dimensional embedding.

sitions in the embedding for alternative training parameters and sub-sampling strategies simulating potential sources of biases, including: i) the possibility that volunteers saw recommendations on devices outside our study, ii) age, gender and ideological skews in our population, iii) changes in the activity and population of the platform during the period of observation (including changes around May 2023 related to change in platform ownership), iv) changes in the structure of the follower graph during the period of observation, and v) alternative parameters of the optimization problem. See the Methods section for a detailed description of the constrained optimization problem, its implementation, parametrization, validation and experiments demonstrating robustness and reliability.

Inferring political and demographic attributes of users

To consider political attributes of users in our study, we rely on a dataset of users in X with inferred political positions on a Left-Right scale in 2023 (28). The authors of this dataset use multidimensional ideology scaling of follower networks on X (3, 29) and political survey data (30) for calibrating positions of large populations of users pooled from followers of Members of Parliament (4). The dataset includes a Left-Right scale calibrated to have three reference points: 0 (the leftmost position for political parties), 10 (the rightmost positions), and 5 standing as the political center. The dataset includes a second political dimension called anti-elite rhetoric, measuring attitudes towards elites and institutions, deemed by authors of the dataset as a second relevant dimen-

sion in the country of our study (France). 8,249 users in this dataset are also in population $\mathcal{E}$ (i.e., 31.1%; see Figure 2.A). Extended Figure S14 & S15 show the positions of users in the dataset from (28) along the two dimensions and metrics assessing the quality of these calibrated positions.

We include additional user attributes in our study, both to control for potential covariates that might explain Left-Right positions of users encoded in the recommender system, and to further examine how these attributes are processed in the embedding space. To infer these attributes, we collected their profile bios (text and image), their followed accounts, and their last 200 published posts (either original or reposted), using the API of the platform. We infer the age and gender of users with the M3 model developed by Wang et al. (23), analyzing profile pictures, screen names, and biographies. The M3 model provides complementary probabilities (i.e., adding to 1) of being either “male” or “female”, allowing us to consider a Male-Female dimension ranging from the highest probability of being classified as male to the highest probability of being classified as female. The M3 model also provides complementary probabilities of being in age groups: 0-18, 19-29, 30-39, 40 and older (Extended Figures S11 & S12 show agreement of inferences using M3 by comparison with those achieved with other models and with attributes self-reported by volunteers).

We also consider the popularity of accounts, measured as the percentile rank of number of followers to account for long-tailed distributions (results are unchanged when considering the logarithm of the number of followers), and attributes capturing the interest that users have in different topics by computing a topic modeling classification of their last 200 posts, leveraging the model developed by Antypas et al. (31),

specifically designed for posts on (then) Twitter. Among these available topics, we specifically consider interest in news as a covariate that might explain political positions encoded by a recommender system.

Attributes in our population exhibit low absolute Spearman correlations (Figure 2.B), with the two highest being between age and interest in news ($\rho = 0.339$; i.e., older users tend to exhibit more interest in news content) and between age and negative attitudes towards elites and institutions ($\rho = -0.263$). Left-Right leaning (our main variable of interest) exhibits a lower absolute correlation with age ($\rho = 0.128$) and gender ($\rho = -0.079$). All the above-cited correlations have p-values < 0.001 .

Measuring attributes encoded by the recommender system

Next, we proceed to the identification and measurement of information on attributes of users that might be encoded in the embedding space. The main hypothesis behind this identification and measurement procedure is that these AI systems might develop representations that match relevant constructs, such as Left-Right ideological leanings. This expectation is motivated by the use of inner product similarities in ranking, based on positions on the embedding, on the part of the recommender.

To identify and measure the information of attributes of users encoded in the embedding, we leverage a search of spatial directions in this space and that induce high correlations with attributes, using Canonical Correlation Analysis (CCA), treating user embeddings as a multivariate dataset and inferred attributes as univariate ones. For each user attribute, a CCA identifies the spatial direction in the embedding space that maximizes correlation between the projected positions of users along the direction and the selected attribute. Because our goal to provide an explanation for the embedding on the semantic provided by attributes, we compute a CCA regression independently for each feature for interpretability. The CCA directions determined in this way, achieve Pearson correlations of different qualities for the selected attributes (see Figure 3.A; all correlations have p-value < 0.0001): Left-Right leaning ($\rho = 0.887$), anti-elite rhetoric ($\rho = 0.863$), interest in news ($\rho = 0.848$), popularity ($\rho = 0.730$), age ($\rho = 0.562$), and gender ($\rho = 0.384$). Extended Table S2 shows all Pearson correlations achieved by directions identified with the CCA for political and demographic attributes, as well as for attributes regarding interests in topics, and platform activity metrics. These inferred CCA directions represent the best possible linear encoding of the attributes of users learned and represented by the recommender system.

To further validate that these directions produce a reliable ordering of users by the value of their attributes, we assessed the quality of their positions projected onto these di-

rections, for Left-Right leaning ($\mathbf{w}_{LR}$), for anti-elite rhetoric ($\mathbf{w}_{AntiElite}$), for age ($\mathbf{w}_{age}$), for gender ($\mathbf{w}_{gender}$), for interest in news ($\mathbf{w}_{News}$), and for popularity ($\mathbf{w}_{pop}$). We assess the spatial ordering of users that follow Members of Parliament of parties with clear ideological leaning (using a Mann-Whitney U test): the center-leaning party *Renaissance* (RE), the Right-leaning party *Rassemblement National* (RN), and the Left-leaning party *La France Insoumise* (FI) (obtaining respectively 27.85%, 23.15% and 21.95% of the vote in the last presidential elections in 2022, followed by a fourth candidate obtaining 7.07% of the vote). Direction $\mathbf{w}_{LR}$ induces an order (Extended Figure S19) of users such that followers of MPs from FI are to the left of those of the followers of MPs from RE ($p < 0.0001$), which are to the left of those of the followers of MPs from RN ($p < 0.0001$). Direction $\mathbf{w}_{AntiElite}$ also induces an order of users (Extended Figure S20) such that followers of MPs from FI and RN (widely regarded as anti-elite parties (32)) are positioned to the right (i.e., displaying stronger anti-elite rhetoric) of followers of MPs from RE ($p < 0.0001$).

Measured by a Mann-Whitney U test, $\mathbf{w}_{gender}$ orders users classified male or female ($p < 0.0001$), and $\mathbf{w}_{age}$ produces a relative pair-wise ordering of age groups ($p < 0.0001$). Classification of users into age and gender groups is performed at a 0.5 probability threshold on the results of the M3 model. To further validate positions along these CCA directions, we computed the accounts followed, and the words and hashtags used by users at their extremes, and showed in the Methods section how they relate to corresponding ideological leaning, demographic group, or topic of interest.

Measuring the alignment between attributes encoded in the embedding space

Is the Left-Right spatial representation of users learned by the recommender system different from the representation of age or gender? To assess the similarity and alignment between positions of users along the inferred attribute directions in the embedding space, we inspect two metrics: 1) the cosine similarity (ranging from -1 to 1, with 0 meaning orthogonality) between each pair of CCA directions, and 2) the correlation between the positions of users along each pair of CCA directions capturing user attributes in the embedding space of the recommender.

CCA directions do not display a high degree of spatial alignment when measured via cosine similarity (see Figure 3.B). The highest (absolute) values for cosine similarity are between age and interest in news (0.304), and age and gender (-0.229), followed by the alignment between popularity and gender (-0.168). Only one pair of CCA directions displays a statistically significant alignment (at $\alpha = 0.01$) measured by the cosine similarity: age and interest in news. We measure the p-value as the proportion of randomized co-

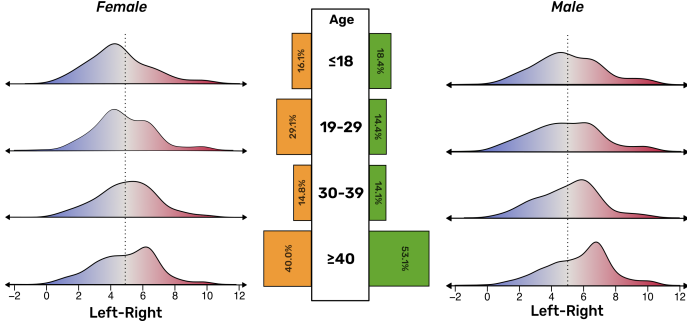
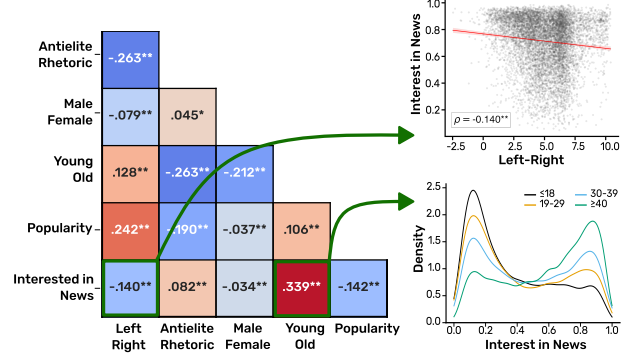
A Demographic and political profile of the population**B** Correlation between features in the population

Figure 2: Characterization of the population in the study. (A) Distribution of Left-Right ideological leaning of X users segmented by age and gender. (B) Pairwise Spearman correlations between the inferred attributes of users (* indicates p-value < 0.01, ** indicates p-value < 0.001). The correlation between Left-Right leaning and degree of interest in news is -0.140^{**} . The density plots for the degree of interest in news of users in different age groups illustrates a comparatively higher correlation between this attribute and age.

sine similarities (i.e., permuting the attributes of users) that exceed the observed similarity between two CCA directions.

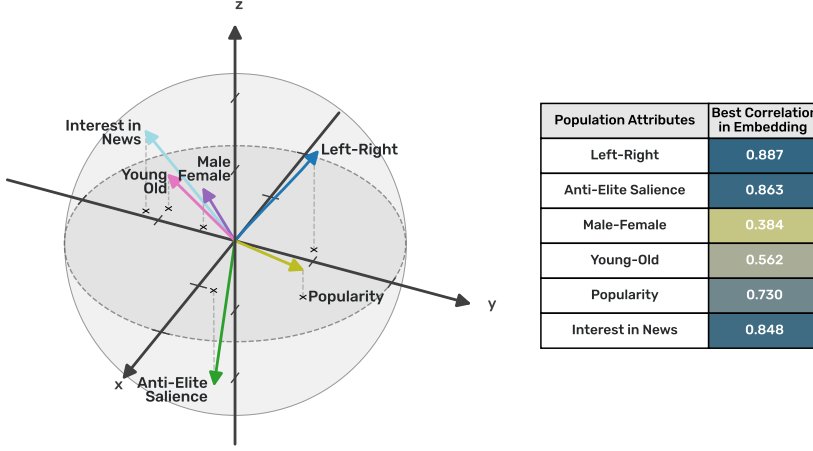
A measurement of the Spearman correlation between positions of users projected onto the CCA dimensions reveals different degrees of similarity among them. The three highest (absolute) correlations observed (all with p-value < 0.001) are for positions projected onto CCA directions: $\mathbf{w}_{age}$ and $\mathbf{w}_{news}$ ($\rho = 0.620$), $\mathbf{w}_{age}$ and $\mathbf{w}_{Anti-elite}$ ($\rho = -0.524$), and $\mathbf{w}_{Popularity}$ and $\mathbf{w}_{LR}$ ($\rho = 0.404$). Importantly, Left-Right leanings in the recommender (captured by $\mathbf{w}_{LR}$) do not induce high correlations with machine representations of age (as captured by $\mathbf{w}_{age}$; $\rho = 0.172$) nor gender (as captured by $\mathbf{w}_{age}$; $\rho = -0.275$) (see Extended Figure S30 for the spatial distribution along these pairs of directions). Extended Figure S29 shows the values of Spearman correlation between positions along all CCA directions.

These results show that the information linearly encoded by the recommender along spatial directions identified with CCA, are largely orthogonal. An attribute of special interest for both the study of online politics and data privacy regulation are Left-Right leanings. Direction $\mathbf{w}_{LR}$ that arranges users by Left-Right ideological leaning, induces projected positions that cannot be predicted with high accuracy by leveraging positions of users along other CCA directions capturing other attributes. In other words, up to linear encoding of attributes in the embedding of the recommender, information on ideological Left-Right leaning is nearly orthogonal from other attributes. While more complex geometrical models of encoded information on attributes might possibly be more explicative (e.g., nonlinear manifolds), our results show that, even considering a linear encoding, the recommender system can be proven to store information on the ideological leaning of users.

Constraining political profiling in recommendations

Even though political attribute processing was not specified in the design objectives of this AI system, our findings show that it does engage in such processing according to definitions contained in several data privacy regulations (e.g., EU’s GDPR forbids non-consenting processing that “reveals” political opinions). Our findings also point to potential ways in which this can be mitigated, by constraining the amount of information on political leanings of users that can be used in the computation of recommendations. The proposed procedure consists in computing the subspace in the embedding that is orthogonal to $\mathbf{w}_{LR}$ (i.e., a 255-dimensional hyper-plane in the embedding space), and projecting all recommendable entities onto it before using inner products in the resulting space to assess similarities and formulate recommendations (Figure 4.A). In the scope of our study, we propose an experiment recommending users on $\mathcal{E}$ to investigate the effects of such a procedure. The new embedding resulting from this procedure is *protected* (34) from having linearly encoded information about ideological leanings of users.

This proposed procedure holds resemblance to that described as “debiasing” of language models by Bolukbasi et al. (35), which allows zeroing-out the components of neutral words in the direction of a “gender subspace” (in their work) to mitigate gender stereotypes in word embeddings. As shown by Gonen and Goldberg (36) and by Ravfogel et al. (34), linearly-encoded attributes in high-dimensional spaces might be encoded in several orthogonal dimensions. This means that, even after projecting positions of users in the embedding onto the subspace that is orthogonal to $\mathbf{w}_{LR}$, in the new resulting embedding, it might be possible to find a second dimension that also encodes (with lower induced absolute correlation) ideological leanings. Therefore, as Rav-

A Directions in the embedding space correlated with attributes**B** Cosine similarities between correlated directions

Antielite Rhetoric	.089				
Male Female	-.117	.019			
Young Old	-.085	-.164	-.229		
Popularity	.059	.142	-.168	.137	
Interested in News	.028	-.033	-.052	.304*	.130
	Left Right	Antielite Rhetoric	Male Female	Young Old	Popularity

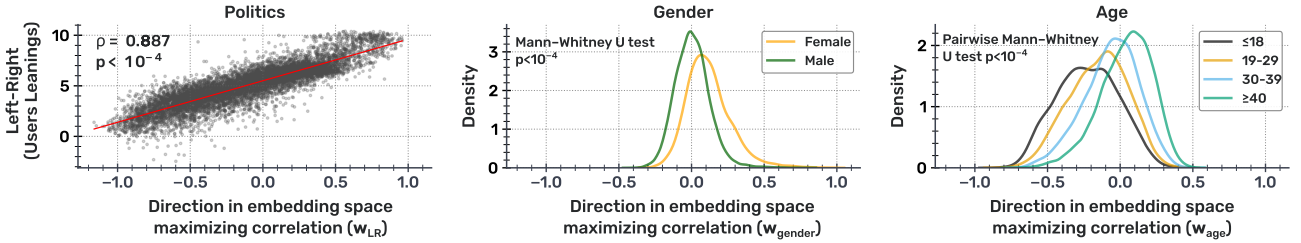
C Comparing population attributes with projection onto correlated direction in the embedding space

Figure 3: (A) Three-dimensional representation of directions in the 256-dimensional embedding space of the recommender system that maximize the correlation with attributes of users, identified using Canonical Correlation Analysis (CCA). The three-dimensional visualization was obtained through a Locally Preserving Projection (33). Vector norms of directions associated with attributes correspond to absolute correlations with these attributes, reported in the adjacent table (all p-values < 0.0001). (B) Absolute cosine similarity computed in the 256-dimensional embedding space between directions in the embedding space identified with CCA and that maximize correlations with attributes of users (* indicates p-value < 0.01). (C) Comparison of attributes of users along the CCA directions in the embedding that maximize correlations with three features: Left-Right leaning w_{LR} , age w_{age} , and gender w_{gender} .

fogel et al. (34), we repeat the identification and projection procedures, iteratively computing a CCA identification of directions encoding Left-Right leanings, until no subsequent identifications yield statistically-significant correlations (set at $\alpha=0.01$). This means that we iteratively remove linearly-encoded ideological information by projecting into new subspaces, until the p-value of the maximal correlation identified with CCA overflows above 0.01.

To evaluate the effect of our procedure we compare recommendations computed using X’s described pipeline based on embeddings (21), using both the original embedding and the new protected embedding void of linearly-encoded ideological leanings. We formulate 50 friend recommendations for each user and evaluate i) the Left-Right ideological diversity of recommended profiles (measured as the standard deviation of Left-Right positions) and ii) recommendation relevance. To assess relevance, we compare, through cosine similarity, the topic distributions of posts shared by friends suggested using the original embeddings versus those suggested using the protected embeddings. Additionally, we specifically examine the interest in news and politics among users suggested by the original versus protected embeddings, where interest is quantified as the proportion of each user’s last 200 posts classified as discussing news using the model by Antypas et al. (31). By selectively removing information that linearly encodes users’ Left-Right ideological leanings, we observe an increase of the Left-Right political diversity of recommendations (Cohen’s $d = 0.477$, 95% CI [0.459, 0.495] using the original embeddings as control. Confidence intervals were determined through user-level bootstrapping.) while maintaining relevance, as shown in Figure 4.B. Specifically, the intervention preserves the topics shared by suggested friends—the cosine similarity of topic distributions between friends suggested by protected and original embeddings is 0.948 (95% CI [0.751, 0.999])—and maintains similar interest in news between recommendation recipients and recommended friends (Cohen’s $d = -0.009$, 95% CI [-0.032, 0.014]).

In summary, our procedure demonstrates that linearly-encoded ideological profiles can be removed in our setting without significantly compromising (at least in the setting of our study) recommendation relevance, as measured by topic-specific metrics. However, the impact on recommendation diversity is significant. Section G.4 of the Supplementary Information examines the effects of our procedure applied to other attributes, showing that Left-Right leanings encoded in w_{LR} is the attribute that, when removed, leads to the most significant change in the set of recommended contacts. Recent literature reports mixed effects of diversified recommendations on political polarization, with some evidence suggesting potential exacerbation of affective polarization (37). These potentially adverse effects remain poorly understood across different national contexts and at varying levels of diversity intervention.

Discussion and conclusions

Our results show the recommender system in X inadvertently learns political orientations of users, meaning that these orientations are stored in the representation space of this AI system computing recommendations and that these orientations have an impact in the computation of recommendations. The encoding of information of political preferences may occur without users volunteering this information, without the encoding being part of the design of the recommender, and unbeknownst to platforms. This possibility has been recently suggested in previous works (15), and our results provide the first empirical evidence for real-world AI systems in a social media platform serving content to large numbers of users, in which AI data processing occurs at the unit of individuals (as opposed to only at the level of contents or posts), and for large numbers of regular individuals. We show that additional spatial directions encode other fundamental socio-demographic and thematic attributes, these directions being largely independent (i.e., not aligned in the embedding space). Our demographic and political geometry of a real-world recommender system representation also provides the first concrete example of the linear representation hypothesis (38–40) in an AI system serving hundreds of millions of users worldwide.

Our exposure data comes exclusively from feeds displayed on browsers. Our study does not account for differences, both in the specificity of the recommender and in the behavior of users, between navigation in the browser and through X’s telephone app.

We present additional experiments in the Methods section demonstrating the limited effects of this potential bias source, and explore alternative cohort compositions and embedding training strategies in Supplementary Information. Additionally, embedding reconstruction is largely informed by the follower network structure, whereas X’s recommendations incorporate additional signals such as user actions on posts (41) (e.g., liking and sharing). Yet, these limitations do not undermine our central conclusion: if selected subsets of input signals enable the recommender to learn ideological representations of users, the overall system retains this capability, at least locally. While additional signals may enable more sophisticated models, the system fundamentally possesses the capacity to learn ideological representations from specific input components. Similarly, parts of the recommender system architecture cannot be included in our study, such as the heuristics that modify recommendations after inner product similarity ranking, and that include diversification, moderation, content balancing, and content proofing (26). However, if the recommendations that we observe in our study still contain the measurable traces of spatial representations of political preferences in the embedding space (despite these post-ranking heuristics), the system as a whole holds again the property of being able to inadvertently learn political positions of users. Future research may leverage

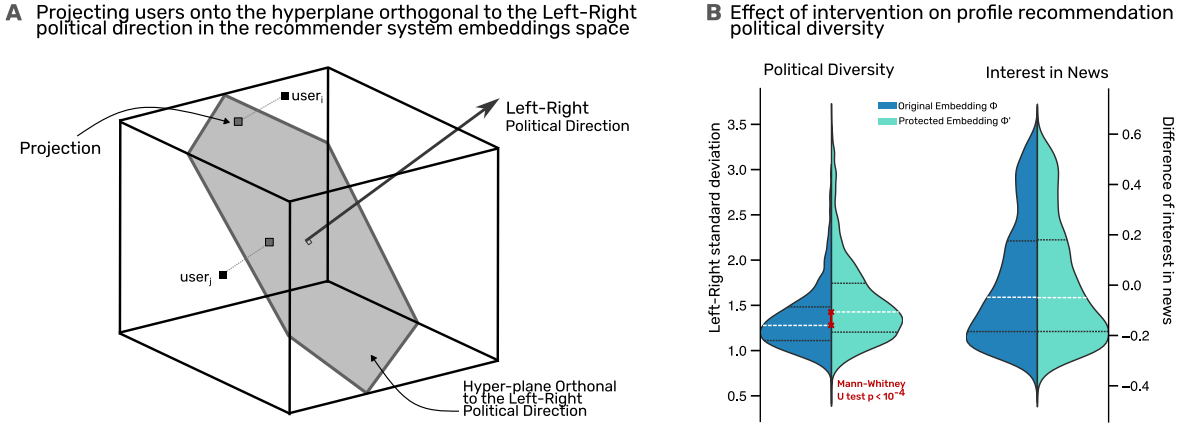


Figure 4: Constraining user recommendations from leveraging left-right political information. (A) Diagram of the intervention procedure. After identifying the direction in the latent space that maximally correlates with users’ left-right political leaning, we iteratively project user representations onto the orthogonal subspace. (B) Effects on recommendations of constraining political information in the embedding of the recommender systems. For each user in the population, we compute 50 friend recommendations before (in blue) and after (in green) the intervention and compute the impact the distribution of the ideological diversity of recommendations as the standard deviation on the Left-Right scale, and the relevance as the similarity in interest in news between recipient of recommendations and recommended friends.

public knowledge on other components of the recommender system, such as the content-based components. Building on our results, future studies might also address whether recommender systems learning of political preferences persists across different languages, national settings, platforms, time periods, and services.

Our results have relevant consequences for platform and data privacy regulation. Several data privacy regulations worldwide prohibit the processing of sensitive personal data, including political opinions, without explicit consent given by users (e.g., European GDPR Article 9.1, Brazilian LGPD Article 11, South Korean PIPA Article 23, Swiss nFADP Article 5, thus pertaining to the rights hundreds of millions of individuals globally). Our results show that social media platforms might inadvertently create and process political profiles as defined by several regulations by simply “revealing” (e.g., EU’s GDPR in Article 9) political opinions and using them in the computation of recommendations, even under uncontroversial definitions such as processing representations of Left-Right leaning for large user populations in scales of political survey instruments. Legal scholars have previously considered this possibility (42, 43), in which biases might be neither in the training data nor in the design of an AI system, but “autonomously developed by the algorithm” (44). Our results provide the first empirical evidence of this possibility. Our results show that user attributes are encoded in approximately orthogonal directions in the embedding space (only age and interest in news show statistically significant alignment) in a recommender system that ranks based on inner product similarity. Our method makes implicit computation of ideological leanings explicit, creating a challenge for privacy regulation by blurring the distinction between active

and passive profiling. If these inadvertently developed profiles fall within the scope of regulation, regulators must define thresholds for what constitutes political opinion—potentially making numerous AI systems non-compliant. If, on the other hand, these profiles fall outside regulatory scope, platforms could exploit AI explainability tools to target users based on political profiles while remaining compliant, effectively nullifying privacy protections in cases like political ad targeting. Even assuming high alignment between sensitive and non-protected attributes in the embedding, contradictions persist. If platforms can legitimately infer sensitive attributes through aligned non-protected attributes, made explicit through a method such as ours, they can select proxy attributes correlated with political positions, undermining regulatory intent. If, on the other hand, sensitive attributes must never be processed regardless of alignment with other attributes, definitions of what constitute political opinion offered by regulators may risk prohibiting several common machine learning practices.

Finally, our work also points to ways in which social platforms and AI operators might seek to mitigate these risks, by developing constrained recommendation policies—i.e., recommendations computed using embeddings in which information about a sensitive attribute has been removed. Our experiments suggest that constraining the political information available for computing recommendations, their relevance is marginally affected, while increasing the political diversity of recommendations.

Materials and Methods

Recommendation and survey data collected via browser plug-in

To collect recommendations on X, we created a browser plug-in (also called a browser extension) that unobtrusively collects “Who To Follow” (WTF) recommendations displayed to volunteer users that installed the extension. We recruited participants following a call for participation launched on mainstream media coverage (newspapers and radio) and on Twitter advertising in the fall of 2022. The study incentivized participation by offering volunteers personalized analytics on their Twitter feed’s political diversity. No monetary incentives were offered to participants. Participants provided informed consent after receiving detailed information about the study objectives and data collection scope. Upon installation of the plug-in, users were asked their gender, their age group, their self-assessed ideological positions on a 11 point scale, and the devices from which they connect to Twitter. See Section B.2.1 for additional details on the recruitment strategy, attrition, and the questions used in this brief survey. Survey responses are only used in our study to validate the inference of socio-demographic and ideological attributes. Because extensive literature has demonstrated the potential sources of biases involved in data donation (45–47), we conducted an assessment of these biases in user attributes in Section D, and an assessment of the impact of these biases on our main results in Section C.2.3 (in experiments with different population compositions) of the Supplementary Information.

From January 1st, 2023 to May 1st, 2024, we gathered over 2,549,008 WTF friend recommendations displayed to 682 unique to volunteers during the period, henceforth referred to as set $\mathcal{V}$. We retained only users who saw at least 30 Who To Follow recommendations and did not opt out of the experiment or request data removal after providing consent. We report in Supplementary Information demographic statistics of our cohort. Extended Figure S4 shows the daily rate of collection of recommendations from users that used our browser plug-in. Participation and frequency of use by volunteers fluctuated during the observation period. We observe an increase in participation around the time of policy changes in the platform linked with changing ownership.

Within the more than 2.5 million WTF recommendations collected, we observed over 105k unique accounts that were recommended, with a highly skewed distribution (see Extended Figure S5 and Extended Table S1 showing the most recommended accounts). 26,338 unique accounts were recommended to at least two volunteers. We will refer hereinafter to this set of users as $\mathcal{U}$, and we define $\mathcal{E} = \mathcal{U} \cup \mathcal{V}$ (users for which we have recommendation data) as the base population for our study, with $|\mathcal{E}| = 26,509$.

User data collected via the API of X

Additionally, we used the API access of X to collect data on users in $\mathcal{E}$, including: their friends or followed accounts (allowing us to establish the follower network subtended by $\mathcal{E}$), the last 200 posts they produced (including reposts), and information from their profiles, including text bio description, name and photo.

Inferring positions of users in the embedding space

The WTF service employs precomputed high-dimensional representations of candidate users as user features (26). X described in scientific articles how Heterogeneous Information Network (HIN) embedding was used to formulate personalized recommendations (21, 27). X’s HIN consists of various entities, such as users or posts, being linked through relationships such as “likes”, “authorship” or “follow” (see Extended Figure S2.A). From observed relationships, X positions entities and relations within an embedding space that is used in downstream tasks to compute rankings of potential entities to recommend.

We used the data collected for this study and publicly available knowledge on the architecture of the recommender and its training procedure (21) to define the problem of inferring the positions ϕ_i of users $u_i \in \mathcal{E}$ in the embedding space used by X’s WTF service. We define this problem as that of estimating the positions of users in the embedding that yield the highest probability of computing recommendations that have been proposed by X to users on the platform, and captured by our browser extension.

We consider the edges e that link a user u_s (recipient of recommendations, the *source* of the link), with all potential other users u_t (the *targets* of the link and that might be recommended to u_s). Embeddings of entities and links are related to similarities used in ranking through a scoring function $f(e) = f(\phi_s, \phi_r, \phi_t)$ for an edge e of type r (in our case “Follow” and “WTF” relations) in the HIN connecting source entity s to target entity t . The model optimizes entities’ embeddings (ϕ_s , ϕ_t , and ϕ_r) by maximizing the log-likelihood $\log \sigma(f(e))$ (with σ being the logistic sigmoid function) for observed edges while minimizing it for unobserved ones (i.e., relations that do not exist in the data). In (21, 27), X reports adopting the *TransE* (for *translating embeddings*) scoring function defined as the sum of inner products: $f(e) = f(s, r, t) = (\phi_s + \phi_r)^T \phi_t$.

To infer the positions of users from $\mathcal{E}$, we leverage the data available for inference (WTF recommendations and the follower network; see Extended Figure S2.B), and cast the problem as an optimization constrained to the system architecture described in X’s scientific publications (21, 27). Specifically, we consider the loss function prescribed by X for the training of the embedding on HIN data, and train

the embedding that minimizes loss for our WTF and follower network data. In a successfully trained model, the inner product between a target account’s embedding ϕ_t and, for instance, the sum of a source account’s embedding and the “Follow” relationship embedding ($\phi_s + \phi_{\text{Follow}}$) should yield high values for actual follower relationships and low values otherwise. To infer embeddings ϕ for users, we consider a convex combination of both relationships through the following loss function $\mathcal{L}$:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{WTF}} + (1 - \alpha) \mathcal{L}_{\text{Follow}}, \quad (1)$$

where each relationship-specific loss component is defined as:

$$\mathcal{L}_r = \sum_{e \in \mathcal{P}_r} \log \sigma(f(e)) + \sum_{e' \in \mathcal{N}_r} \log \sigma(-f(e')). \quad (2)$$

$\mathcal{P}_r$ and $\mathcal{N}_r$ denote the sets of positive and negative (i.e., not observed in our data) edges for a relationship type $r \in \{\text{“Follow”}, \text{“WTF”}\}$. Our formulation allows us to consider the inference of the positions of users embedding space by using only observed recommendations (i.e., setting $\alpha = 1$), but also exploring improvements on the inference by integrating knowledge on the architecture of the embedding, using subsets of the signals that the WTF use during training (i.e., follower networks, setting $\alpha < 1$).

Sampling of negative edges for the inference

Combining our two datasets, we obtain the set of positive edges for the “Follow” and “Who to Follow” relationships, resulting in $|\mathcal{P}_{\text{Follow}}| = 4,441,670$ “Follow” edges and $|\mathcal{P}_{\text{WTF}}| = 110,164$ “Who to Follow” edges. As third party, we lacked access to the exact set of candidate accounts scored by X’s algorithmic systems, from which the “Who to Follow” list is computed after ranking, we then relied on sampling of negative edges, i.e., edges e that are not observed in the data. Specifically, we generated negative edges by corrupting positive ones, replacing target accounts with randomly sampled accounts. This sampling strategy is aligned with the default setting of PyTorch-BigGraph, developed by Meta (48) and used by X in (21), and includes (in equal proportion):

- **Uniform sampling:** target users are randomly selected from the entire pool of accounts with equal probability;
- **Prevalence-based sampling:** target users are sampled with probability proportional to their frequency in the positive edge set, reflecting their natural occurrence in the recommendations;

- **Second-neighborhood sampling:** target users are selected from friends of friends, i.e. being two hops away in the user-to-user network, capturing the local network structure.

Both uniform and prevalence-based sampling present disadvantages when used separately. Uniform sampling risks creating a model that learns to score edges primarily based on source and target node frequency rather than meaningful relationships between nodes. Prevalence-based sampling may not sufficiently penalize high scores assigned to rare entities. We corroborate the robustness of the trained embeddings to alternative negative sampling strategies in the SI in Section C.2.2. To further emulate X’s WTF pipeline, we include non-recommended second-order friends (26); excluding first-degree connections i.e. direct friends. The homophilic nature of X/Twitter networks (49, 50) motivates the search for new ties from friends of friends rather than random strangers. In our data, 62.4% of observed WTF recommendations are second-degree neighbors. Among recommended second neighbors, only 15.8% are followed by only one of the user’s friends, while half are already followed by at least 8 friends of the user to which recommendations are being proposed. The degree distribution (see Extended Figure S7) showcases that recommended second-degree neighbors tend to have substantially more first-degree friend connections compared to non-recommended ones. Consequently, prevalence sampling from non-recommended second-order friends provides particularly valuable negative examples, as these represent plausible but ultimately non-recommended connections —effectively serving as “hard negatives” for the model. Finally, in curating sets of negative WTF ($\mathcal{N}_{\text{WTF}}$) and “Follow” ($\mathcal{N}_{\text{Follow}}$) edges, we adopt a negative-to-positive ratio of 3:1, comparable with other works (51), resulting in $|\mathcal{N}_{\text{Follow}}| = 3|\mathcal{P}_{\text{Follow}}|$ and $|\mathcal{N}_{\text{WTF}}| = 3|\mathcal{P}_{\text{WTF}}|$.

Computing the inference of positions in the embedding

Following X’s implementation (27), we train a 256-dimensional embedding for users and relationship minimizing loss function in (1). The embedding was Glorot initialized (52) to prevent vanishing or exploding gradients during training, and optimized with Adagrad (53). We randomly segmented the curated edge set into 80%, 10%, and 10% splits for training, hyperparameter tuning, and final evaluation. The splitting was performed independently for each relation type of edge to maintain consistent Follow/WTF ratios across all sets.

We tuned the hyperparameter $\alpha \in [0, 1]$ in (1) on a 10% holdout validation dataset. Due to the nondeterministic nature of stochastic gradient descent, we trained the model ten times for each α value over three epochs, evaluating the ROC AUC on the validation dataset for each replicate (see Extended Figure S8). The model that best predicted

the relations in our dataset was achieved at $\alpha = 62.6\%$, yielding an AUC-ROC of 0.7562 ± 0.0018 (standard deviation over ten replicates). Training the model solely on WTF ($\alpha = 1$) or Follow ($\alpha = 0$) data would have yielded AUC ROC values below 0.65 (Extended Figure S8). In what follows, we consider the embedding positions of the model trained with the optimal value $\alpha = 62.6\%$ (assigning each “Who to Follow” training edge 1.67 times the weight of each “Follow” edge) over eight epochs, and three independent runs.

Validating inferred positions in the embedding space

We inferred $\Phi \in \mathbb{R}^{|\mathcal{E}| \times 256}$, the positions of users from $\mathcal{E}$ in the embedding, along with “Who to Follow” ϕ_{WTF} and “Follow” ϕ_{Follow} relationships embeddings. On a 10% held-out test set, the learned embeddings achieve an AUC-ROC of 70.0% for predicting “Who to Follow” recommendations when using non-recommended second-order friends as negative edges. When evaluating positive test recommendations against a pool of randomly sampled 100 non-recommended second-order friends, the model achieves a precision@1 of 72.5%. Matching Twitter’s “Who to Follow” interface which typically displays a slate of three recommendations, we evaluate precision@3, reaching 69.1%, while precision@10 equals 64.3%.

Unlike standard recommendation systems that predict user reactions to suggestions, our model predicts X’s recommendation pipeline outputs. This shift in objective—predicting model behavior rather than user behavior—explains our model’s strong performance relative to literature benchmarks (54). Overall, embeddings Φ successfully predict X’s “Who to Follow” recommendations observed with our plug-in, indicating that studying the inferred embedding space provides a good approximation for Twitter’s internal representation of entities.

Simulating sources of bias (reliability of the embedding)

Next, we conducted experiments to assess the robustness of our procedure against several possible sources of bias. Because our subsequent analysis relies on identifying directions in the embedding space through Canonical Correlation Analysis (CCA), we assess the robustness of learned embeddings up to non-singular affine transformations, as CCA is invariant to such transformations (55). We measure the average coefficient of determination R^2 (56), as well as the average cosine similarity and inter-quantile value [2.5-97.5], between the reference embeddings Φ of each user and each variant $\tilde{\Phi}$ considered in different experiments, after applying the optimal nonsingular affine transformation determined through least squares optimization.

First, we conducted experiments assessing the stability of the embedding for different parameters determining the training. We performed 10 training runs using the original training dataset, saving the learned representations after 3, 5, and 10 epochs. Original embeddings were learned at 3 epochs, as determined optimal on the validation dataset. The coefficients of determination R^2 between Φ and $\tilde{\Phi}$ show consistent stability across non-deterministic training runs, with highest values at 3 epochs ($R^2 = 0.84759 \pm 0.00052$), and an average pairwise cosine similarity of 0.949 [0.873, 0.973]. Embedding stability decreases thereafter, with R^2 values dropping at 5 epochs ($R^2 = 0.83757 \pm 0.00062$) and 10 epochs ($R^2 = 0.77506 \pm 0.00066$). We also compared embeddings for different values for α in (1). We performed five training runs on the original training dataset for 3 epochs, setting α to 42.6%, and 87.6% (the original embeddings Φ were learned with $\alpha = 62.6\%$, optimal value on the validation dataset), achieving validation AUC-ROC scores of 0.7503 ± 0.0019 and 0.7494 ± 0.0019 respectively (standard deviation over ten replicates, and R^2 between embeddings Φ and variants $\tilde{\Phi}$ of 0.8417 ± 0.0041 for $\alpha = 42.6\%$ and 0.7762 ± 0.0061 for $\alpha = 87.6\%$ (standard deviation over five replicates). Complementarily, the average pairwise cosine similarity was 0.946 [0.856, 0.971] and 0.932 [0.822, 0.965]. See further details and additional experiments in SI Section C.2.2, including experiments with alternative positive-to-negative edge ratios, negative sampling strategies and degree threshold in the selection of users.

Second, we conducted experiments assessing the stability of the embedding against different socio-demographical compositions, and possible changes occurred in the platform during the observation period. Our browser plug-in only captures recommendations displayed on Twitter’s desktop interface. In our data, we observe that each account is suggested 5.11 times on average to the same volunteer. We thus postulate that most potentially missed recommendations for volunteers are redundant with desktop ones and thus already included in the training dataset, prompting us to experiment with different compositions of data using our observed recommendations. Assuming volunteers spend 50% of their X consumption on desktop browsers and 50% on other devices, we estimate that we may miss up to 19.6% of unique accounts recommended to volunteers. To simulate this effect, we augment each volunteer’s dataset by adding 19.6% of *artificial non-observed* positive edges, randomly sampled following the same rationale as for the previous curation of negative edges. We trained variant embeddings $\tilde{\Phi}$ using identical hyperparameters as reference embeddings Φ (3 epochs, $\alpha = 62.6\%$) on 10 random samplings of additional “Who to Follow” edges. The average coefficient of determination R^2 between embeddings Φ and $\tilde{\Phi}$ was 0.8417 ± 0.0033 , and the average cosine similarity of 0.946 [0.855, 0.970]. In a more extreme scenario, where users spend 75% of their Twitter consumption on non-observed devices (meaning that we

would be missing potentially 39.2% of recommendations) the R^2 between Φ and 10 variants $\tilde{\Phi}$ with 39.2% artificial non-observed positive edges remained high at 0.8177 ± 0.0031 , and the average cosine similarity at 0.942 [0.840, 0.969].

In our study volunteers are self-selected. Upon enrollment, volunteers may indicate their age, gender, and political preference on a continuous scale from far-Left to far-Right (see Extended Figure S9), revealing that respondents tended to be Left-leaning males over 35 years old. To evaluate the impact of demographic skew, we constructed alternative training datasets of recommendations served only to subsets of volunteers who completed the survey. Using 3 epochs and $\alpha = 62.6\%$ (as with the original embedding), we trained variant embeddings $\tilde{\Phi}$ over the following three datasets: a) 284 self-identified male volunteers, b) 186 volunteers aged over 35, and c) 218 Left-leaning volunteers. The coefficient of determination R^2 between Φ and $\tilde{\Phi}$ was 0.654 for dataset (a), 0.552 for (b), and 0.611 for (c). Similarly, the average pairwise cosine similarity was for dataset (a) 0.870 [0.698 0.934], for (b) 0.825 [0.599 0.917] and for (c) 0.852 [0.655 0.926].

Our dataset of recommendations spans a 16-month collection period, with varying collection rates (see Extended Figure S4). A notable increase in “Who to Follow” recommendations occurred in May 2023, around the time of policy changes on the platform. We hypothesize that these changes raised interest on the part of users to know more about the quality of the recommender, increasing the rate of volunteers, as they were offered personalized analyses. To assess potential effects of changes in population, we split the dataset into two periods: recommendations collected before and after May 1st, 2023. We then computed inferences of embeddings using our previously described methodology. The embeddings learned from these temporal splits show significant alignment with the original embeddings Φ , achieving R^2 values of 0.784 and 0.773 for the pre-May and post-May 2023 periods, respectively. The average cosine similarity was 0.921 [0.669 0.967] and 0.927 [0.726 0.968] respectively.

In our training dataset, we aggregated all accounts followed by volunteers, irrespective of when they started following them, with one notable exception: if a WTF suggestion turned into a friend, we solely considered the WTF observation, as the signal was scarcer for these cases. Consequently, some accounts in the training dataset labeled as “friends” were not followed when the first WTF recommendations were collected. Similarly, users in our sample might have unfriended/unfollowed some accounts included in our study. To assess how robust our embeddings Φ are against these changes, we compared volunteers’ friends lists at the beginning and at the end of our observation period, allowing us to derive friend addition and removal rates. This revealed that, at most, 36.9% of friends in the training set were either a) not yet followed (26.7%) at the experiment’s start, or b) removed from the friend pool (10.2%) since then. We

thus randomly removed 36.9% of friends from the training dataset and computed the corresponding alternative embeddings $\tilde{\Phi}$, repeating this process three times. The comparison between Φ and $\tilde{\Phi}$ achieved a coefficient of determination R^2 of 0.80217 ± 0.00028 (standard deviation over three replicates), and an average cosine similarity of 0.931 [0.841 0.965]. See further details and additional experiments in SI Section C.2.3, including the removal of edges between several groups of users and training of embeddings using data from different time-periods

Political attributes of users

We leverage a dataset of ideologically positioned X users developed by Ramaciotti et al. (57) to investigate representations of ideology in embeddings Φ . This dataset contains 494 065 Twitter users interested in French politics, positioned along two political scales or dimensions: 1) a Left-Right ideology dimension and 2) a dimension measuring attitudes towards elites and institutions. The dataset comes with a detailed documentation (28) explaining the methodology used and the quality metrics showcasing the accuracy of inferred political positions.

The methodology relies on two steps. Step (1) uses ideological scaling (22) of follower networks, first developed for Liberal-Conservative ideological spatialization of Twitter users (3), and subsequently developed to capture additional political dimensions (4). The population in (57) is chosen among users that follow at least three Members of Parliament (MPs) on X/Twitter (as Barberà did in his 2015 studies (3, 29)), selecting 494 065 users that have an interest in national politics. Next, the follower network subtended by these users is collected, and a political homophily process is proposed as an explanation for the observed following links connecting MPs and their users:

$$P(U_i \rightarrow MP_j | \alpha_i, \beta_j, \gamma, \varphi_i, \varphi_j) = \text{logistic} \left(\alpha_i + \beta_j - \gamma \|\varphi_i - \varphi_j\|^2 \right), \quad (3)$$

where, $U_i \rightarrow MP_j$ stands for the event in which user U_i follows MP_j , α_i is the level of *activity* of user U_i measured in number of followed accounts, β_j is the level of *popularity* of MP_j , measured in number of followers, γ is a scale variable, and φ_i and φ_j are the multidimensional positions of U_i and MP_j in some unobservable multidimensional space. This data generation process is homophilic in the sense that $P(U_i \rightarrow MP_j)$ increases as the distance between unobservable positions $\|\varphi_i - \varphi_j\|$ in some latent space decreases. Step (1) involves using observed follower networks between MPs and their followers to infer the parameters of the model in (3). The reader is referred to the work of Barberà on Twitter (29) for the detailed methodology on this inference and to

the documentation of the dataset we use (28) for the details of the application of the methods to the data of this study.

Step (2) uses political survey data to calibrate multidimensional positions φ_i resulting from the Bayesian inference for the parameters of (3), as $P(U_i \rightarrow MP_j)$ is invariant to isometries on positions φ_i and φ_j . For calibration and identification, the dataset from (28) uses party positions available for political parties, from the Chapel Hill Expert Survey (CHES) (30), created with the responses from hundreds of experts in political science, who position political parties along relevant dimensions. Creating proxies for party positions as the centroids of MPs φ_j , the dataset from (28) is compounded with an affine transformation going from the parameter space resulting from the Bayesian inference using (3), onto the space of the dimensions of the CHES data. The leading political dimension of the CHES data is the Left-Right dimension (ranging from 0 to 10, from Left to Right, for positions of parties). Because the scale used for calibration is constructed for positions of political parties, users can have positions to the left of position 0 (the leftmost position for parties) and to the right of position 10 (the rightmost position for parties). Previous research has shown that a second relevant dimension in France is the dimension measuring attitudes towards elites and institutions (58). Because of its importance and availability in the dataset we also leverage this second dimension in our study. Extended Figure S14 shows the density of users in the population along these two dimensions, and the positions of MPs and parties used for the calibration. Extended Figure S15 shows validations metrics proposed in (28) for positions of users along these two dimensions, based on annotations of profiles of users. See Section D.4 of the SI for a detail presentation of this dataset.

An analysis of this dataset reveals that 8 249 users in the population $\mathcal{E}$ of our study (31.1%) also have political positions in the dataset from (28). Extended Figure S16 shows that the distributions of users from $\mathcal{E}$ that have both, Left-Right and Anti-elite rhetoric positions do not deviate significantly from the French population delimited in the construction of the data set from (28).

Estimating demographic attributes of users

We analyze profile pictures, screen names, usernames, and biographical descriptions to infer demographic attributes for users our study, applying the Multi-modal, Multilingual, and Multi-attribute (M3) model, developed by Wang et al. (23), on the 24 425 accounts in $\mathcal{E}$ that have available information (see Extended Figure S11 for the gender and age distributions of profiles inferred by M3).

We assessed the accuracy of these inferences by comparing inferred age and gender with self-declared values (for volunteers), yielding an F1-score of 0.84 for gender and an average F1-score of 0.82 for the age groups. Section D.2 of the SI provides further details of this assessment, as well as

other forms of assessment, based on comparisons with inferences with a second mode (FairFace (59)), on analyses comparing inferred ages and dates of creations of accounts, and finally on analyses characterizing the most salient words used in posts by gender and age groups.

Estimating additional attributes of users

We also established additional attributes of users: their popularity measured in number of followers (as this is probably relevant in recommendations), and the degree of interest of users in different topics. Interest in different topics discussed on X is known to be part of the features considered by the content (e.g., tweet) recommender system (41). We established the number of followers using metadata provided by the API for each user, ranging from 0 to 197 million among users in $\mathcal{E}$, with a median of 7 987. On 27th October 2024, we collected the last 200 tweets (published or retweeted) from 25 080 users in $\mathcal{E}$ (excluding private and deleted accounts), totaling 2.3 million tweets. We leveraged the multi-topic classification model `cardiffnlp/tweet-topic-21-multi` developed by Antypas et al. (31) specifically conceived for tweets, to categorize each tweet into 19 generic topics including arts & culture, news & social concern, and science & technology. Non-English tweets were translated using Opus-MT model before topic inference. We measure the degree of interest of a user for a topic as the proportion of tweets per topic for each user across all 19 categories considered by (31). “News & Social Concern”, hereinafter simply *news*, is the most discussed topic among embedded users (Extended Figure S13 shows the distribution of interests in all topics). As we did with age groups, we report the 25 most overrepresented words in tweets classified under each topic, compared to their expected frequency under uniform random distribution across topics, showing that our metric captures interest in topics from (31) in our population (see Extended List S2).

Finally, we computed additional attributes of interest, including (see Extended Figure S10 for a full description): follower-to-friend ratio (assesses network reciprocity, with median 0.097), mean number of daily posts per user (as the total posts (status) divided by days since account creation, with median 2.21 posts per day) and mean number of daily likes (as the total likes divided by days since account creation, with median 1.40 likes per day). We also identified in X profiles whether users had the legacy *verification* status (i.e., the blue check-mark sign signaling whether the account is “authentic, notable, and active”), typically granted (at the time of our collection) to public figures, journalists, and influential organizations before the platform’s verification policy changes. Up to 30.1% of the accounts in $\mathcal{E}$ were, as of early 2023, verified. We complete our attributes with an estimation of the linguistic complexity displayed by users in their posts, computed as the average word-length in posts (after pre-processing to remove non-linguistic elements such

as mentions, hashtags, emojis, URLs; see Extended List S3 for the 15 words with the highest frequency in the first and last decile of users according to the word-length metric).

Measuring attributes encoded in the embedding space

Let us $\Phi \in \mathbb{R}^{n \times d}$ denote the matrix of user embeddings, where each row $\phi_u^\top \in \mathbb{R}^d$ represents the embedding vector of user $u \in \{1, \dots, n\}$ in a d -dimensional embedding space; in our case $d = 256$ and $n = |\mathcal{E}| = 26,509$. Given an attribute vector $\mathbf{a} \in \mathbb{R}^n$, where a_i corresponds to a measured characteristic of user i (political leaning, age, gender or any other of the inferred attributes), we aim to identify the direction in the embedding space such that projections along the direction that maximally correlate with this attribute.

For any unit vector $\mathbf{w} \in \mathbb{R}^d$, the correlation between the projection of user embeddings onto the direction spanned by $\mathbf{w}$ and the attribute vector $\mathbf{a}$ is given by:

$$\rho_{\mathbf{a}}^\Phi(\mathbf{w}) = \frac{(\Phi\mathbf{w})^\top \mathbf{a}}{\|\Phi\mathbf{w}\|_2 \|\mathbf{a}\|_2} \quad (4)$$

where $\|\cdot\|_2$ denotes the Euclidean norm. We seek to find $\mathbf{w}_{\mathbf{a}}$ such that the projection $\Phi\mathbf{w}_{\mathbf{a}}$ of users' embeddings Φ onto $\mathbf{w}_{\mathbf{a}}$ maximally correlate with $\mathbf{a}$:

$$\mathbf{w}_{\mathbf{a}} = \underset{\mathbf{w} \in \mathbb{R}^d: \|\mathbf{w}\|_2=1}{\operatorname{argmax}} \rho_{\mathbf{a}}^\Phi(\mathbf{w}) \quad (5)$$

This optimization problem is a special case of Canonical Correlation Analysis (CCA), where one seeks to find a linear combination that maximizes correlation between a univariate dataset (the attribute vector $\mathbf{a}$) and a multivariate dataset (the user embeddings Φ). The vector $\mathbf{w}_{\mathbf{a}}$ can then be readily obtained by solving the CCA eigenvalue problem. To establish statistical significance of the identified directions, we employ a permutation test with $N = 10^4$ repetitions. For each iteration, we generate $\tilde{\mathbf{a}}$ by randomly permuting the elements of $\mathbf{a}$, compute $\mathbf{w}_{\tilde{\mathbf{a}}}$ through CCA, and measure the resulting correlation $\rho_{\tilde{\mathbf{a}}}^\Phi(\mathbf{w}_{\tilde{\mathbf{a}}})$. By permuting the feature vector, we obtain the null distribution of correlations expected if the embedding space under consideration was not meaningfully encoding the feature of interest. The one-sided p-value is subsequently determined as the proportion of permuted absolute correlations that exceed the absolute observed correlation $\rho_{\mathbf{a}}^\Phi(\mathbf{w}_{\mathbf{a}})$.

Figure 3. A shows the Pearson correlation induced by the CCA directions capturing the main attributes in our study (all p-values < 0.0001). Extended Table S2 shows the Pearson correlation for all attributes estimated for our study.

Validating directions in the embedding capturing attributes

After identifying through CCA the directions in the embedding space that induce positions that maximally correlates

with relevant attributes, we conduct additional validations showing that these CCA directions capture attributes. This is needed because of two reasons. 1) Even if a direction produces a high correlation in ordering and positioning users according to attributes for users, the variance or disturbance error of an underlying linear model might still be high. 2) Secondly, not every user in $\mathcal{E}$ has known values for all attributes. Accordingly, we need to control that the CCA directions induce an ordering of users also for those that have no known attributes. The CCA direction for Left-Right leanings was computed relying on the 8 249 users from $\mathcal{E}$ (31.1%) that had a position on the Left-Right scale, and we need to control that the direction orders the rest of users in $\mathcal{E}$ correctly. To validate the positions on CCA directions independently from the data we used inferring a direction $\mathbf{w}_{\mathbf{a}}$, we analyze the friends and tweet content of users at both extremes of the direction. Specifically, we identify overrepresented friends (followed accounts), words, and hashtags in the tweets produced by users in the first and last deciles of positions along $\mathbf{w}_{\mathbf{a}}$. We determine overrepresented entities by comparing their observed frequency in the target pool to their expected frequency under uniform random distribution along the $\mathbf{w}_{\mathbf{a}}$ direction. To protect privacy of users, we only included followed accounts with more than 50K followers and followed by at least 10% of the users in the considered decile. Section F of the SI shows the results of these validations for identified directions in the embedding capturing ideology, age, gender, and thematic interests among other attributes.

Measuring the alignment between directions in the embedding

The measurements of alignment and dependence between positions along CCA directions presented in the article include Spearman correlations (Extended Figure S29) and cosine similarities in Figure 3.B. To assess the statistical significance of cosine similarities, we draw values of cosine similarities under a null hypothesis. Since the null distribution of attribute directions $\mathbf{w}_{\mathbf{a}}$ depends on the distribution of user embeddings in $\mathbb{R}^d$, we use permutation tests, rather than assuming a uniform hyper-sphere distribution, to assess the significance of similarities between $\mathbf{w}_{\mathbf{a}}$ and $\mathbf{w}_{\mathbf{b}}$, the embedding directions associated with features $\mathbf{a}$ and $\mathbf{b}$. The one-sided p-value is computed as the proportion of randomized cosine similarities that exceed the observed similarity between $\mathbf{w}_{\mathbf{a}}$ and $\mathbf{w}_{\mathbf{b}}$, where the randomized similarities are computed between embedding directions $\mathbf{w}_{\tilde{\mathbf{a}}}$ and $\mathbf{w}_{\tilde{\mathbf{b}}}$ associated with shuffled features $\tilde{\mathbf{a}}$ and $\tilde{\mathbf{b}}$.

Computing politically-constrained recommendations

The method proposed for constraining the amount of information on political preferences of users available to the

recommender borrows from the method called “debiasing” by Bolukbasi et al. (35) for language models. We adapt this methodology to “protect” (34) user attributes from being used by recommender systems. Consider a user attribute $\mathbf{a}$ to be protected. Using CCA, we first identify the direction $\mathbf{w}_\mathbf{a}$ for which the projection of positions of users in the embedding, i.e., $\Phi \mathbf{w}_\mathbf{a}$, maximally correlates with $\mathbf{a}$. We then project embedded positions of users Φ onto $\mathbf{w}_\mathbf{a}^\perp$, the subspace orthogonal to $\mathbf{w}_\mathbf{a}$, removing linearly-encoded information about $\mathbf{a}$ along this direction. In high-dimensional spaces, attributes may be potentially encoded in multiple orthogonal directions (36). Therefore, we repeat this process, iteratively reducing the effective dimension of the embedding, identifying each time the quality of the correlation achieved by the new CCA direction, until no spatial direction in the embedding induces a significant correlation with the attribute selected to be constrained or protected. For our experiment, we seek to protect Left-Right leanings of users, and we set the threshold of significance at 0.01, meaning that we consider that no further linear encoded is significant when the p-value of the correlation overflows above this value; corresponding to Pearson correlation of 0.19408, according to the distribution obtained through permutation. Importantly, this procedure only removes linearly-encoded information. However, since user recommendations are formulated through inner products between user embeddings, our procedure effectively constrains recommendations from leveraging attribute $\mathbf{a}$.

Evaluating politically-constrained recommendations

Holding out 10% of observed WTF recommendations for evaluation, we find that the protected embeddings are 20.6% less informative (in terms of normalized mutual information) about whether a given recommendation was served to a user compared to the original (unprotected) embeddings. Subsequently, we use this protected embedding to compute friend recommendations to characterize relevant properties: relevance and accuracy of recommendations. To compute friend recommendations for each user $s \in \mathcal{E}$, we first select all users $t \in \mathcal{E}$ that s does not already follow, and rank each potential recommendation (i.e., of user t to user s) by evaluating the TransE scoring function (51): $f(s, WTF, t) = (\phi_s + \varphi_{WTF})^T \phi_t$, where φ_{WTF} represents the WTF relation embedding computed during training.

We measure the ideological diversity of friend recommendations made to each user as the standard deviation of the Left-Right position of the recommended friends. To measure thematic relevance of recommendations, we inspect the relative interest in news content that recommended friends have when compared to the users receiving recommendations (among types of contents, news content is both relevant and in risk of being affected by signals re-

lated ideological leaning). Interest in news content is measured as the fraction of tweets published or shared that were classified as related to “News & Social Concern” by the `cardiffnlp/tweet-topic-21-multi`. For each user we measure the mean distance between its own interest in news and those of proposed recommendations. Figure 4 reports results based on 50 recommendations per user s . Extended Figure S31 shows the effect of the protected embedding on the diversity of recommendations for users separated into three groups according to ideological leaning: Left-leaning (ideological leaning below 4.5), Center-leaning (ideological leaning between 4.5 and 7), and Right-leaning users (ideological leaning above 7).

References

1. A. Amini, A. P. Soleimany, W. Schwarting, S. N. Bhatia, D. Rus, Uncovering and mitigating algorithmic bias through learned latent structure, in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (2019), pp. 289–295.
2. T. Fagni, S. Cresci, Fine-grained prediction of political leaning on social media with unsupervised deep learning. *Journal of Artificial Intelligence Research* **73**, 633–672 (2022).
3. P. Barberá, Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political analysis* **23** (1), 76–91 (2015).
4. P. Ramaciotti, et al., Inferring attitudinal spaces in social networks. *Social Network Analysis and Mining* **13** (1), 14 (2022).
5. P. Lorenz-Spreen, L. Oswald, S. Lewandowsky, R. Herwig, A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature human behaviour* **7** (1), 74–101 (2023).
6. E. Bakshy, S. Messing, L. A. Adamic, Exposure to ideologically diverse news and opinion on Facebook. *Science* **348** (6239), 1130–1132 (2015).
7. P. Bouchaud, D. Chavalarias, M. Panahi, Crowdsourced audit of Twitter’s recommender systems. *Scientific Reports* **13** (1) (2023), doi:10.1038/s41598-023-43980-4, <http://dx.doi.org/10.1038/s41598-023-43980-4>.
8. S. Milli, et al., Engagement, user satisfaction, and the amplification of divisive content on social media. *PNAS nexus* **4** (3), pgaf062 (2025).
9. K. Duskin, J. S. Schafer, J. D. West, E. S. Spiro, Echo Chambers in the Age of Algorithms: An Audit of Twitter’s Friend Recommender System, in *ACM Web Science*

- Conference, Websci '24 (ACM) (2024), p. 11–21, doi:10.1145/3614419.3643996.
10. B. Nyhan, *et al.*, Like-minded sources on Facebook are prevalent but not polarizing. *Nature* **620** (7972), 137–144 (2023).
 11. F. Huszár, *et al.*, Algorithmic amplification of politics on Twitter. *Proceedings of the National Academy of Sciences* **119** (1), e2025334119 (2022).
 12. C. Jia, M. S. Lam, M. C. Mai, J. T. Hancock, M. S. Bernstein, Embedding Democratic Values into Social Media AIs via Societal Objective Functions. *Proceedings of the ACM on Human-Computer Interaction* **8** (CSCW1), 1–36 (2024), doi:10.1145/3641002, <http://dx.doi.org/10.1145/3641002>.
 13. L. Rheault, C. Cochrane, Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political analysis* **28** (1), 112–133 (2020).
 14. J. Milbauer, A. Mathew, J. Evans, Aligning multidimensional worldviews and discovering ideological differences, in *Proceedings of the 2021 conference on empirical methods in Natural Language Processing* (2021).
 15. T. Faverjon, P. Ramaciotti, Discovering ideological structures in representation learning spaces in recommender systems on social media data, in *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining* (2023), pp. 400–406.
 16. C. Sandvig, K. Hamilton, K. Karahalios, C. Langbort, Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* **22** (2014), 4349–4357 (2014).
 17. L. Van Vliet, P. Törnberg, J. Uitermark, The Twitter parliamentarian database: Analyzing Twitter politics across 26 countries. *PLoS one* **15** (9), e0237073 (2020).
 18. A. Jungherr, Twitter use in election campaigns: A systematic literature review. *Journal of information technology & politics* **13** (1), 72–91 (2016).
 19. H. Kwak, C. Lee, H. Park, S. Moon, What is Twitter, a social network or a news media?, in *Proceedings of the 19th international conference on World wide web* (2010), pp. 591–600.
 20. Twitter, Twitter's Recommendation Algorithm, https://blog.x.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm, accessed: 2024-12-29.
 21. A. El-Kishky, *et al.*, Twihin: Embedding the twitter heterogeneous information network for personalized recommendation, in *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining* (2022), pp. 2842–2850.
 22. J. Clinton, S. Jackman, D. Rivers, The statistical analysis of roll call data. *American Political Science Review* **98** (2), 355–370 (2004).
 23. Z. Wang, *et al.*, Demographic inference and representative population estimates from multilingual social media data, in *The world wide web conference* (2019), pp. 2056–2067.
 24. Twitter, Discovering Who To Follow, https://blog.x.com/official/en_us/a/2010/discovering-who-to-follow.html, accessed: 2025-01-06.
 25. A. Goel, *et al.*, The who-to-follow system at Twitter: strategy, algorithms, and revenue impact. *Interfaces* **45** (1), 98–107 (2015).
 26. Twitter, Twitter/The-Algorithm/Twitter's Recommendation Algorithm, <https://github.com/twitter/the-algorithm/tree/main> (2023).
 27. X. Zhang, *et al.*, Twihin-bert: A socially-enriched pre-trained language model for multilingual tweet representations at twitter, in *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining* (2023), pp. 5597–5607.
 28. P. Ramaciotti, J. Royo-Letelier, J.-P. Cointet, A. Pournaki, Attitudinally-positioned European sample dataset. *Social Media for Democracy, Sciences Po Paris* p. 18 p. (2024).
 29. P. Barberá, J. T. Jost, J. Nagler, J. A. Tucker, R. Bonneau, Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science* **26** (10), 1531–1542 (2015).
 30. S. Jolly, *et al.*, Chapel Hill Expert Survey trend file, 1999–2019. *Electoral Studies* **75**, 102420 (2022), doi:10.1016/j.electstud.2021.102420, <http://dx.doi.org/10.1016/j.electstud.2021.102420>.
 31. D. Antypas, *et al.*, Twitter Topic Classification, in *Proceedings of the 29th International Conference on Computational Linguistics* (International Committee on Computational Linguistics, Gyeongju, Republic of Korea) (2022), pp. 3386–3400.
 32. G. Ivaldi, Populism in France, in *Populism Around the World. A Comparative Perspective*, D. Stockemer, Ed. (Springer), pp. 27–48 (2018), doi:10.1007/978-3-319-96758-5_3, <https://shs.hal.science/halshs-01889832>.

33. X. He, P. Niyogi, Locality Preserving Projections, in *Advances in Neural Information Processing Systems*, S. Thrun, L. Saul, B. Schölkopf, Eds. (MIT Press), vol. 16 (2003).
34. S. Ravfogel, Y. Elazar, H. Gonen, M. Twiton, Y. Goldberg, Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, J. Tetreault, Eds. (Association for Computational Linguistics, Online) (2020), pp. 7237–7256, doi:10.18653/v1/2020.acl-main.647.
35. T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, A. T. Kalai, Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings, in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett, Eds. (Curran Associates, Inc.), vol. 29 (2016).
36. H. Gonen, Y. Goldberg, Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them, in *Proceedings of the 2019 Workshop on Widening NLP*, A. Axelrod, D. Yang, R. Cunha, S. Shaikh, Z. Waseem, Eds. (Association for Computational Linguistics, Florence, Italy) (2019), pp. 60–63, <https://aclanthology.org/W19-3621>.
37. C. A. Bail, *et al.*, Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences* **115** (37), 9216–9221 (2018).
38. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* **26** (2013).
39. S. Arora, Y. Li, Y. Liang, T. Ma, A. Risteski, A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics* **4**, 385–399 (2016).
40. N. Elhage, *et al.*, Toy models of superposition. *arXiv preprint arXiv:2209.10652* (2022).
41. V. Satuluri, *et al.*, Simclusters: Community-based representations for heterogeneous recommendations at twitter, in *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining* (2020), pp. 3183–3193.
42. R. Xenidis, Beyond bias: algorithmic machines, discrimination law and the analogy trap. *Transnational Legal Theory* **14** (4), 378–412 (2023).
43. P. Ramaciotti, Depolarizing And Moderating Social Media With AI: Tools and guidelines leveraging representation spaces. *AI4Democracy, IE Center for the Governance of Change* (2024).
44. L. Grozdanovski, In search of effectiveness and fairness in proving algorithmic discrimination in EU law. *Common Market Law Review* **58** (1) (2021).
45. J. Strycharz, C. Meppelink, B. Zarouali, T. Araujo, H. Voorveld, The blind spot in data donations: who is (not) willing to donate digital data in social scientific research. *Computational Communication Research* **6** (2), 1 (2024).
46. Z. Kmetty, *et al.*, Determinants of willingness to donate data from social media platforms. *Information, Communication & Society* **28** (7), 1324–1349 (2025).
47. Z. Kmetty, *et al.*, Determinants of willingness to donate data from social media platforms in the US and Hungary. *Information, Communication & Society* **28** (7), 1324–1349 (2025).
48. A. Lerer, *et al.*, PyTorch-BigGraph: A Large-scale Graph Embedding System (2019), <https://arxiv.org/abs/1903.12287>.
49. E. Colleoni, A. Rozza, A. Arvidsson, Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of communication* **64** (2), 317–332 (2014).
50. Y. Halberstam, B. Knight, Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter. *Journal of public economics* **143**, 73–88 (2016).
51. A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko, Translating Embeddings for Modeling Multi-relational Data, in *Advances in Neural Information Processing Systems*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Weinberger, Eds. (Curran Associates, Inc.), vol. 26 (2013).
52. X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, Y. W. Teh, M. Titterton, Eds. (PMLR, Chia Laguna Resort, Sardinia, Italy), vol. 9 of *Proceedings of Machine Learning Research* (2010), pp. 249–256, <https://proceedings.mlr.press/v9/glorot10a.html>.
53. J. Duchi, E. Hazan, Y. Singer, Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research* **12** (61), 2121–2159 (2011), <http://jmlr.org/papers/v12/duchi11a.html>.

54. W. Hu, *et al.*, Open graph benchmark: datasets for machine learning on graphs, in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20 (Curran Associates Inc., Red Hook, NY, USA) (2020).
55. A. Mathai, S. Provost, H. Haubold, *Chapter 10: Canonical Correlation Analysis* (Springer International Publishing), p. 641–677 (2022), doi:10.1007/978-3-030-95864-0_10.
56. D. Stewart, W. Love, A general canonical correlation index. *Psychological bulletin* **70** (3p1), 160 (1968).
57. P. Ramaciotti, J. Royo-Letelier, J.-P. Cointet, A. Pournaki, Attitudinally-positioned European sample dataset, <https://doi.org/10.6084/m9.figshare.25288210.v1> (2024).
58. P. Ramaciotti, J.-P. Cointet, G. M. Zolotoochin, Unfolding the dimensionality structure of social networks in ideological embeddings, in *Proceedings of the 2021 IEEE/ACM international conference on advances in social networks analysis and mining* (2021), pp. 333–338.
59. K. Karkkainen, J. Joo, FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation, in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)* (IEEE) (2021), p. 1547–1557, doi:10.1109/wacv48630.2021.00159.

Acknowledgments

This work has been partially funded by the “European Polarisation Observatory” (EPO) of CIVICA Research (co-)funded by EU’s Horizon 2020 programme under grant agreement No 101017201, by European Union Horizon program project “Social Media for Democracy” under grant agreement No 101094752 (www.some4dem.eu), by the *Very Large Research Infrastructure* (TGIR) Huma-Num of CNRS, Aix-Marseille Université and Campus Condorcet, and by Project Liberty Institute project “AI-Political Machines” (AIPM). P.B. acknowledges financial support from the CFM Foundation for sciences.

Supplementary Information

Recommender system in X inadvertently profiles ideological positions of users

Paul Bouchaud and Pedro Ramaciotti

A	X's personalized recommendations	S3
A.1	The "Who to Follow" service	S3
A.2	Heterogeneous Information Network representation for recommendation	S4
B	Inferring users' embeddings in the recommender system	S5
B.1	Loss function defining the learning task	S5
B.2	Datasets used for training and approximation of the embedding	S5
B.2.1	Crowdsourced (plug-in) dataset of recommendations served in the platform and participant survey	S6
B.2.2	Dataset of follower networks obtained with the platform's API	S8
B.3	Negative sampling	S8
B.4	Training an approximation of the embedding of X's recommender system	S10
B.4.1	Implementation	S10
B.4.2	Tuning hyperparameter α in loss function $\mathcal{L}$	S10
C	Validating the inferred embeddings Φ for users	S11
C.1	Testing the performance of embeddings in predicting recommended users	S11
C.2	Testing the robustness of embeddings Φ under alternative data sub-sampling and model parameter strategies	S12
C.2.1	Evaluating similarity of embeddings under affine transformations	S12
C.2.2	Testing the robustness of embeddings for alternative training parameters	S12
C.2.3	Testing the robustness of embeddings for different data sub-sampling strategies	S13
D	Inferring users' features	S15
D.1	User activity metrics	S15
D.2	Inference of demographic features	S15
D.2.1	Gender	S18
D.2.2	Age	S18
D.3	Content-based attributes	S19
D.4	Ideological Positions of Users	S22
D.4.1	Politically-positioned X population dataset from (57)	S22
D.4.2	Intersection between the politically-positioned dataset from (28) and our population of study $\mathcal{E}$	S23
E	Inferring directions in the embedding space that reproduce user features	S23
F	Validating inferred embedding directions	S27
F.1	Inferred Left-Right political leaning direction $\mathbf{w}_{LR}$	S27
F.2	Inferred anti-elite rhetoric direction $\mathbf{w}_{AntiElite}$	S29
F.3	Inferred gender direction $\mathbf{w}_{gender}$	S31
F.4	Inferred age direction $\mathbf{w}_{age}$	S33
F.5	Inferred directions in the embedding space of the recommender system and that capture interest in different topics as defined by (31)	S35
F.5.1	News $\mathbf{w}_{News}$	S35
F.5.2	Sports $\mathbf{w}_{Sports}$	S35
F.5.3	Science & Technology $\mathbf{w}_{SciencesTech}$	S35
F.6	Inferred directions in the embedding space of the recommender system and that the lexical complexity $\mathbf{w}_{complexity}$	S38
F.7	Summary of correlations of user attributes in the embedding spaces of the recommender system	S40
G	Constraining political position information encoded in the recommender system	S44
G.1	Iterative projections onto subspace orthogonal to feature directions $\mathbf{w}_a$ in the embedding space Φ	S44
G.2	Protocol for the evaluation of accuracy loss of protected embeddings Φ'	S44
G.3	Results: assessing the effects on accuracy and diversity of recommendations	S45
G.4	Constraining other attributes information encoded in the recommender system	S46

A X's personalized recommendations

A.1 The “Who to Follow” service

The “Who to Follow” service aims to enhance Twitter’s user engagement and network growth by recommending accounts that users don’t currently follow but may find interesting, based on shared interests and common connections (1). This feature is important for the platform to sustain platform engagement through user discovery and crucial for new accounts. By the platform’s own report, by 2015, 35% of new Twitter users started following friend recommendations on their first day (2). Launched in 2010, Twitter’s “Who to Follow” service encompasses two complementary components (3): 1) recommending accounts that users may be “Interested in” based on content affinity, and 2) suggesting “similar users” who exhibit analogous following patterns, grounded in the homophily principle (4). With Twitter’s vast user base in the hundreds of millions, this dual approach sought to surface relevant accounts from the platform’s extensive network, and has been reported to drive over one-eight of the new connections on Twitter (2).

To accomplish such a task, rather than operating as a monolithic system, Twitter’s “Who to Follow” service integrates recommendations from diverse algorithms —a blend of 20 different algorithms in 2013 according to Gupta et al. (3). An overview of the service is displayed in Extended Figure S1. The service combines, for instance, Personalized PageRank (5) and SALSA (6) ranking algorithms with heuristics based on the assumption of social homophily, such as pooling candidates from most common second-degree neighbors.

Acknowledging that not all follow relationships carry equal significance in terms the interest and preferences of users, Twitter measures the predicted “engagement per impression” to evaluate recommendation quality, with the objective of increasing the overall engagement on the platform (3).

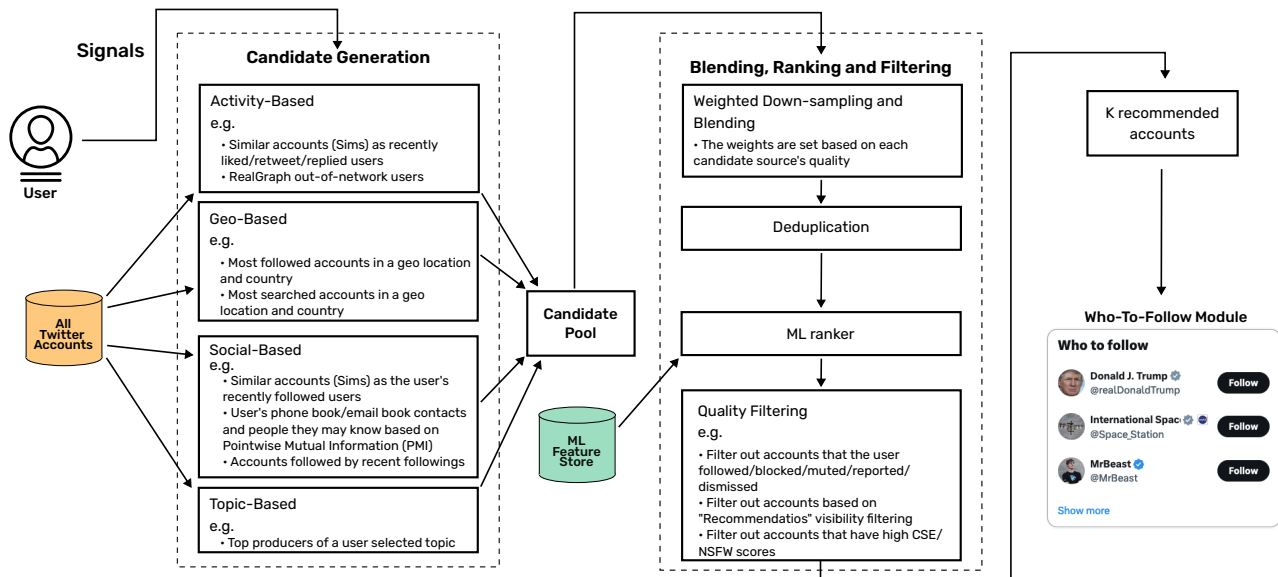


Figure S1: Diagram (adapted from (7)) of Twitter/X “Who to Follow” (WTF) account recommendation system that computes the three friend recommendations that are proposed on the right-hand lateral bar in the website of X. The WTF recommender system can be organized in three main sub-parts: 1) candidate generation (left part of the figure), 2) integration of features computed by machine learning models (at the center of the figure), and 3) candidate mixing and ranking (on the right part of the figure).

Additional details about Twitter’s more recent architecture of the “Who to Follow” profile recommendation system were disclosed in April 2023 together with partial release of the code base of a few key components (8). The recommender system integrates 17 components contributing candidates for profile recommendation (9), encompassing previously established approaches like SimCluster (10), SALSA, and Personalized PageRank, alongside additional heuristics, including: considering accounts with which users have recently interacted (through explicit actions such as likes and retweets, or implicit signals

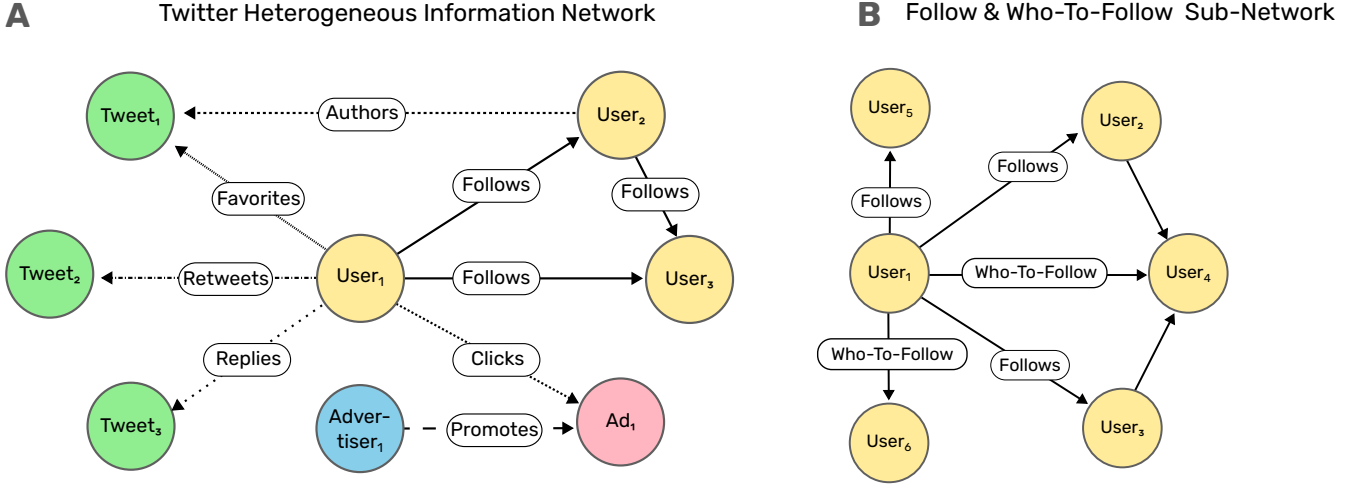


Figure S2: Diagram of the representation of the data used for training the recommender system in X, adapted from (13) (A) X’s Heterogeneous Information Network (HIN, a graph abstraction model for Knowledge Graph data) showing interconnections between entities that are relevant for recommendation (tweets, users, advertisers, advertisements) through diverse relations (authorship, follows, interactions, shares or clicks). (B) Part of the HIN representation relevant for the inference of the embedding space in our study, including “Follow” and “Who to Follow” relationships between users form a sub-network analyzed in this work.

like profile visits), and accounts two-random hops from users in the network of interactions (11). The system also considers promoted (advertiser) accounts and contacts from users’ address books, when synced with Twitter.

Each candidate account undergoes scoring by machine learning models that compute a weighted sum between two probabilities (9):

$$P(\text{user follows account} | \text{account recommended to user}),$$

and

$$P(\text{user gives positive engagement to account} | \text{user follows account}).$$

The accounts for which these probabilities are estimated then pass through refinement heuristics before a ranking of the highest-scoring candidates, which are finally recommended to users.

A.2 Heterogeneous Information Network representation for recommendation

In its candidate scoring process, the “Who to Follow” service employs feature enrichment to capture accounts’ social environment (12). To support this process Twitter researchers (13, 14) implemented a Heterogeneous Information Network (HIN) for personalized recommendations, generating dense entity embeddings, lying in a high-dimensional space, that serve downstream ranking models and candidate generation.

The HIN integrates diverse entity types, such as tweets, users, and advertisers, connected through varied relationships such as authorship, clicks, retweets, and follows; as illustrated on Extended Figure S2. Each entity and edge type is encoded as a learnable embedding vector, optimized to predict specific types of entity-entity links. Extended Figure S2A illustrates the HIN data representation leveraged by X’s recommender system, as shown in the article by X’s researchers (13).

X’s HIN-based recommender system uses observed occurrences of relations between different entities to train an embedding space for all entities and relations on which to compute rankings of potential entities to recommend. The ranking relies on a scoring function f encapsulating inner product similarity between entities in the embedding space. Consider a scoring function $f(e) = f(\phi_s, \varphi_r, \phi_t)$ for an edge e of type r in the HIN connecting source entity s to target entity t . The model optimizes entities’ embeddings ϕ_s, ϕ_t and relation’s embeddings φ_r by maximizing the log-likelihood $\log \sigma(f(e))$ (with σ being the logistic sigmoid function) for observed edges while minimizing it for unobserved ones (i.e., relations that do not exist in the

data). In (13, 14), Twitter report adopting, for model simplicity, the TransE (translating embeddings) scoring function (15). For an edge $e = (s, r, t)$, TransE is defined as the sum of inner products:

$$f(e) = f(s, r, t) = \phi_s^T \phi_t + \varphi_r^T \phi_t. \quad (6)$$

This function operates by adding the relation embedding φ_r to the source entity embedding ϕ_s and computing its inner product with the target entity embedding ϕ_t . In a successfully trained model, the inner product between a target account’s embedding ϕ_t and, for instance, the sum of a source account’s embedding and the “Follow” relationship embedding ($\phi_s + \varphi_{\text{Follow}}$) should yield high values for actual follower relationships and low values otherwise.

B Inferring users’ embeddings in the recommender system

B.1 Loss function defining the learning task

We are interested in this article in inferring users’ positions in this HIN-based embedding space. We use the detailed specification of the machine learning procedures provided by X’s researchers (13) to define the task of inferring embeddings that best approximate those that were learned and leveraged by the platform’s recommender system. As displayed in Extended Figure S2B, we consider the sub-network solely comprised of Twitter accounts connected through two types of edges corresponding to the empirical data available for our study: “Follow” (following links from one user to its followees or friends) or “Who to Follow” (link connecting a user to those that have been recommended by the platform). Our goal is to learn user and edge type embeddings that effectively predict Twitter’s “Who to Follow” recommendations to users that we observe empirically.

Integrating “Follow” relationships in the inference of positions ϕ in the embedding serves two purposes: it aligns with Twitter’s architecture, where user-to-user connections drive 9 of the 17 candidate sources for profile recommendations (8), and it mitigates the sparsity of “Who to follow” observations. As shown in the training section below (Section B.4), integration “Follow” relationships enhance the model’s ability to predict “Who to Follow” recommendations.

To learn embeddings ϕ for users, we consider a convex combination of both relationships through the following loss function $\mathcal{L}$:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{WTF}} + (1 - \alpha) \mathcal{L}_{\text{Follow}}, \quad (7)$$

where each relationship-specific loss component is defined as:

$$\mathcal{L}_r = \sum_{e \in \mathcal{P}_r} \log \sigma(f(e)) + \sum_{e' \in \mathcal{N}_r} \log \sigma(-f(e')) \quad (8)$$

In this formulation, $\mathcal{P}_r$ and $\mathcal{N}_r$ denote the sets of positive and negative edges for a relationship type r with $r \in \{\text{“Follow”}, \text{“Who To Follow”}\}$. “Who To Follow” is hereinafter also shortened as “WTF”.

Both to generate and rank candidate accounts, Twitter’s systems use account features to assess which candidates relate most strongly to a given user. Through the loss function $\mathcal{L}$, the account representations we learned from our experimental data aim to approximate Twitter’s internal representations ϕ . Two accounts will be placed closely in our representation if Twitter recommended one account to a specific user, indicating it was a top-scoring candidate according to Twitter’s internal user representation system.

B.2 Datasets used for training and approximation of the embedding

Our inference of users’ embeddings relies on two data sources: B.2.1) “Who To Follow” recommendations gathered via a crowdsourcing initiative and B.2.2) “Follow” relationships collected through Twitter’s API.

B.2.1 Crowdsourced (plug-in) dataset of recommendations served in the platform and participant survey

The main hurdle in the inference of embedded positions using $\mathcal{L}$ in Eq. (7) stems from acquiring “Who To Follow” recommendation data. We address this through a custom browser extension that unobtrusively collects “Who To Follow” recommendations displayed to volunteer users that installed the extension as they scroll their feed.

This browser extension was installed by self-recruited participants following mainstream media coverage (newspapers and radio) and Twitter advertising in fall of 2022. The study incentivized participation by offering volunteers personalized analytics on their Twitter feed’s political diversity, delivered upon sufficient data collection. No monetary incentives were offered to participants. Participants were required to answer a brief survey, shown in Extended Figure S3, asking their gender, age group, their “ideological position” on a 101-point scale (from 0 to 100) with political parties arranged from Left to Right for reference and orientation, whether they connect to Twitter from their phone or their computer or both (on a gliding 21-point scale from 0 to 100 with intervals of length 5). The parties used for reference are LFI (*La France Insoumise*, or Indomitable France), PS (*Parti Socialiste*, or Socialist Party), *Rénaissance* (Macron’s party, previously known as *En Marche!*), LR (*Les Républicains*, or The Republicans), and RN (*Rassemblement National*, Le Pen’s party previously known as *Front National*). The ordering of parties is taken from the Chapel Hill Expert Survey (16) and are equidistant. The survey also asks a question related to YouTube use (not used in this study). The answers to this survey are not used directly in the conclusions of our study, which relies on inferred socio-demographic and ideological attributes. We administered this survey to produce validation metrics for the inferred attributes of users. The following box provides the questions presented to the respondents, translated into English (Extended Figure S3 for a screenshot).

Translation of the survey used to compute quality metrics for inferred user attributes.

If you want to help us further...

We would be grateful if you could answer a few optional questions (you can complete the survey partially if you want):

- *Your gender:* ☐ Non-Binary ☐ Female ☐ Male
- *Your age:* ☐ <25 years ☐ 25-34 years ☐ 35-49 years ☐ > 50 years
- *To which political party do you feel the closest?*
(gliding scale)
LFI PS Renaissance LR RN
☐ None ☐ I don’t wish to answer
- *You use Twitter:*
(gliding scale)
Only on PC Equally on PC and on my mobile Only on my mobile
- *You use YouTube:*
(gliding scale)
Only on PC Equally on PC and on my mobile Only on my mobile

Save Close

All participants provided informed consent after receiving detailed information about the study objectives and data collection scope. Only after obtaining consent did the browser extension begin monitoring participants’ Twitter interactions. The reader is referred to the Ethics and Compliance section of the main article for the reference to the Ethics Review Board and legal data deposited associated with the data used in this study.

From January 1st, 2023 to May 1st, 2024, we gathered over 2,549,008 “Who To Follow” profile recommendations displayed to 682 unique Twitter users logged-in using our browser plug-in. To ensure the robustness of our results, we only include users having received 30 “Who To Follow” profile recommendations, henceforth referred to as volunteers from the set of volunteers $\mathcal{V}$. Extended Figure S4 shows the daily rate of collection of recommendations from users that used our browser plug-in.

Participation fluctuated with media coverage of the project, peaking in spring 2023 with over ten thousand “Who To Follow” recommendations collected daily; as displayed on Extended Figure S4. Users could freely withdraw from the study by removing the browser extension, leading to natural attrition in collection rates during periods of reduced public outreach. We observe an increase in participation around the time of policy changes in the platform linked with changing ownership.

Si vous voulez nous aider encore plus ...

Nous vous serions reconnaissant de bien vouloir répondre à ces quelques questions facultatives (vous pouvez ne répondre qu'à certaines si vous le souhaitez) :

- Votre genre: ☐ Non-Binaire ☐ Femme ☐ Homme
- Votre âge: ☐ <25 ans ☐ 25-34 ans ☐ 35-49 ans ☐ >50 ans
- De quel parti politique vous sentez le plus proche idéologiquement ?

☐ LFI ☐ PS ☐ Renaissance ☐ LR ☐ RN

☐ Aucun ☐ Ne souhaite pas me prononcer

- Vous utilisez Twitter:

☐ Seulement sur Ordinateur ☐ Autant sur ordinateur que sur Mobile ☐ Seulement sur Mobile

- Vous utilisez Youtube:

☐ Seulement sur Ordinateur ☐ Autant sur ordinateur que sur Mobile ☐ Seulement sur Mobile

Figure S3: Screenshot of the survey presented to volunteers of the data donation program (in French), asking: their gender, age group, their “ideological position” on a 101-point scale with political parties arranged from Left to Right for reference and orientation, whether they connect to Twitter from their phone or their computer or both. The survey also asks the same question for YouTube (not used in this study).

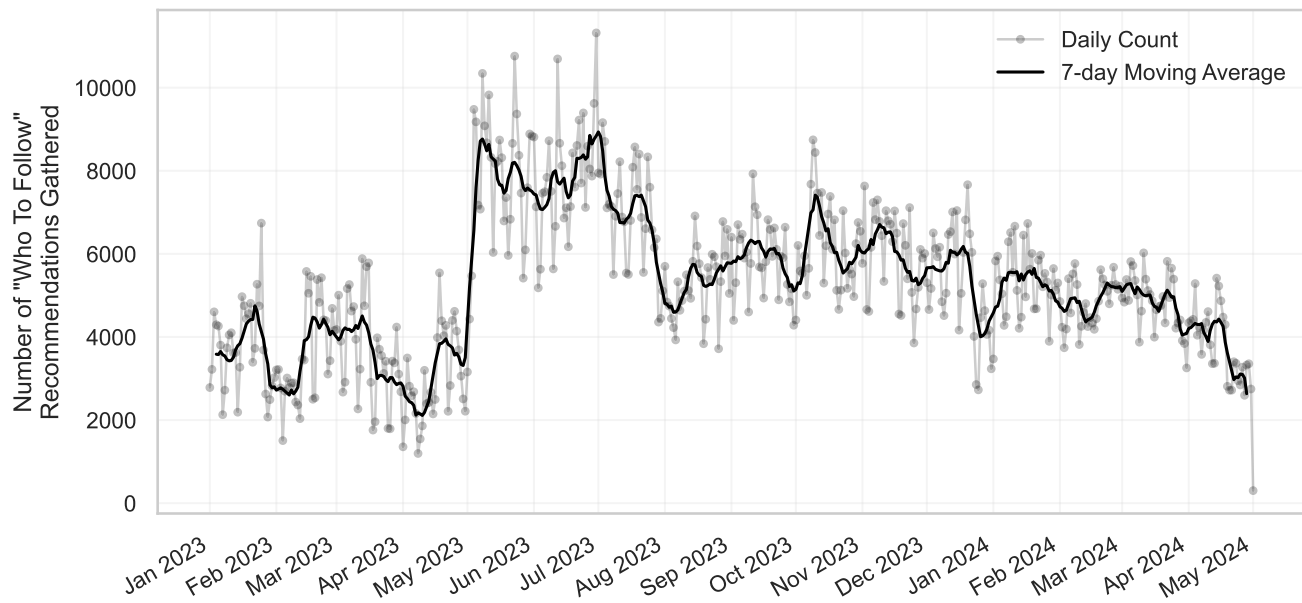


Figure S4: Evolution of the number of friend recommendations (WTF) seen by volunteers in our study and collected with our browser plug-in. The observation period of the study extended from January 1st, 2023, to May 1st, 2024. The number of collected recommendations depends on the number of volunteers participating in the project, which saw an increase in May 2024 due to the visibility of changes in content moderation and data accessibility brought about by the change in ownership of the platform. We include tests of accuracy and reliability tests in the inference of the embedding space, and that use different temporal windows in sampling, to demonstrate that the inference is robust throughout the whole period of observation.

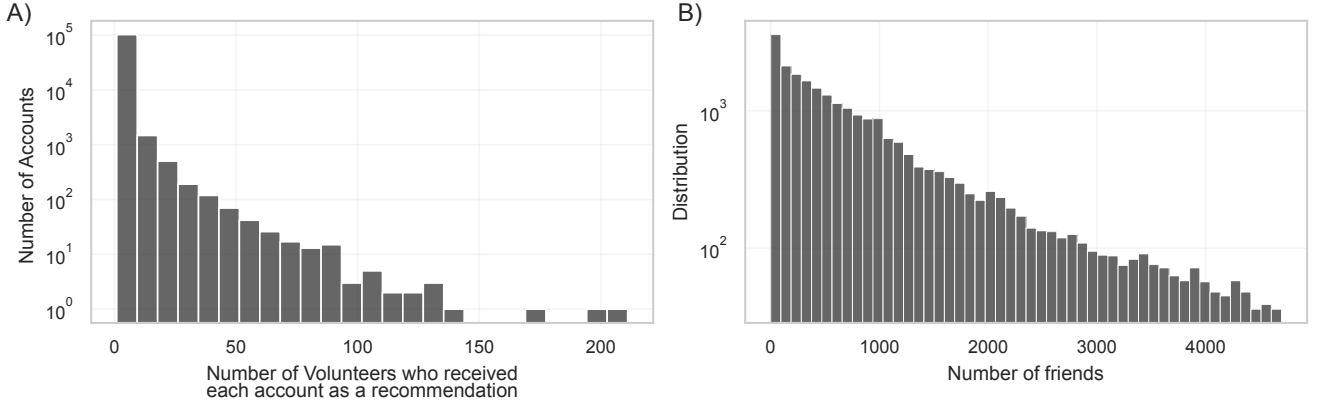


Figure S5: (A) Distribution of accounts based on how many volunteers received them as recommendations. (B) Distribution of the number of friends accounts for each user in $\mathcal{E}$.

Table S5: Most recommended accounts seen by our volunteers, with the rate or fraction of volunteers that saw the recommendation during the observation period.

Twitter Handle	Recommendation Rate	Description
@elonmusk	35.4%	Owner, CTO and Exec. Chairman of Twitter
@BFMTV	30.8%	French News Broadcast Channel
@MrBeast	28.1%	Popular YouTube Content Creator
@EmmanuelMacron	22.3%	President of France (2017-2027)
@GabrielAttal	20.7%	Former Prime Minister of France (2024)

Below, in Section C.2.3 we show specific experiments (experiment C.2.3.iii) aiming at testing the impact of, and ensuring that this exogenous factor does not affect the validity of our results.

Within the more than 2.5 million “Who to Follow” recommendations collected, we observed 105k unique Twitter accounts that were recommended, with a highly skewed distribution (see Extended Figure S5). For instance, while 74.5% of accounts were recommended to only one volunteer, Elon Musk’s account was recommended to 35.4% of volunteers who did not follow him already. Extended Table S5 presents the five accounts recommended to the highest fraction of non-following volunteers. Besides Twitters’ new owner, we find a major French news broadcast channel, a YouTube creator, followed by the French president and a former prime minister.

Overall, 26 338 unique Twitter accounts were recommended by Twitter to at least two volunteers. We will refer hereinafter to this set of users as $\mathcal{U}$.

B.2.2 Dataset of follower networks obtained with the platform’s API

To complement the “Who To Follow” profile recommendations collected through crowdsourcing, we systematically collected the list of friends of each user in $\mathcal{E} = \mathcal{V} \cup \mathcal{U}$ via the Twitter API. Among 26 509 users in $\mathcal{E}$, 24 918 had public accounts and at least one friend. Our volunteers $v \in \mathcal{V}$ followed an average of 661 accounts, with a median of 320. This collection allows us to integrate the follower network subtended by users in $\mathcal{E}$ in the loss function $\mathcal{L}$ for the inference of their embeddings ϕ .

We display in Figure S6 the distribution of degrees (number of Follow and Who to Follow relationships) for each user. In particular, we observe that 99.1% of accounts have a degree larger than 5, and 95.6% larger than 10.

B.3 Negative sampling

Combining our two datasets (described in Sections B.2.1 & B.2.2), we obtain the set of positive edges for the “Follow” and “Who to Follow” relationships, resulting in $|\mathcal{P}_{\text{Follow}}| = 4\,441\,670$ “Follow” edges and $|\mathcal{P}_{\text{WTF}}| = 110\,164$ “Who to Follow” edges. As we lacked access to the exact set of candidate accounts scored by Twitter’s algorithmic systems, from which the

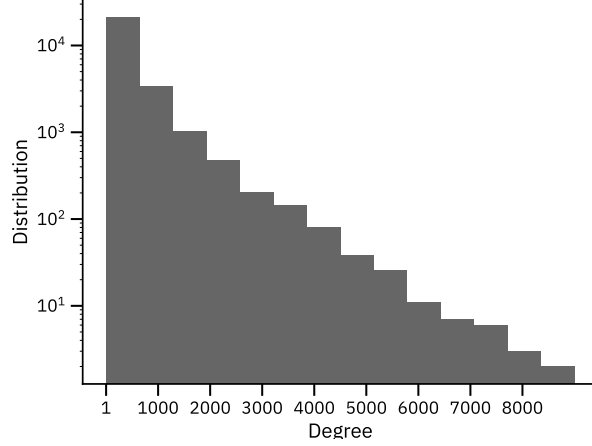


Figure S6: Distribution of accounts’ degree.

“Who to Follow” list is computed after ranking, we rely on negative edge sampling. Specifically, we generated negative edges by corrupting positive ones, replacing target accounts with randomly sampled accounts. Building upon established literature on Knowledge Graph embeddings (17) and the impact of negative sampling on model quality (18), we extended the negative sampling strategy of PyTorch-BigGraph (19), used by Twitter in (13).

Specifically, we combine three sampling methods in equal proportion:

- **Uniform sampling:** Target users are randomly selected from the entire pool of accounts with equal probability;
- **Prevalence-based sampling:** Target users are sampled with probability proportional to their frequency in the positive edge set, reflecting their natural occurrence in the recommendations;
- **Second-neighborhood sampling:** Target users are selected from friends of friends, i.e. being two hops away in the user-to-user network, capturing the local network structure.

Both uniform and prevalence-based sampling approaches present distinct limitations when used independently. Uniform sampling risks creating a model that learns to score edges primarily based on source and target node frequency rather than meaningful relationships between nodes. Conversely, prevalence-based sampling may not sufficiently penalize high scores assigned to rare entities. PyTorch-BigGraph (19), used by Twitter in (13), addresses these limitations by combining, in equal proportion, targets sampled using both strategies.

Informed by Twitter’s own algorithms (8), we complemented the negative sampling by incorporating non-recommended second-order friends and excluding first-degree connections (i.e. direct friends). The homophilic nature of social networks (4) (and in particular Twitter networks (20–22)) motivates the search for new ties from friends of friends rather than random strangers. We observe in Twitter’s algorithmic pipeline different candidate sources tied to this notion of homophily: either by computing similarity between accounts through Follow2vec, Twitter’s adaptation of Word2Vec over the user-user follower network (23), embedding closely users tied to similar accounts (24), or by explicitly looking for second-order connections through two-hop random walks (11) or second-order friends following the source users (25).

An analysis of the empirical “Who to Follow” recommendations reveals that 62.4% of recommended accounts are second-degree neighbors to the user receiving the recommendations. Among recommended second neighbors, only 15.8% are followed by a single of the user’s friends, while half are followed by at least 8 friends. The degree distribution shown in Extended Figure S7 showcases that recommended second-degree neighbors tend to have substantially more first-degree friend connections compared to non-recommended ones.

Consequently, prevalence sampling from non-recommended second-order friends provides particularly valuable negative examples, as these represent plausible but ultimately non-recommended connections —effectively serving as “hard negatives” for the model.

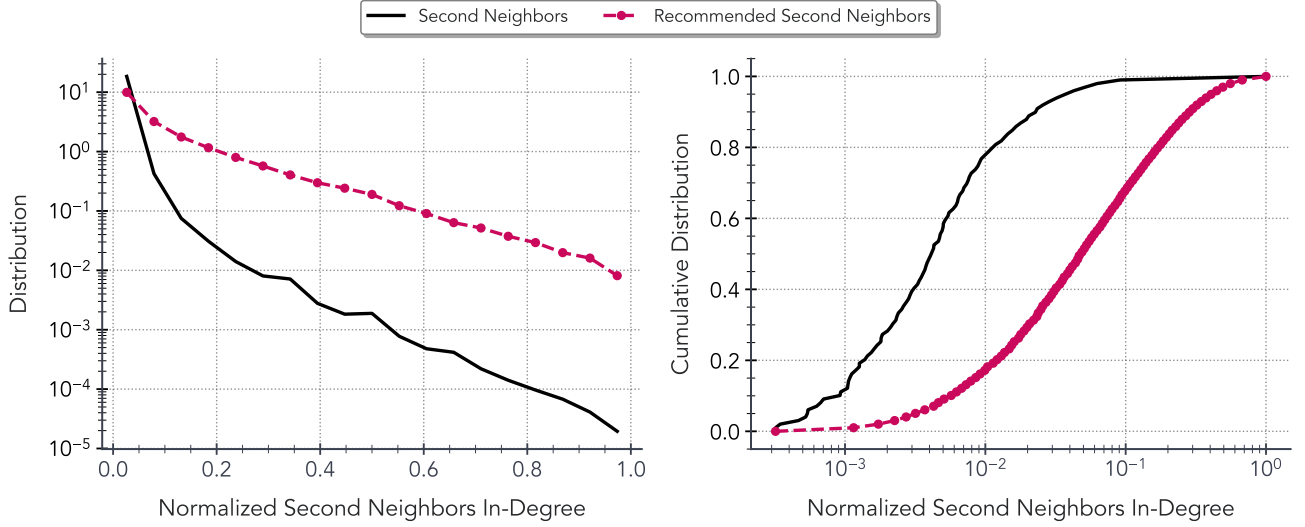


Figure S7: Analysis of the degree distribution of second-degree neighbors. Distribution (left figure) and Cumulative Distribution (right figure) of the in-degree of second-degree neighbors, with respect to first-degree neighbors, normalized per volunteer by the largest in-degree. The distribution for recommended second-degree neighbors is shown as a dashed red line.

Finally, we curate the negative edge sets $\mathcal{N}_{\text{Follow}}$ and $\mathcal{N}_{\text{WTF}}$ by combining uniform, prevalence-based, and second-neighborhood sampling methods in equal proportions. We adopt a negative-to-positive ratio of 3:1 for both “Follow” and “Who to Follow” relationships, ratio of the same magnitude than literature and industry standards (15), resulting in $|\mathcal{N}_{\text{Follow}}| = 3|\mathcal{P}_{\text{Follow}}|$ and $|\mathcal{N}_{\text{WTF}}| = 3|\mathcal{P}_{\text{WTF}}|$.

B.4 Training an approximation of the embedding of X’s recommender system

B.4.1 Implementation

Following Twitter’s implementation (14), we trained 256-dimensional users and relationship embeddings under the loss function defined in Equation (7). The model parameters were Glorot initialized (26), in order to prevent vanishing or exploding gradients during training, and optimized with Adagrad (27) in PyTorch.

We randomly segmented the curated edge set into 80%, 10%, and 10% splits for training, hyperparameter tuning, and final evaluation. The splitting was performed independently for each relation type to maintain consistent Follow/WTF ratios across all sets.

B.4.2 Tuning hyperparameter α in loss function $\mathcal{L}$

To optimize prediction of “Who to Follow” edges, we tuned the hyperparameter $\alpha \in [0, 1]$ from loss function on a 10% holdout validation dataset. The parameter α in Equation 7, balances in the loss the contributions of profile recommendations and “Follow” relationships. Setting $\alpha = 1$ trains the model exclusively on “Who to Follow” data, while $\alpha = 0$ uses only “Follow” relationships.

Due to the nondeterministic nature of gradient descent, we trained the model ten times for each α value over three epochs, evaluating the ROC AUC on the validation dataset for each replicate. We observe on Extended Figure S8 that the AUC-ROC —measuring the probability that the model ranks a randomly chosen positive instance higher than a randomly chosen negative one— increases with α until reaching a peak of $.7562 \pm .0018$ (standard deviation over ten replicates) at $\alpha = 62.6\%$, after which it declines.

The relative scarcity of profile recommendation data, especially in comparison with follow relationships, limits predictive power when training solely on user recommendations, i.e. $\alpha = 1$. The tuning of α demonstrates the benefits of heterogeneous information network embeddings, even for single-entity-single-relationship downstream tasks such as account-to-account follow recommendations, corroborating Twitter’s observations (13) that guided their implementation (8).

In what follows, all training uses the optimal value $\alpha = 62.6\%$, which assigns each “Who to Follow” training edge 1.67 times the weight of each “Follow” edge. Also, we trained the model over eight epochs, and three independent runs, measuring

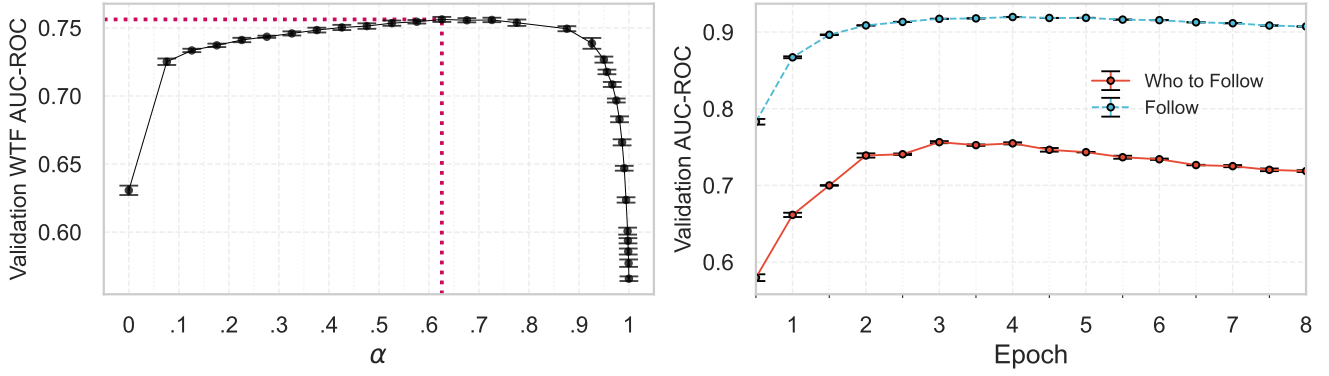


Figure S8: Accuracy of the trained algorithm and its embedding space in predicting observed data for different training parameters. (Left figure) AUC-ROC scores for “Who to Follow” predictions as a function of α , with error bars showing standard deviation across ten replications. (Right figure) AUC-ROC values for “Who to Follow” (in orange) and “Follow” (in cyan) predictions versus training epoch

the AUC-ROC on the validation dataset every half epoch. As shown in Extended Figure S8, the optimal number of epochs was determined to be three.

C Validating the inferred embeddings Φ for users

In this section, we provide different evidence to the quality of these embeddings. First, in Section C.1 we validate our computed embeddings by showing that, if used for recommendation of friends using X’s prescribed recommender system methods, they are able to predict the users that were recommended with high accuracy. Then, in Section C.2, we show that these embeddings are robust under a diversity of sub-sampling strategies addressing possible biases that might have been introduced by our research design.

C.1 Testing the performance of embeddings in predicting recommended users

After training for three epochs on a combination (1:3) of positive and negative edges from “Who to Follow” recommendations and follower relationships, weighted (1.67:1), we learned $\Phi \in \mathbb{R}^{|\mathcal{E}| \times 256}$, the accounts embeddings, along with “Who to Follow” φ_{WTF} and “Follow” φ_{Follow} relationships embeddings. Our evaluation approach, while constrained by our lack of definitive negative set of candidates considered by Twitter’s model, demonstrates strong predictive performance.

On a 10% held-out test set, the learned embeddings achieve an AUC-ROC of 70.0% for predicting “Who to Follow” recommendations when using non-recommended second-order friends as negative edges. When evaluating positive test recommendations against a pool of randomly sampled 100 non-recommended second-order friends, the model achieves a precision@1 of 72.5%. Matching Twitter’s “Who to Follow” interface which typically displays slate of three recommendations, we evaluate precision@3, reaching 69.1%, while precision@10 equals 64.3%.

We emphasize a key distinction in our evaluation: unlike standard recommendation systems that predict user reactions to suggestions, our model predicts Twitter’s recommendation pipeline outputs. This shift in objective—predicting model behavior rather than user behavior—may explain our model’s strong performance relative to literature benchmarks (28).

Overall, the X account embeddings Φ , learned following the delineated approach, successfully predict Twitter’s “Who to Follow” recommendations, indicating that studying the inferred embedding space provides a good approximation for Twitter’s internal representation of entities.

C.2 Testing the robustness of embeddings Φ under alternative data sub-sampling and model parameter strategies

The learned user embeddings Φ , hereinafter named *reference embeddings*, depend on both the collected data and the training procedure described above. Abusing notation, we denote by $\tilde{\Phi}$ different embeddings that result from considering different data sub-sampling strategies and different hyper-parametrizations. First, in Section C.2.1 we specify how we compare alternative embeddings $\tilde{\Phi}$ with the reference embedding Φ . While the training procedure was optimized to maximize the model’s predictive performance for “Who to Follow” relationships between accounts, the stability of the resulting embeddings under various perturbations remains crucial for drawing reliable conclusions. Consequently, we conducted two types of robustness checks, training embeddings $\tilde{\Phi}$ (C.2.2) with alternative training parameters, and (C.2.3) on alternative datasets.

C.2.1 Evaluating similarity of embeddings under affine transformations

Considering our subsequent analysis, identifying interpretable directions in the embedding space through canonical correlation analysis (CCA; see more details below in Section E), we assess the robustness of learned embeddings up to nonsingular affine transformations, as CCA exhibits invariance to such transformations (29). We measure the coefficient of determination, R^2 , between the reference embeddings Φ and each variant $\tilde{\Phi}$ after applying the optimal nonsingular affine transformation determined through least squares optimization.

C.2.2 Testing the robustness of embeddings for alternative training parameters

Experiment C.2.2.i) Different number of epochs. We performed 10 training runs using the original training dataset, saving the learned representations $\tilde{\Phi}$ after 3, 5, and 10 epochs. Reference embeddings Φ were learned at 3 epochs, as determined optimal on the validation dataset.

Results of C.2.2.i) The coefficients of determination, R^2 , between reference embeddings Φ and linearly transformed variants $\tilde{\Phi}$ show consistent stability across non-deterministic training runs, with highest values at 3 epochs ($R^2 = .84759 \pm .00052$). Embedding stability decreases thereafter, with R^2 values dropping at 5 epochs ($R^2 = .83757 \pm .00062$) and 10 epochs ($R^2 = .77506 \pm .00066$). This decline reflects a shift from learning the overall network structure to noisy over-fitting the training data, as evidenced by the validation metrics in Extended Figure S8. This interpretation is further supported by the reduced consistency between embedding replicates —the average pairwise R^2 between 10 replicates trained for 10 epochs drops to $.6410 \pm .0013$, compared to $.8496 \pm .0018$ for embeddings learned after 3 epochs. Yet, overall, reference embeddings Φ do explain a substantial fraction of variance in variant embeddings $\tilde{\Phi}$, even after extended training.

Experiment C.2.2.ii) Intermediate values for α . We performed five training runs on the original training dataset for 3 epochs, considering two additional α values. While the reference embeddings Φ were learned with $\alpha = 62.6\%$ (optimal value on the validation dataset), alternative embeddings $\tilde{\Phi}$ were trained with $\alpha = \{42.6\%, 87.6\%\}$, achieving validation AUC-ROC scores of $.7503 \pm .0019$ and $.7494 \pm .0019$ respectively (standard deviation over ten replicates).

Results of experiment C.2.2.ii) The average coefficient of determination, R^2 , between reference embeddings Φ and variants $\tilde{\Phi}$ was $.8417 \pm .0041$ for $\alpha = 42.6\%$ and $.7762 \pm .0061$ for $\alpha = 87.6\%$ (standard deviation over five replicates). The high R^2 values indicate that embeddings learned with different α values remain structurally similar, up to non-singular linear transformations, to the reference embeddings Φ , despite slightly suboptimal validation performance.

Experiment C.2.2.iii) Training only on recommendation ($\alpha = 1$). We trained for 3 epochs considering, from the original dataset, only “Who to Follow” recommendation signal, excluding “Follow” signal i.e. $\alpha = 1$. Beyond R^2 , and considering the particular status of $\alpha = 1$, we performed the entire subsequent analysis, finding interpretable directions in the embedding space aligned with features of interest.

Results of experiment C.2.2.iii) Through Canonical Correlation Analysis, we found in the new embedding space directions that correlated with the features of interest, yet to a lesser extent than for the optimal embedding Φ . Specifically, the Pearson correlations between the optimal direction and the features are: 0.35 for Left-Right leaning, 0.34 for Anti-Elite Rhetoric, 0.15 for Male-Female, 0.26 for Age, 0.42 for Popularity (log of number of followers), and 0.32 for Interest in News. These results confirm that even when training solely on the observed recommendations, the embeddings encode users’ political leaning.

Experiment C.2.2.iv) Alternative negative/positive edges ratio. Rather than the 3-1 ratio between negative and positive edges, we train embeddings with a 1-1 ratio, randomly subsampling the initial training set.

Results of experiment C.2.2.iv) Through Canonical Correlation Analysis, we found in the new embedding space a direction that correlates with the Left-Right political alignment, the variable of interest in this article, to a comparable measure with the original embeddings. In particular, we observe a strong correlation of $\rho = 0.802$, $p < 10^{-4}$.

Experiment C.2.2.v) Alternative negative edges sampling strategy Rather than the sampling strategy for the negative edges used in the original setup, we experiment with a random baseline. Specifically, negative edges are randomly sampled between users, under the only constraint of not being a positive one.

Results of experiment C.2.2.v) Through Canonical Correlation Analysis, we found in the new embedding space a direction that correlates with the Left-Right political alignment, the variable of interest in this article, to a comparable measure with the original embeddings. In particular, we observe a strong correlation of $\rho = 0.783$, $p < 10^{-4}$.

Experiment C.2.2.vi) Alternative degree threshold Rather than considering accounts recommended to at least two volunteers, we increase the threshold to assess the robustness of our results and consider only those recommended to at least 10 volunteers.

Results of experiment C.2.2.vi) Through Canonical Correlation Analysis, we identified in the new embedding space a direction that correlates with Left-Right political alignment at a level comparable to the original embeddings, with a strong correlation of $\rho = 0.807$, $p < 10^{-4}$.

C.2.3 Testing the robustness of embeddings for different data sub-sampling strategies

In this section, we consider four possible sources of biases in our collected served recommendations and address them by comparing the embeddings Φ with alternative embeddings $\tilde{\Phi}$ computed with subsamples addressing these potential sources.

Experiment C.2.3.i) Users seeing recommendations outside our experiment (on other devices). The extension only captures recommendations displayed on Twitter’s desktop interface, meaning volunteers may have been exposed to more recommendations than those collected. From the profile recommendations gathered on desktop, we observe that each account is suggested 5.11 times on average to the same volunteer (computation limited to the 435 volunteers who viewed at least 25 “Who To Follow” triplets in their feed). Inspecting Twitter’s codebase reveals no fundamental reason why mobile profile recommendations would significantly differ from desktop ones. Therefore, we postulate that most missed recommendations were redundant with desktop ones and thus already included in the training dataset. Nevertheless, as we cannot, by construction, assess this observational blind spot, we trained embeddings on corrupted training datasets to validate our findings.

Assuming volunteers spend 50% of their Twitter consumption on desktop browsers where we collect profile recommendations, and 50% on other devices, and assuming that the pool of recommended accounts is shared across devices, we estimate that we may miss up to 19.6% of unique accounts recommended to volunteers. To simulate this effect, we augmented each volunteer’s dataset by adding 19.6% of *artificial non-observed* positive edges, randomly sampled following the same rationale as for the previous curation of negative edges. We trained variant embeddings $\tilde{\Phi}$ using identical hyperparameters as reference embeddings Φ (3 epochs, $\alpha = 62.6\%$) on 10 random samplings of additional “Who to Follow” edges.

Results of experiment C.2.3.i) The average coefficient of determination, R^2 , between variant embeddings $\tilde{\Phi}$, linearly transformed, and the reference embeddings Φ was $.8417 \pm .0033$. In a more extreme scenario where users spend 75% of their Twitter consumption on non-observed devices —where we would be missing potentially 39.2% of unique account recommendations, assuming cross-device recommendation pools and accounts recommended 5.11 times on average to volunteers— the R^2 between reference embeddings and 10 variants trained with 39.2% artificial non-observed positive edges remained high at $.8177 \pm .0031$. These high R^2 values demonstrate the robustness of the reference embeddings Φ to missing observations.

Experiment C.2.3.ii) Addressing population biases. A second limitation of our “Who to Follow” recommendation dataset arises from volunteer self-selection, creating a sample non-representative of X’s overall user base. Our data donation initiative was advertised through mainstream media, including a call for participants in LeMonde (a major French newspaper), radio coverage on France Culture (a French public radio), and X advertisements. During extension installation, volunteers could optionally indicate their age, gender, and political preference on a continuous scale from far-left to far-right. Extended Figure S9 presents these survey results as both raw data and weighted by the number of “Who to Follow” recommendations gathered per user. While our cohort spans the full spectrum of age, gender, and political leanings, respondents tended to be left-leaning, males, over 35.

To evaluate demographic skew impact, we constructed alternative training datasets of profiles suggested to subsets of volunteers who completed the survey. We applied our previous curation filters to each subset, including only accounts suggested by at least two volunteers while incorporating their follow relationships. Sample size limitations restricted our analysis to single demographic variables rather than intersections, as multi-dimensional stratification would yield insufficient subsets.

Using identical hyperparameters as for the reference embeddings Φ (3 epochs, $\alpha = 62.6\%$), we trained variant embeddings $\tilde{\Phi}$ over the following four datasets:

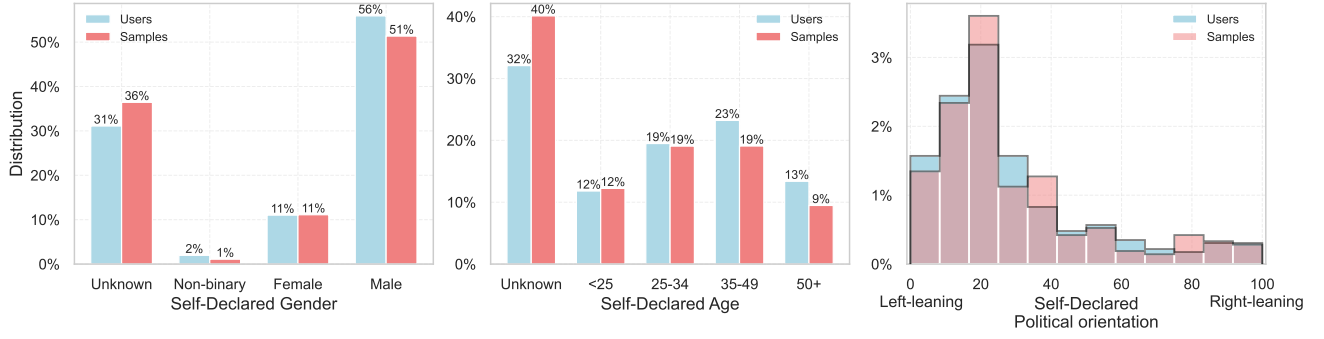


Figure S9: Demographic and political characterization of volunteers in the study using surveys administered at enrollment via the browser plug-in. Distribution of volunteers’ self-declared demographics (in blue) and weighted by “Who to Follow” samples collected per volunteer (in red), including gender, age, and political leaning. “Unknown” category reflects unfilled optional survey fields, 45.9% of volunteers did not disclose a political preference.

- 284 self-identified male volunteers, yielding 49 319 “Who to Follow” recommendations and 11 454 embedded accounts.
- 186 volunteers aged over 35, yielding 23 868 “Who to Follow” recommendations and 6 542 embedded accounts.
- 218 Left-leaning volunteers (scoring below 40 on the political spectrum), yielding 37 042 “Who to Follow” recommendations and 9 132 embedded accounts.
- 57 Right/far-Right leaning volunteers (scoring above 65 on the political spectrum), yielding 5 997 “Who to Follow” recommendations and 1 947 embedded accounts.

Results of experiment C.2.3.ii) Linear transformation of embeddings learned from male-volunteer induced datasets achieved a coefficient of determination, R^2 , of .654 with reference embeddings Φ . Left-leaning-volunteer and above-35-volunteer induced datasets yielded R^2 values of .611 and .552, respectively. The R^2 values contrast notably with higher values obtained from variations in training parameters and observational data, highlighting the importance of demographic information encoded in the learned account representations. Finally, even when considering a narrow, highly skewed cohort of only Right/far-Right leaning volunteers, the learned embeddings encode Left-Right political leaning. Through a CCA we find a direction whose user projection onto it correlates with left-right leaning with a Spearman correlation of $\rho = 0.76, p < 10^{-4}$. Overall, we note that the reference embeddings Φ learn from the complete volunteer cohort—spanning diverse gender, age, and political orientations—captured significant structure of the variant embeddings $\tilde{\Phi}$, maintaining R^2 values above .55. We analyze the demographic information encoded in these learned account representations in the following sections.

Experiment C.2.3.iii) Heterogeneity of the population across the period of observation. Our dataset of recommendations spans a 16-month collection period, with varying collection rates as shown in Extended Figure S4. A notable increase in “Who to Follow” recommendations occurred in May 2023. To assess potential temporal effects, we split the dataset into two periods: recommendations collected before and after May 1st, 2023. We then learned embeddings using our previously described methodology.

Results of experiment C.2.3.iii) The embeddings learned from these temporal splits show strong alignment with the reference embeddings Φ , achieving R^2 values of .784 and .773 for the pre-May and post-May 2023 periods, respectively.

Experiment C.2.3.iv) Changing follower graph during the period of observation. An additional consequence of our extended observational window is that the volunteer pool of friends changed over time. In our training dataset, we aggregated all accounts followed by volunteers, irrespective of when they started following them, with one notable exception: if a “Who to Follow” suggestion turned into a friend, we solely considered the Who to Follow observation, as the signal was scarcer for these cases. Consequently, some accounts in the training dataset labeled as “friends” were not followed when the first Who to Follow recommendations were collected. Similarly, some accounts may no longer be friends by the experiment’s end if users unfollowed them. To assess how robust our embeddings Φ are to temporal dynamics, we tracked volunteers’ friend lists throughout the experiment, allowing us to derive friend addition and removal rates. This revealed that, at most, 36.9% of friends in the training set were either a) not yet followed (26.7%) at the experiment’s start 16 months ago or b) removed from

the friend pool (10.2%) since then. We then randomly removed 36.9% of friends from the training dataset and learned variant embeddings $\tilde{\Phi}$, repeating this process three times.

Results of experiment C.2.3.iv) Linear transformation of embeddings $\tilde{\Phi}$ learned from those variant training dataset, achieved a coefficient of determination, R^2 , of $.80217 \pm .00028$ (standard deviation over three replicates) with reference embeddings Φ .

Our experiments demonstrate that the reference embeddings Φ remain stable (modulo linear transformations, as described in Section C.2.1) despite substantial changes in training parameters and dataset composition.

Experiment C.2.3.v) Removing ties between users and Members of Parliament The ideological leaning of users is inferred from following patterns with members of parliament, and because such follow relations appear in the training dataset — specifically 2.9% of edges in the training dataset are follow edges between members of parliament and users whose ideology has been scaled— we assess the robustness of our results to the removal of such ties. We explicitly remove such ties in this experiment and then learn the embeddings using our previously described methodology.

Results of experiment C.2.3.v) A CCA between the new embeddings and the Left-Right political leaning of users results in a Spearman correlation of $\rho = 0.816$, $p < 10^{-4}$. In comparison, as reported in the main text, we observe a correlation of $\rho = 0.887$, $p < 10^{-4}$ when considering the full training dataset.

Experiment C.2.3.vi) Removing ties between volunteers. Because our method of recruitment could allow snowballing effects to emerge, e.g., a volunteer recommending a friend to take part in the experiment, in this experiment we remove the ties (follow or WTF recommendation) between volunteers, removing less than 0.03% of edges. We then learn the embeddings using our previously described methodology.

Results of experiment C.2.3.vi) A CCA between the new embeddings and the Left-Right political leaning of users results in a Spearman correlation of $\rho = 0.829$, $p < 10^{-4}$. In comparison, as reported in the main text, we observe a correlation of $\rho = 0.887$, $p < 10^{-4}$ when considering the full training dataset.

D Inferring users’ features

Having learned embeddings Φ for each X user in $\mathcal{E}$, we now curate a set of user features that we will later leverage in attributing semantics to this embedding space. Specifically, we define the following families of user features: D.1) user activity metrics (e.g., popularity, number of posts and likes), D.2) demographic features, D.3) topics of interest and linguistic features, and D.4) ideological leanings of users.

D.1 User activity metrics

Using X’s API, we collect account metadata for each user in $\mathcal{E}$:

- *Follower count*: measures social capital, ranging from 0 to 197 million with a median of 7,987 followers.
- *Follower-to-friend ratio*: assesses network reciprocity, with a median of 0.097.
- *Daily Posts count*: user total posts (status) divided by days since account creation, with a median of 2.21 posts per day.
- *Daily Likes count*: user total likes divided by days since account creation, with a median 1.40 of likes per day.

These metrics, whose distributions are shown in Extended Figure S10, span multiple orders of magnitude. We normalized each metric as the fraction of users in $\mathcal{E}$ with strictly higher values, yielding scores in $[0,1]$.

Leveraging an independent data collection from early 2023, we examined the Legacy verification status of 24 214 users in $\mathcal{E}$. This status, formerly known as the blue check-mark, indicated whether an account was “authentic, notable, and active” (30), typically granted to public figures, journalists, and influential organizations before the platform’s verification policy changes. Up to 30.1% of the accounts were, as of early 2023, verified.

D.2 Inference of demographic features

Leveraging account metadata collected for Twitter activity metrics, we analyze profile pictures, screen names, usernames, and biographical descriptions to infer accounts’ demographic attributes. Specifically, we applied the Multimodal, Multilingual, and Multi-attribute (M3) model, developed by Wang et al. (31), on the 24 425 accounts in $\mathcal{E}$ with available information. We display on Extended Figure S11 the gender and age distributions of profiles inferred by M3.

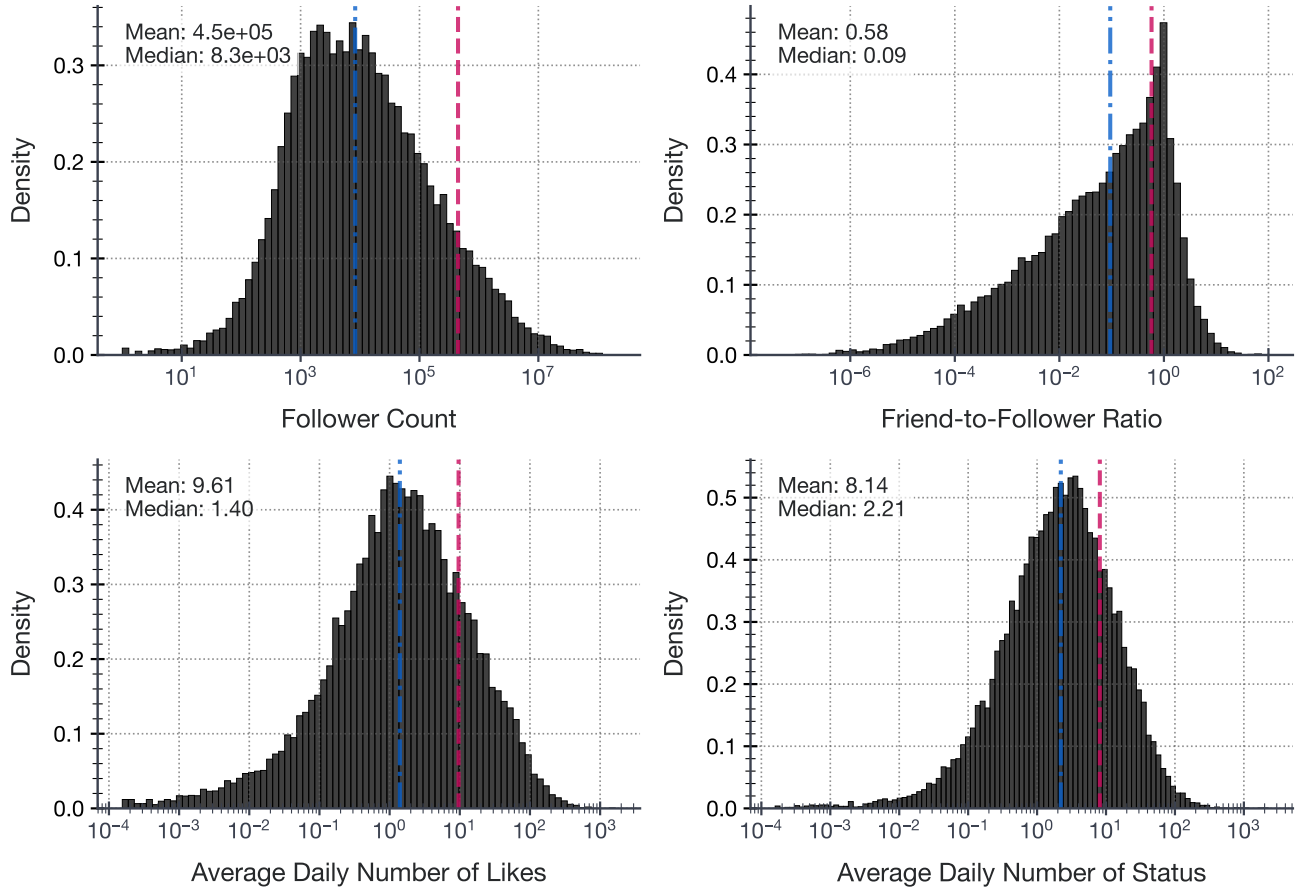


Figure S10: Distribution of X users in our study by attributes related to activity on the platform, on a logarithmic scale, with mean values indicated by dashed red lines and median by dashed blue lines: follower count (top left), friend-to-follower ratio (top right), average daily number of likes (bottom left), and average daily number of posts/statuses (bottom right).

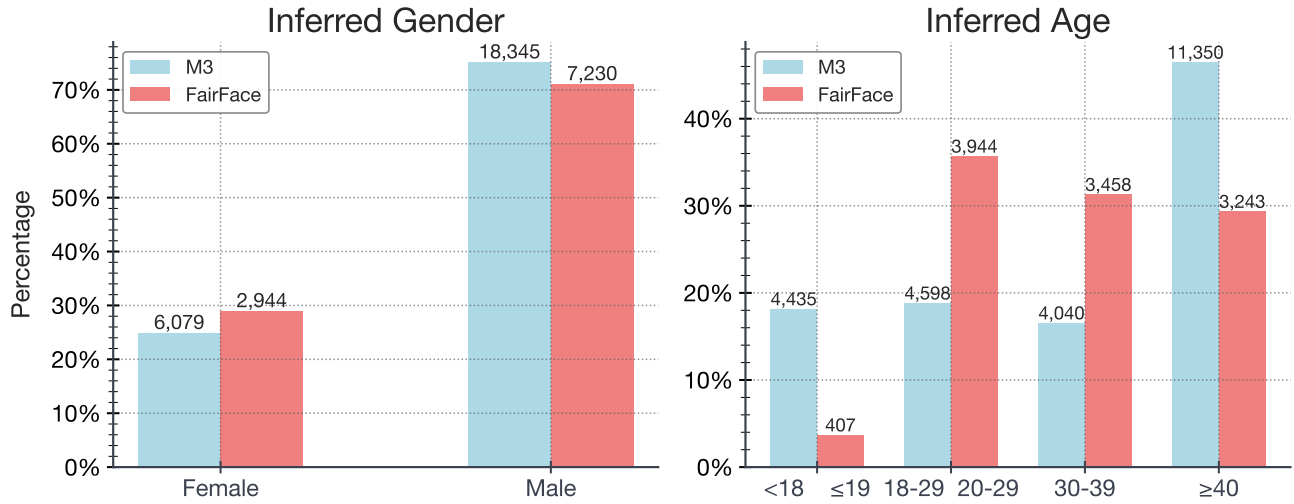


Figure S11: Comparison of the inference of age and gender for users in the study (using M3 model by Wang et al., (31)), with an alternative method of inference, the FairFace model (developed by Karkkainen and Joo (32)), for robustness. Gender and age distributions predicted by M3 (in blue) for 24 425 accounts in $\mathcal{E}$ and by FairFace (in red) for the subset of 10 932 accounts with a *single* visible profile face. One limitation of our study lies in the fact that users may choose profile names, descriptions, and photos that do not match their demographic groups, choosing, for instance, avatars (e.g., cartoons, historic personalities), obfuscated representations (e.g., disguises, photos where they do not appear clearly), or even misleading representations (e.g., seeking to appear to other users as being from a demographic group other than their own). We argue that this limitation does not impact the main findings of our study, as these occurrences do not affect the overall trend between features in the absence of reason to believe that their frequency should be dependent on the main variables against which we test independence to demographic features: namely, political leaning.

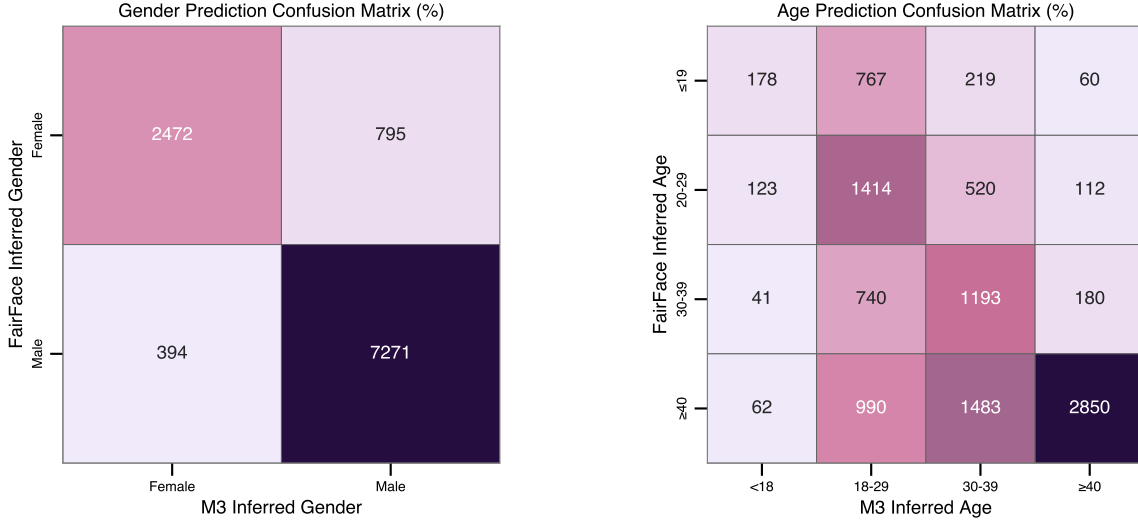


Figure S12: Confusion matrices comparing the demographic inference achieved by M3 and by FairFace models on users in our study, for gender (left figure) and age (right figure) for 10 932 accounts with visible profile faces in $\mathcal{E}$.

Furthermore, as to corroborate M3’s inference, we estimate the age and gender of 10 932 accounts in $\mathcal{E}$ having a *single* visible face in the profile pictures. These estimates were derived through the FairFace model, developed by Karkkainen and Joo (32). This model was selected for being open-weights and offering state-of-the-art performance. Moreover, its curated training dataset enables consistent performance across race and gender groups (32), contrasting with typical closed-source solutions (33, 34). We show the confusion matrices for the inference of gender and age in Extended Figure S12.

Finally, we compare the inferred age and gender with volunteers’ (optional) self-declarations. We observe an overall alignment, with an F1 score of 0.84 for gender and an average F1 score of 0.82 for age buckets.

One limitation of our study lies in the fact that users may choose profile names, descriptions, and photos that do not match their demographic groups, choosing, for instance, avatars (e.g., cartoons, historic personalities), obfuscated representations (e.g., disguises, photos where they do not appear clearly), or even misleading representations (e.g., seeking to appear to other users as being from a demographic group other than their own). We argue that this limitation does not impact the main findings of our study, as these occurrences do not affect the overall trend between features in the absence of reason to believe that their frequency should be dependent on the main variables against which we test independence to demographic features: namely, political leaning.

D.2.1 Gender

For the 10 932 accounts where FairFace detected a *single* face, the gender inferred by FairFace strongly agrees with M3’s inference with a Cohen’s $\kappa = .731$.

Automated gender inference systems tend to disproportionately misclassify individuals that do not identify as male or female (35). We interpret these inferred gender as an algorithmic extension enabling large-scale aggregated trends in demographic analysis rather than definitive individual gender identification, i.e., a prediction of the most probable binary gender label that a team of annotators (akin to that involved in the training of M3) would provide if prompted to do so.

D.2.2 Age

Having aligned FairFace age bins into the closest M3 stratification levels (< 18 from M3 mapped to the ≤ 19 from FairFace, 18-29 from M3 mapped to 20-29 from FairFace), we observe a poor agreement between both models with Cohen’s $\kappa = .326$.

This poor agreement reflects the models’ distinct objectives: FairFace estimates age from profile pictures, while M3 infers account owners’ ages using also text description from the profiles. Manual inspection reveals accounts displaying children

receive young age estimates from FairFace but young-parent age ranges from M3 (average first-time parent age was 28.5 in 2015 (36) corresponding to the second class bin in M3 classification). Similarly, profiles classified as ≥ 40 years by FairFace but 18-29 by M3 displayed celebrities like Elon Musk or Ronaldo on their profile picture. Given M3’s specific training for social media demographics, we utilize its predictions for subsequent analyses.

The validity of M3’s age categorization is supported by distinctive patterns in account characteristics, with each age group showing characteristic median account creation years (≤ 18 : 2018; 19-29: 2015; 30-39: 2013; ≥ 40 : 2013) and terms used in users’ biographical descriptions. We report distinctive vocabulary by extracting the 10 words, longer than 5 characters, with the highest chi-squared statistics, comparing their frequency in each age bin to their overall use in the population $\mathcal{E}$ (see Extended List S1).

Extended List S1. *Distinctive vocabulary of age groups computed by extracting the 10 words, longer than 5 characters, with the highest chi-squared statistics, comparing their frequency in each age bin to their overall use in the population $\mathcal{E}$.*

≤ 18 : YouTube, Twitch, Player, Minecraft, Streaming, Videos, Discord, Youtuber, Streamer, Gaming
 19-29 : Student, Player, Professional, Enthusiast, Artist, Contact, Candidate, Engineer, Creator, Feminist
 30-39 : Journalist, Doctor, Opinions, Co-Founder, Manager, Adjoint, Director, Teacher, Senior, Redactor
 ≥ 40 : Official Account, President, France, MP, Association, University, Politics, Minister, District, Director

D.3 Content-based attributes

Beyond activity metrics and demographic inference from Twitter profiles metadata, we characterized user content production patterns. On 27th October 2024, we collected the last 200 tweets (published or retweeted) from 25 080 users in $\mathcal{E}$ through the Twitter API, excluding private and deleted accounts, totaling 2.3 million tweets.

This additional dataset enabled computation of content-related metrics. Specifically, we leveraged the multi-topic classification model `cardiffnlp/tweet-topic-21-multi` developed by Antypas et al (37), to categorize tweets into 19 generic topics including arts & culture, news & social concern, and science & technology. Non-English tweets were translated using Opus-MT models (38, 39), selected for their strong validated performance and their open-weights nature ensuring reproducibility. We calculated the proportion of tweets per topic for each user across all 19 categories considered by (37).

We display on Extended Figure S13 the distribution of those metrics, observing that “News & Social Concern” is the most discussed topic among embedded users, with some exclusively discussing this topic. The following main topics in decreasing frequency are “Diaries & Daily Life”, “Sports”, “Science & Technology” and “Business & Entrepreneurs”.

To ease topics interpretation and corroborate their validity, we report in Extended List S2 below the 25 most overrepresented words in tweets classified under each topic, compared to their expected frequency under uniform random distribution across topics. Extended List S2 contains only 4+ letter words (lower-cased and translated into English when required using Opus-MT model (39)).

Extended List S2. *25 most overrepresented words in tweets classified under each topic, compared to their expected frequency under uniform random distribution across topics from (37). List contains only 4+ letter words (lower-cased and translated into English when required using Opus-MT model (39)).*

- **Arts & Culture:** paintings, feast, theatre, artworks, painter, tolkien, ballet, illustrations, exhibitions, retrospective, exhibit, nouveau, angoul, musicians, animations, shakespeare, arts, gogh, poem, curator, biography, treasures, photographs, exhibition, festival
- **Business & Entrepreneurs:** binance, staking, paypal, multinationals, pharmaceutical, cryptocurrency, economies, ceos, sellers, subsidiary, remuneration, investors, mckinsey, consulting, profitability, outlook, investor, margin, defi, stablecoin, blackrock, eligible, founders, banking, trading
- **Celebrity & Pop Culture:** lana, celine, portman, hathaway, meryl, grammy, zendaya, vmas, cotillard, starring, billie, perry, cobain, hollywood, eminem, brad, pitt, natalie, dion, thee, cooper, comedian, andrew, lipa, scarlett
- **Diaries & Daily Life:** subway, stayed, shower, crush, laundry, habits, sagittarius, cashier, eats, darling, toilet, pisces, rude, reflexes, wakes, kindness, jokes, notification, fever, conditioning, kisses, asleep, sleep, boyfriend, girlfriends

- **Family:** *dads, grandparents, mothers, cousin, eidmubarak, parent, daughters, moms, niece, mubarak, sister, grandmother, widows, grandpa, adha, daughter, father, mother, fathers, aunt, uncle, uncles, sisters, nephew, sibling*
- **Fashion & Style:** *fashion, jewelry, dior, leather, wardrobe, clothing, outfit, makeup, necklace, nike, pants, clothes, heels, accessories, bracelet, runway, lingerie, christianlouboutin, outfits, shirts, dress, perfume, vinted, grazia, shoe*
- **Film Tv & Video:** *batman, emmy, deadpool, succession, saga, starring, biopic, nolan, superman, amour, slayer, ahsoka, christopher, uploaded, toriyama, anatomy, pixar, loki, runner, wick, titanic, arcane, theaters, rings, venom*
- **Fitness & Health:** *injections, pediatric, cardiovascular, endometriosis, resuscitation, addictions, parkinson, chronic, fitness, therapeutic, respiratory, cardiac, autism, cancers, surgery, vaccinated, worldmentalhealthday, dental, injection, disorders, liver, oncology, masks, transplant, practitioner*
- **Food & Dining:** *pastry, bakery, ingredients, foods, cook, culinary, sauce, flavors, meal, juice, gourmet, burger, meals, cookies, vegetables, cooking, cuisine, kebab, chefs, eats, recipe, cheese, restaurants, cocktails, pizzas*
- **Gaming:** *warcraft, ubisoft, indiegame, warhammer, xbox, gameplay, fallout, warzone, kart, doom, skins, tabletop, pokemon, mario, nintendoswitch, gaming, overwatch, gamedev, sims, playstation, fortnite, playable, creed, zelda, gear*
- **Learning & Educational:** *classroom, graduates, orientation, bachelor, genetet, schooling, aesh, autism, sociology, graduate, maths, vocational, fellowship, interdisciplinary, nursing, diploma, exams, principal, supervisors, college, dean, trainers, doctoral, baccalaureate, teacher*
- **Music:** *musicians, lana, electro, guitar, piano, sings, celine, vocals, tyler, eminem, billboard, grammy, concert, tiakola, tunes, singles, hiphop, eurovision, billie, music, olympia, concerto, vinyl, lipa, perform*
- **News & Social Concern:** *burkina, blacks, judiciary, vietnam, insufficient, sudden, accountable, thailand, rage, committing, widespread, punished, relative, investigators, ppda, deprived, colonization, bastille, interruption, reservists, synagogue, vape, kirk, remuneration, golan*
- **Other Hobbies:** *jesus, christ, shall, salvation, prayer, sins, worship, blessed, resurrection, baptized, holy, righteous, lord, foryou, elder, saints, mary, remission, saves, thee, messiah, heaven, gospel, bible, mercy*
- **Relationships:** *boyfriend, romantic, married, dads, darling, girlfriends, happybirthday, girlfriend, wedding, crush, cousin, couples, marry, daddy, breakup, birthday, cherish, bride, laughs, lily, marriage, kisses, bestie, inseparable, thanksgiving*
- **Science & Technology:** *github, methane, labs, ipad, firefox, rust, genai, artificialintelligence, mozilla, embedded, astronomy, computing, macos, ryzen, kubernetes, reactors, encryption, benchmarks, datasets, defi, browser, cyberattack, ransomware, planets, phishing*
- **Sports:** *derby, coaches, athletics, squad, vinicius, neymar, allezlesbleus, athlete, djokovic, maxime, whistle, uefa, mexicopg, flick, khelif, eurosport, bayern, cristiano, zidane, anceldotti, leon, azerbaijangu, bellingham, dauphin, chants*
- **Travel & Adventure:** *hiking, hotels, booking, portes, trips, boarding, tourist, thailand, traveling, airports, travellers, baggage, travel, beaches, trip, travelling, thisisliving, shuttle, wuhan, bikes, infotrafic, booked, gate, pavement, rides*
- **Youth & Student Life:** *schooling, orientation, aesh, classroom, boarding, bachelor, exam, exams, baccalaureate, college, vocational, genetet, nursery, principal, diploma, elementary, graduates, kindergarten, school, graduation, autism, primary, classes, ulis, colleges*

Finally, to assess linguistic sophistication in users' original tweets beyond topical content, we compute the average word-length; after pre-processing to remove non-linguistic elements (mentions, hashtags, emojis, URLs). We report the 15 words with the highest chi-squared statistics, comparing their frequency in the first and last decile of users average word-length to their overall use in the population in Extended List S3.

Extended List S3. Linguistic sophistication in users' original tweets beyond topical content, shown by computing the average word-length. After pre-processing to remove non-linguistic elements (mentions, hashtags, emojis, URLs), we report through the 15 words with the highest chi-squared statistics, comparing their frequency in the first and last decile of users average word-length to their overall use in the population.

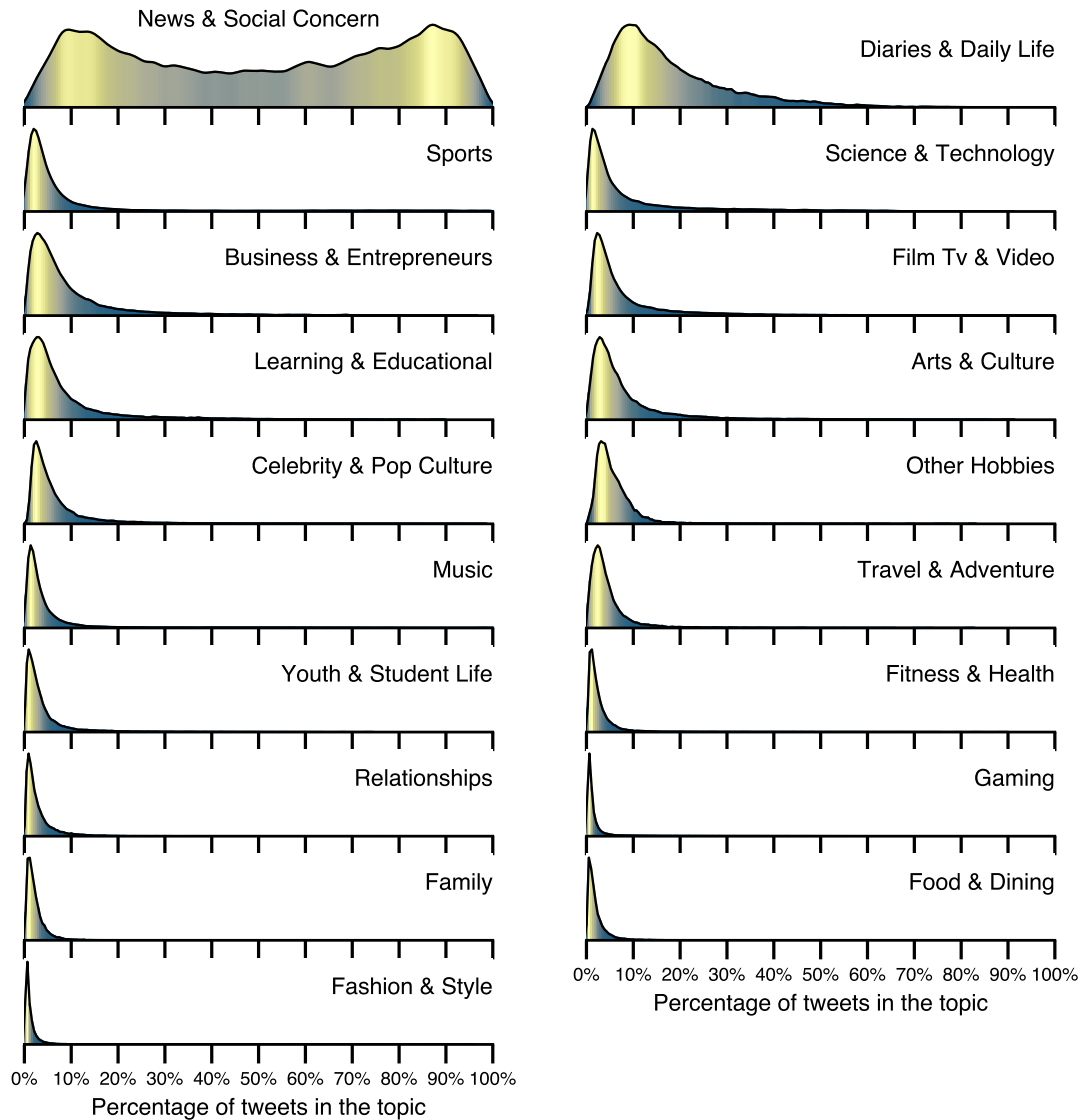


Figure S13: Distribution of users by percentage of tweets that they produce on each topic, according to the topic classifier from (37). Most users produce a low percentage of their tweets on contents that are classifiable into topics from (37). The exception is the topic “News % Social Concern”: users are more evenly distributed when organized by the % of tweets their produce in this topic, and slightly bimodal, concentrated around 10% and 90% of tweets in this topic.

- **Low Average Word-Length:** *jveux, jte, drole, jcrois, etais, wsh, jss, jpense, fr, karmine, karmine corp, oe, jvais, ptite, tq*
- **High Average Word-Length:** *applications, pollinators, exchange, UMR, young talents, cyberattacks, webinars, ArtificialIntelligence, RSE, white paper, innovative, access, innovative, apply, emerging*

Remark: Most terms of the reported “Low Average Word-Length” are jargon or slang: e.g., jveux (for je veux, meaning “I want that”), jte (for je te, meaning “I verb ; you”), drole (drôle, translating to “funny”), jcrois (for je crois, meaning “I think so”). The term wsh has an intricate meaning. It might come from wesh, which is itself a term derived from the word oui in French. It is also a word from Algerian Arab (the largest community of foreign residents), commonly used in the expression We[c/s]h rak?, translatable to How is it going?.

D.4 Ideological Positions of Users

One of the central objectives of our study is to identify whether the embedding space Φ of the recommender system encodes information about political leaning of users. To achieve this goal, we leverage a dataset of politically positioned X users developed by Ramaciotti et al. (40). We first present the dataset in Section D.4.1 to then analyze its intersection with the population $\mathcal{E}$ of our study in Section D.4.2.

D.4.1 Politically-positioned X population dataset from (40)

This dataset contains 494 065 Twitter users interested in French politics, positioned along two political scales or dimensions: 1) a Left-Right dimension and 2) a dimension measuring attitudes towards elites and institutions. The dataset comes with a detailed documentation (41) explaining the methodology used and the quality metrics showcasing the accuracy of inferred political positions. The methodology relies on a two step approach: 1) multidimensional ideology scaling of follower networks on Twitter, and 2) calibration of political positions using political survey data.

Step (1) uses ideological scaling (42, 43) of follower networks. Ideology scaling of choice data has its roots in political science research in the 1980s (44), and was first developed for single-dimensional Liberal-Conservative ideological spatialization of Twitter users in the US by Barberà (45). Ideology scaling spatialization has been subsequently developed to capture additional political dimensions (46). The main motivation for this new dimension is that European settings, and in particular France for our study, cannot be reduced to a single ideological dimension (e.g., Left-Right) in empirical political spatial analyses (47). The population used for this ideological spatialization is chosen among users that follow at least three Members of Parliament (MPs) on X/Twitter (as Barberà did in his 2015 studies (45, 48, 49)). This allows us to select a large number of users that have an interest in national politics. Next, the follower network subtended by these users is collected, and a political homophily process is proposed as an explanation for the observed following links connecting MPs and their users:

$$P(U_i \rightarrow MP_j | \alpha_i, \beta_j, \gamma, \varphi_i, \varphi_j) = \text{logistic} \left(\alpha_i + \beta_j - \gamma \|\varphi_i - \varphi_j\|^2 \right), \quad (9)$$

where, $U_i \rightarrow MP_j$ stands for the event in which user U_i follows MP_j , α_i is the level of *activity* of user U_i measured in number of followed accounts, β_j is the level of *popularity* of MP_j , measured in number of followers, γ is a scale variable, and φ_i and φ_j are the multidimensional positions of U_i and MP_j in some unobservable multidimensional space. This data generation process is homophilic in the sense that $P(U_i \rightarrow MP_j)$ increases as the distance between unobservable positions $\|\varphi_i - \varphi_j\|$ in some latent space decreases. Step (1) involves using observed follower networks between MPs and their followers to infer the parameters of the model in equation Eq. (9). The reader is referred to the work of Barberà on Twitter (48) for the detailed methodology on this inference, to the work of Ramaciotti et al. (46) for an application to French Twitter data, and to the documentation of the dataset we use (41) for the details of the application of the methods to the data of this study.

Step (2) uses political survey data to calibrate multidimensional positions φ_i for the population of study resulting from the Bayesian inference for the parameters of Eq. (9). This step is required because, in Eq. (9), the probability $P(U_i \rightarrow MP_j)$ is invariant to isometries on positions φ_i and φ_j : i.e., for any isometry T , $\|\varphi_i - \varphi_j\| = \|T(\varphi_i - \varphi_j)\|$. In particular, this identification problem means that positions φ on any dimension cannot be taken to stand as indicators for political positions on an identifiable issue or ideology (ideology in the sense of an organized structure of attitudes (50, 51)), because isometries T include flips, translations, and rotation of the latent space. For calibration and identification, the dataset from (41) uses party positions available for political parties, from the Chapel Hill Expert Survey (CHES). The CHES dataset (16) is created with the responses from hundreds of experts in political science, who position political parties along dimensions such as Left-Right,

attitudes towards elites and institutions, Liberal-Conservative, or attitudes towards immigration, to name a few. Creating proxies for party positions as the centroids of MPs φ_j , the dataset from (41) is compounded with an affine transformation going from the parameters space resulting from the Bayesian inference using Eq. (9), onto the space of the dimensions of the CHES data. The leading political dimension of the CHES data is the Left-Right dimension (ranging from 0 to 10, from Left to Right, for positions of parties). Because the scale used for calibration is constructed for positions of political parties, users can have positions to the left of position 0 (the Left-most position for parties) and to the right of position 10 (the Right-most position for parties). Previous research by the authors of the dataset has shown that a second relevant dimension in France and in several countries is the dimension measuring attitudes towards elites and institutions (52), which authors explain that motivates the inclusion of this dimension in the dataset. Because of its importance and availability we also leverage this second dimension in our study. Extended Figure S14 from (41) shows the density of users in the population along these two dimensions, and the positions of MPs and parties used for the calibration.

The authors of the dataset provide quality metrics computed using text data written by users embedded in these two political dimensions. They collected the profile bios of users and proposed an annotation protocol based on LLM classification (using model Zephyr 7B β (53)) to identify a subset of users that could be attributed with the labels “left”, “right”, “elite”, or “populist”. The quality metrics for the positions measure the degree to which the inferred political dimensions can separate these labeled users: i.e., the quality with which the Left-Right dimension can separate users labeled as “left” and “right”, and the quality with which the Anti-elite rhetoric dimension can separate users labeled as “populist” and “elite”. Extended Figure S15 from (41) shows the quality metrics of political positions along these two dimensions. This separation is measured as the accuracy metrics (precision, recall and F1-score) resulting from using the political dimensions for classification of labeled users. It is worth noting that individuals displaying populist discourse (e.g., subscribing to a vision of society divided into the “people” and “elites”) may also be part of elite political or economic groups. The validation proposed by the authors of the dataset (41) relies on the fact that most individuals that display high anti-elite and anti-establishment views will not be themselves part of the elite, and that if the proposed validation metrics hold, it will be *despite* the fact there may be some elite individuals displaying anti-elite and/or anti-establishment rhetoric.

D.4.2 Intersection between the politically-positioned dataset from (41) and our population of study $\mathcal{E}$

Before using the dataset from (41), we compare our population $\mathcal{E}$ with that included in this dataset. An analysis of the dataset reveals that 8 249 users in the population $\mathcal{E}$ of our study (31.1%) have also political positions in the dataset from (41). Furthermore, we compare in Extended Figure S16 the distributions of users from $\mathcal{E}$ that have both, Left-Right and Anti-elite rhetoric positions, and labels in (41). Similarly, we specify the validation of the scales to the embedded population in Extended Figure S17. The reader is referred to the documentation of the dataset we use (41) for the details.

The core demographic and political makeup of our population of study is described in Figure 2 in the main article, with Extended Figure S18 providing a more direct comparison within demographic and political variables.

E Inferring directions in the embedding space that reproduce user features

Let $\Phi \in \mathbb{R}^{n \times d}$ denote the matrix of user embeddings, where each row $\phi_u^\top \in \mathbb{R}^d$ represents the embedding vector of user $u \in \{1, \dots, n\}$ in a d -dimensional embedding space; in our case $d = 256$ and $n = |\mathcal{E}| = 26\,509$. Given an attribute vector $\mathbf{a} \in \mathbb{R}^n$, where a_i corresponds to a measured characteristic of user i , we aim to identify the direction in the embedding space that maximally correlate with this attribute.

For a unit vector $\mathbf{w} \in \mathbb{R}^d$, the correlation between the projection of user embeddings onto $\mathbf{w}$ and the attribute vector $\mathbf{a}$ is given by:

$$\rho_{\mathbf{a}}^{\Phi}(\mathbf{w}) = \frac{(\Phi \mathbf{w})^\top \mathbf{a}}{\|\Phi \mathbf{w}\|_2 \|\mathbf{a}\|_2} \quad (10)$$

where $\|\cdot\|_2$ denotes the Euclidean norm.

We seek to find $\mathbf{w}_{\mathbf{a}}$ such that the projection $\Phi \mathbf{w}_{\mathbf{a}}$ of users' embeddings Φ onto $\mathbf{w}_{\mathbf{a}}$ maximally correlates with $\mathbf{a}$:

$$\mathbf{w}_{\mathbf{a}} = \arg \max_{\mathbf{w} \in \mathbb{R}^d: \|\mathbf{w}\|_2=1} \rho_{\mathbf{a}}^{\Phi}(\mathbf{w}) \quad (11)$$

This optimization problem is a special case of Canonical Correlation Analysis (CCA), where one seeks to find a linear combination that maximizes correlation between a univariate dataset (the attribute vector $\mathbf{a}$) and a multivariate dataset (the user embeddings Φ). The vector $\mathbf{w}_{\mathbf{a}}$ can then be readily obtained by solving the CCA eigenvalue problem.

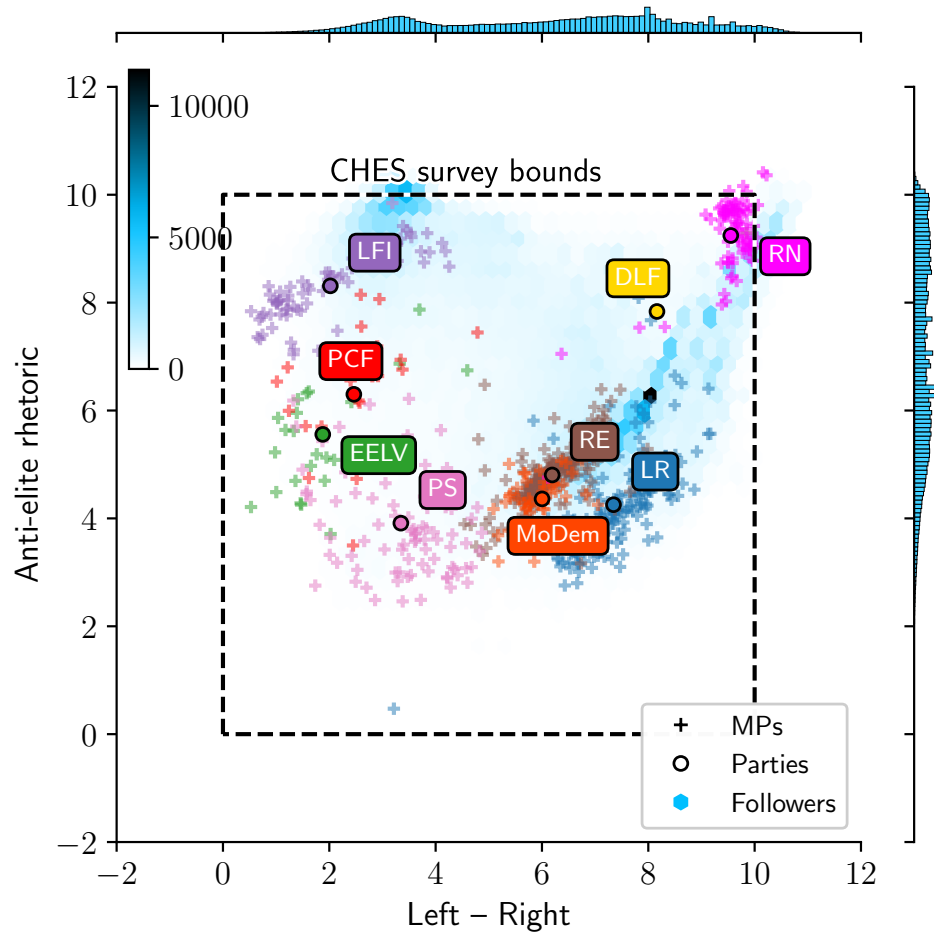


Figure S14: Figure taken from the methodology report of the dataset of politically positioned users from the Twitter French political Twittosphere and that we use in this study (41). Users (followers of MPs, density shown in light blue) are embedded along two political dimensions: Left-Right and an Anti-elite rhetoric dimension. Positions in the dimension are calibrated using data from the Chapel Hill Expert Survey (CHES) (16), and that provide positions of parties as reference points in space. Positions of MPs and political parties used in the calibration are shown in colors. Along the Left-Right scale 0 is the left-most position for political parties, while 10 is the right-most, and 5 is the political center. Along the Anti-elite rhetoric scale 0 stands for no salience of anti-establishment and anti-elite rhetoric, while 10 stands for extreme salience of anti-establishment and anti-elite rhetoric.

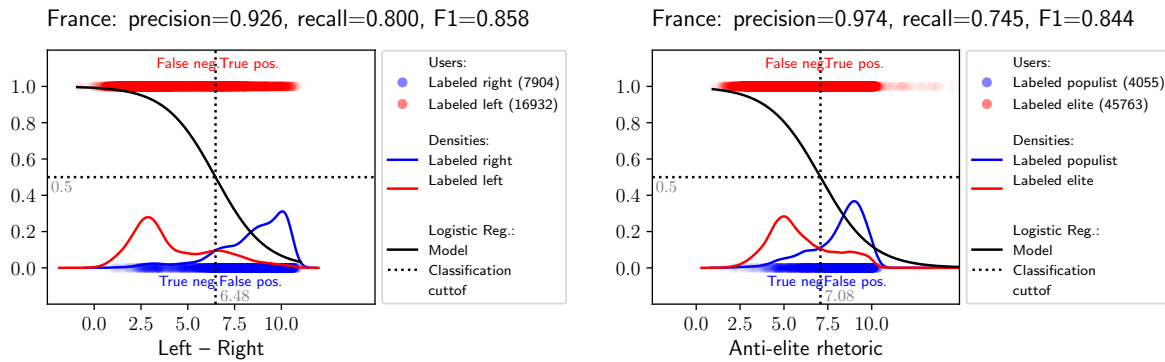


Figure S15: Figures taken from the methodological report of the dataset of politically positioned users from the Twitter French political Twittosphere (41), reporting the validation of spatial positioning of users along the political dimensions. Density of users labeled as identified with “left”, “right”, “elite”, or “populist” along dimensions Left-Right (left figure) and Anti-elite rhetoric (right figure). Labeling was computed with annotations produced by the LLM Zephyr 7B β (53). The figure shows the goodness of fit of logistic regressions as classifiers of pairs of labels using the political dimension on which they are embedded. Individuals displaying populist discourse (e.g., subscribing to a vision of society divided into the “people” and “elites”) may also be part of elite political or economic groups. The validation proposed by the authors of the dataset (41) relies on the fact that most individuals that display high anti-elite and anti-establishment views will not be themselves part of the elite, and that if the proposed validation metrics hold, it will be *despite* the fact there may be some elite individuals displaying anti-elite and/or anti-establishment rhetoric.

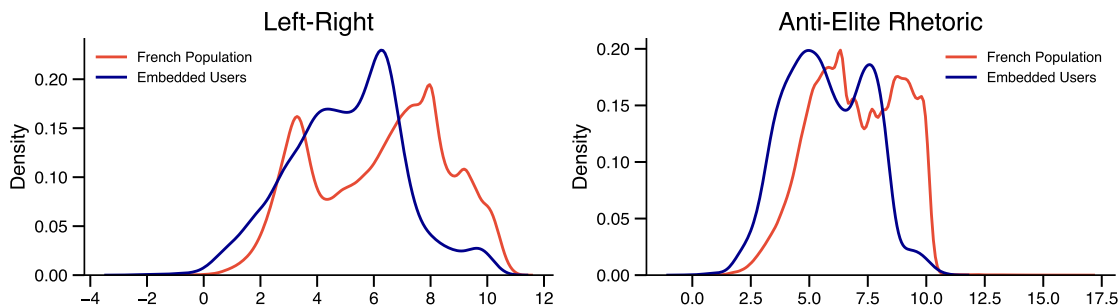


Figure S16: Comparison of positions along the two political dimensions for the population available from the data from (41), and the intersection of 8 249 users in the population $\mathcal{E}$ of our study (31.1%). Distribution of users Left-Right (left figure) leaning and Anti-Elite rhetoric (right figure) within the embedded population $\mathcal{E}$ (in red) compared to the distribution within the overall French Twitter population (in blue) from the dataset of politically-positioned Twitter users from (41).

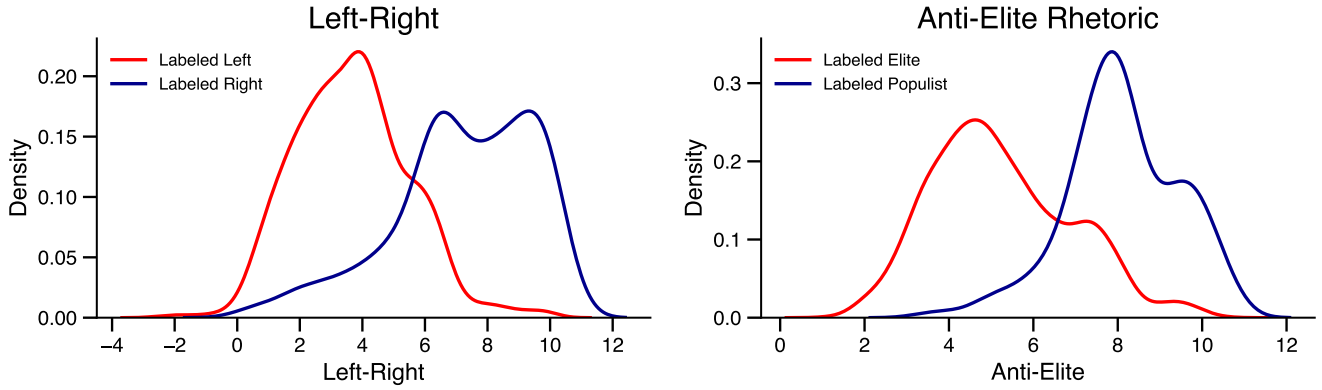


Figure S17: Validation of the positions of 8 249 users in the population $\mathcal{E}$ of our study (31.1%) that are also in dataset from (41), by checking that labeled users are also correctly ordered by the two political dimensions. Distribution of users Left-Right leaning (left figure) and Anti-Elite rhetoric (right figure) within the embedded population $\mathcal{E}$ segmenting those labeled either as “Left” or “Right”, and either as “Elite” or “Populist” by a LLM Zephyr 7B β (53).

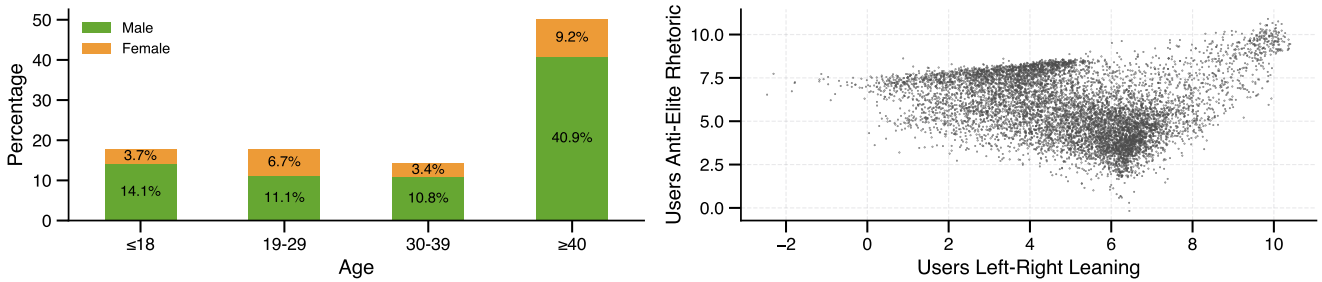


Figure S18: Summary of the main attributes of users in our population. Distribution of inferred age and binary gender for the embedded population $\mathcal{E}$ (left figure), Scatter plot of embedded users’ positions on the left-right and anti-elite rhetoric scales (right figure).

To establish statistical significance, we employ a permutation test with $N = 10^4$ repetitions. For each iteration, we generate $\tilde{\mathbf{a}}$ by randomly permuting the elements of $\mathbf{a}$, compute $\mathbf{w}_{\tilde{\mathbf{a}}}$ through CCA, and measure the resulting correlation $\rho_{\tilde{\mathbf{a}}}^{\Phi}(\mathbf{w}_{\tilde{\mathbf{a}}})$. This procedure yields the null distribution of correlations expected under the hypothesis of no meaningful relationship. The one-sided p-value is subsequently determined as the proportion of permuted absolute correlations that exceed the absolute observed correlation $\rho_{\mathbf{a}}^{\Phi}(\mathbf{w}_{\mathbf{a}})$.

Similarly, since the null distribution of feature directions $\mathbf{w}_{\mathbf{a}}$ depends on the distribution of user embeddings in $\mathbb{R}^d$, we use permutation tests, rather than assuming a uniform hypersphere distribution, to assess the significance of similarities between $\mathbf{w}_{\mathbf{a}}$ and $\mathbf{w}_{\mathbf{b}}$, the embedding directions associated with features $\mathbf{a}$ and $\mathbf{b}$. The one-sided p-value is computed as the proportion of randomized cosine similarities that exceed the observed similarity between $\mathbf{w}_{\mathbf{a}}$ and $\mathbf{w}_{\mathbf{b}}$, where the randomized similarities are computed between embedding directions $\mathbf{w}_{\tilde{\mathbf{a}}}$ and $\mathbf{w}_{\tilde{\mathbf{b}}}$ associated with shuffled features $\tilde{\mathbf{a}}$ and $\tilde{\mathbf{b}}$.

F Validating inferred embedding directions

After identifying through CCA the direction $\mathbf{w}_{\mathbf{a}}$ in the embedding space $\mathbb{R}^d$ spanned by learned account embeddings Φ that maximally correlates with a given feature of interest $\mathbf{a}$, we validate its interpretation as genuinely capturing this feature. This is needed because of two reasons. 1) Even if a direction produces a high correlation in ordering and positioning users according to attributes for users, the variance or disturbance error of an underlying linear model might still be high. 2) Secondly, not every user in $\mathcal{E}$ has known values for all attributes. Accordingly, we need to control that the CCA directions induce an ordering of users also for those that have no known attributes. The CCA direction for Left-Right leanings was computed relying on the 8 249 users from $\mathcal{E}$ (31.1%) that had a position on the Left-Right scale, and we need to control that the direction orders the rest of users in $\mathcal{E}$ correctly.

To validate the positions in the embedding independently from the data we used inferring a direction $\mathbf{w}_{\mathbf{a}}$, we analyze the friends and tweet content of users at both extremes of the direction $\mathbf{w}_{\mathbf{a}}$. Specifically, we identify overrepresented friends, words, and hashtags in the social connections and tweet content of users in the first and last deciles of $\Phi\mathbf{w}_{\mathbf{a}}$ (hereafter referred to as lower and upper users over $\mathbf{w}_{\mathbf{a}}$).

We determine overrepresented entities by comparing their observed frequency in the target pool to their expected frequency under uniform random distribution along the $\mathbf{w}_{\mathbf{a}}$ direction. To protect user privacy and ensure representativity, we solely report popular accounts with over 50k followers, followed by at least 10% of the users in the considered decile. More precisely, for each inferred dimension $\mathbf{w}_{\mathbf{a}}$ corresponding to a feature $\mathbf{a}$, we report the popular accounts, having more than 50k followers, that are over-followed by the low projecting (the 10% of users having the lowest positions on $\mathbf{w}_{\mathbf{a}}$) and high projecting (the 10% of users having the highest positions on $\mathbf{w}_{\mathbf{a}}$) users in $\mathcal{E}$. In the case of words and hashtags, the calculation is done on the basis of those included in the tweets produced by users $\mathcal{E}$.

F.1 Inferred Left-Right political leaning direction $\mathbf{w}_{LR}$

We analyze the relationship between users' political leanings and their projection onto the maximally correlated embedding direction $\mathbf{w}_{LR}$. As displayed on Extended Figure S19, we observe a strong Pearson correlation $\rho = 0.887$, $p < 10^{-4}$.

To validate the interpretation of this direction, we examine how users' following patterns of Members of French Parliament (MPs) vary along $\mathbf{w}_{LR}$. For each user, we compute the fraction of followed MPs from three political categories in the French National Assembly: Left-leaning (*La France Insoumise*, LFI), Center (*La République en Marche*, LREM), and Right-leaning (*Rassemblement National*, RN). The bottom panel of Extended Figure S19 displays these fractions as a function of users' projection onto $\mathbf{w}_{LR}$. LREM (now rebaptized as *Renaissance*), RN, and LFI are the parties having produced the first three front-runners in the first round of the last presidential election in France in 2022, obtaining correspondingly, 27.9% (Emmanuel Macron), 23.2% (Marine Le Pen) and 22% (Jean-Luc Mélenchon) of the vote (the next and fourth candidate obtaining 7.1%)

Users projecting low onto $\mathbf{w}_{LR}$ (below -1) predominantly follow Left-leaning MPs, with minimal following of Center MPs and virtually no following of Right-leaning MPs. Conversely, users projecting high onto $\mathbf{w}_{LR}$ exhibit an opposite pattern, with Right-leaning MPs constituting 59% of their followed parliamentarians. Users projecting onto $\mathbf{w}_{LR}$ at around 1.1 predominantly follow MPs from the Center, who comprise 56% of their followed parliamentarians.

Following our previously outlined methodology, we report in Extended Lists S4 & S5 the accounts, words and hashtags overrepresented in users' following networks and content production at opposite ends of the $\mathbf{w}_{LR}$ spectrum further validate this interpretation.

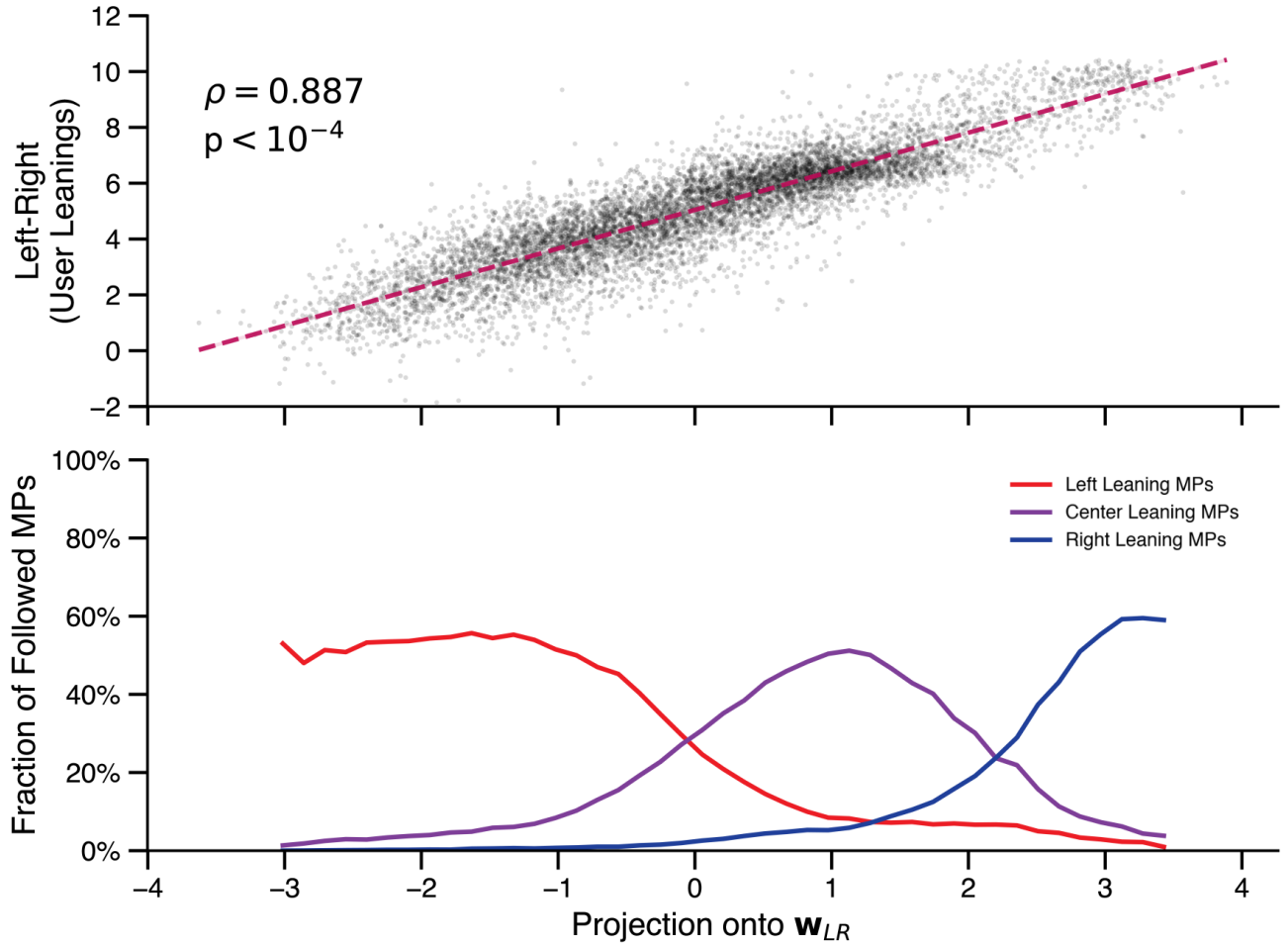


Figure S19: Positioning of users by the recommender along direction $\mathbf{w}_{LR}$ and validation. (top figure) Scatter plot showing users' Left-Right political leaning against their projection onto $\mathbf{w}_{LR}$, the embedding direction maximizing feature correlation; $\rho = .887$, $p < 10^{-4}$. (bottom figure) Average fraction of followed Members of Parliament from main Left-leaning (LFI), Center (LREM), and Right-leaning (RN) parties, as a function of users' projection onto $\mathbf{w}_{LR}$.

Extended List S4. Accounts followed, words used, and hashtag used by 10% of users with the highest probability of begin Left-leaning according to the embedding space of recommender system (i.e., having low values along $\mathbf{w}_{LR}$).

- **Popular followed accounts:** @AvecJLM, @FiAssemblee, @PODEMOS, @europeangreens, @sebastiendelogu, @NUPES_2022_, @Linsoumission, @CarolineFiat54, @LachaudB, @GabrielBoric, @alma_dufour, @TrouveAurelie, @Ugobernalicis, @GreensEFA, @PCF
- **Popular words:** austerity, ruralities, sleeping in street, perjury, rebellious, parliamentary niche, restoration of wealth tax, pension repeal, capital income, gaza west bank
- **Popular hashtags:** #jde2024 (Summer Days of Ecologists), #noconfidencemotion, #impeachmentmarch, #nfp (New Popular Front), #victorynfp, #wevotenfp, #amfis2024 (Far-left-wing Summer University), #october1strike, #newpopularfront, #ceasefire, #popularunion, #macronimpeachment, #sanofi (pharmaceutical company), #eelv (Green Party)

Extended List S5. Accounts followed, words used, and hashtag used by 10% with the highest probability of begin Right-leaning according to the embedding space of recommender system (i.e., having high values along $\mathbf{w}_{LR}$).

- **Popular followed accounts:** @gabriellecluzel, @tegnererik, @IdLibertes, @tanguy-france, @stanislasrig, @JulienOdoul, @LesPatriotes, @matteosalvinimi, @AmauryBucco, @ChdOrnellas, @GregToussaint4, @Stephane_Ravier, @kevinbossuet, @Bruno_Attal_, @HistoiredeFran7
- **Popular words:** islamization, illegal immigrants, sovereign powers, remigration, narco, family reunification, cathedrals, allah akbar
- **Popular hashtags:** #deportation, #hdpros (Right-Leaning TV Show), #drugs, #europe1 (French Radio), #reconquête (Far-Right Policy Party), #angryfarmers, #sinwar

F.2 Inferred anti-elite rhetoric direction $\mathbf{w}_{AntiElite}$

Following the same approach as for Left-Right positions, we plot in Extended Figure S20 users' positions on the anti-elite rhetoric dimensions against their projection onto $\mathbf{w}_{AntiElite}$. This projection exhibits a strong Pearson correlation, $\rho = 0.863$, $p < 10^{-4}$, with users' anti-elite salience scores.

Users projecting low onto $\mathbf{w}_{AntiElite}$, do not display salient sentiments against elites and institutions, predominantly follow center-leaning MPs, who constitute 73% of their followed parliamentarians. This proportion decreases along $\mathbf{w}_{AntiElite}$, while the fraction of far-right-leaning MPs in users' followed MPs progressively increases.

Following the previously outlined methodology, we report in Extended Lists S6 & S7 the accounts, words and hashtags overrepresented in users' following networks and content production at opposite ends of the $\mathbf{w}_{AntiElite}$ spectrum further validate this interpretation.

Extended List S6. Accounts followed, words used, and hashtag used by 10% with the lowest probability of displaying anti-elite rhetoric according to the embedding space of recommender system (i.e., having low values along $\mathbf{w}_{AntiElite}$).

- **Popular followed accounts:** @FaireEnsemblefr, @FrontRep, @DeputesEnsemble, @thalesgroup, @JeunesMacron, @RPFranceUE, @JBLemoyne, @EmmWargon, @tcabarrus, @regionbretagne, @FranceTravail, @olivierdussopt, @StanGuerini, @J_Denormandie, @ENGIEgroup
- **Popular words:** competitive, Air Force, DGA (Directorate General of Armament'), national days, brothers in arms, industrial sovereignty, civil security, circular economy
- **Popular hashtags:** #decarbonization, #SMEs (Small/Medium Enterprises), #choosefrance (investment in France initiative), #france2030 (innovation investment plan), #territories, #cyber (cybersecurity), #municipalities, #cybermonth, #frenchtech (French tech ecosystem), #HR (Human Resources), #companies, #defense

Extended List S7. Accounts followed, words used, and hashtag used by 10% with the highest probability of displaying anti-elite rhetoric according to the embedding space of recommender system (i.e., having high values along $\mathbf{w}_{AntiElite}$).

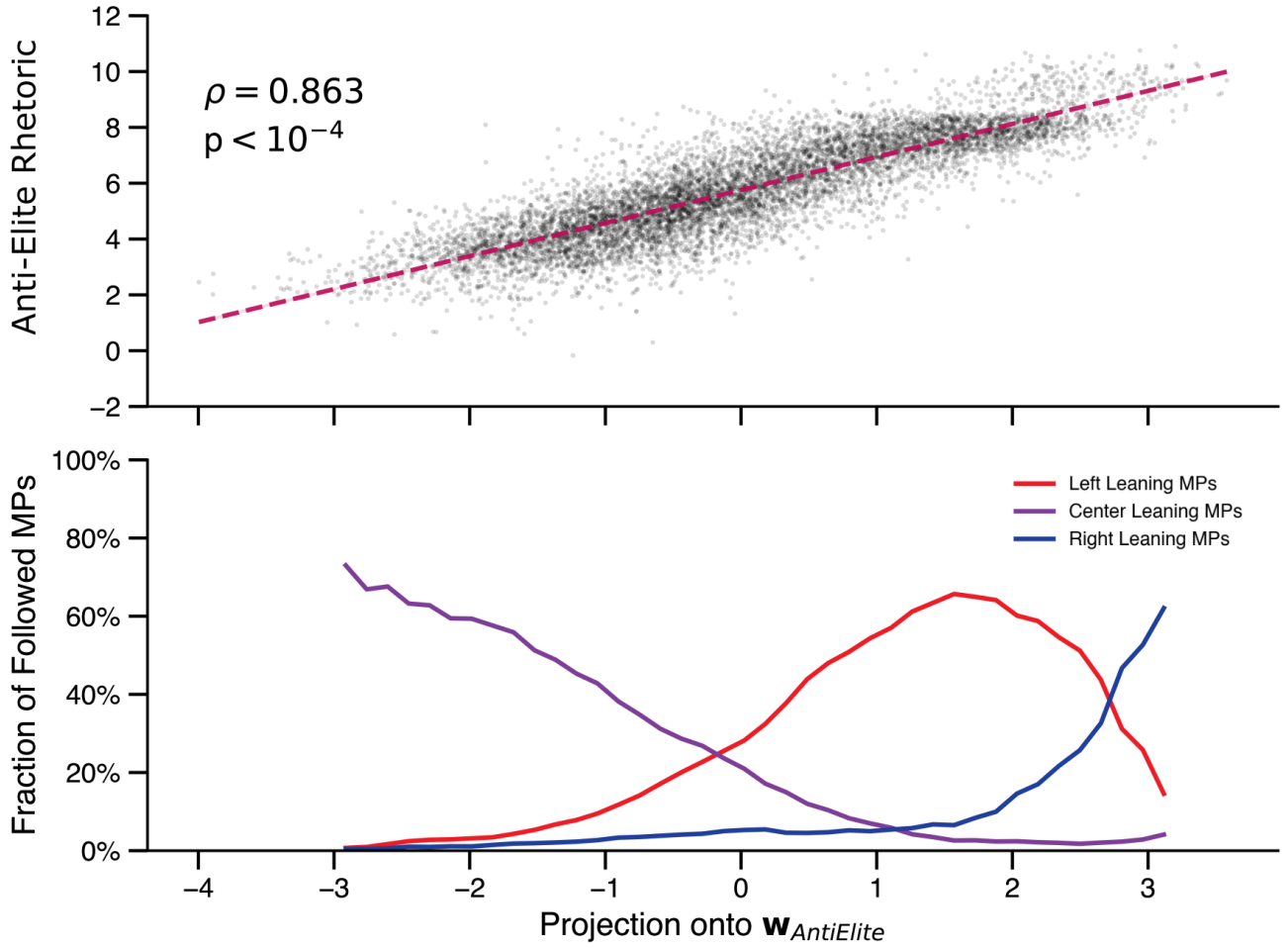


Figure S20: Positioning of users by the recommender along direction $\mathbf{w}_{AntiElite}$ and validation. (Top) Scatter plot showing users' Anti-Elite Saliency against their projection onto $\mathbf{w}_{AntiElite}$, the embedding direction maximizing feature correlation; $\rho = .863$, $p < 10^{-4}$. (Bottom) Average fraction of followed Members of Parliament from Left-leaning (LFI), Center (LREM), and Right-leaning (RN) parties, as a function of users' projection onto $\mathbf{w}_{LR}$.

- **Popular followed accounts:** @louisfouch3, @BanounHelene, @StatChrisCotton, @xazalbert, @MichelJeanDomi1, @ReinfoCovidOff, @CStrategies, @VictorSinclair3, @verity.france, @Laissonslespre1, @myriampalomba, @LaurentOzon, @gregtabibian, @CaudeHenrion, @Beatrice_Rosen
- **Popular words:** remigration, politico-media, pedophilia, Islamization, UEJF (Union of French Jewish students), Soros, scheming, family reunification
- **Popular hashtags:** #deportation, #stopgenocide, #reconquête (far-right French political party), #impeachment, #zelen-sky, #angryfarmers, #zemmour (French far-right politician), #mercosur, #macronimpeachment, #impeachmentmarch, #amfis2024 (Far-left-wing summer university)

F.3 Inferred gender direction $\mathbf{w}_{\text{gender}}$

The direction $\mathbf{w}_{\text{gender}}$ in the embedding space induced by user embeddings Φ that maximally correlates with the estimated probability of users being “female” (binary classification, with complementary probability of being “male”), achieves a Pearson correlation of $\rho = .384$, $p < 10^{-4}$. When framed as a classification task for these two groups, projections onto $\mathbf{w}_{\text{gender}}$ predict user gender with an AUC-ROC of $.731 \pm .044$ (standard deviation, 10-fold cross validation). We display in Extended Figure S21 the distribution of projections onto $\mathbf{w}_{\text{gender}}$ for users inferred as “female” (M3 prediction above 0.5) and “male”.

A Mann–Whitney U test reveals that the positions of users along $\mathbf{w}_{\text{gender}}$, classified as “female” are greater than the positions of users accounts classified as “male”, with $p \ll 10^{-4}$.

Following the previously outlined methodology, we report the accounts, words and hashtags overrepresented in users’ following networks and content production at opposite ends of the $\mathbf{w}_{\text{gender}}$ spectrum further validate this interpretation.

Extended Figure S21 shows the fraction of users following at least one of the 15 popular accounts, reported below, as a function of their projection onto $\mathbf{w}_{\text{gender}}$. The fraction of users following inferred female-popular accounts increases along $\mathbf{w}_{\text{gender}}$, while those following inferred male-popular accounts decreases to a lesser extent along this direction. As before, we report the accounts followed, and words and hashtags used by users at the extremes of direction $\mathbf{w}_{\text{gender}}$ in Extended Lists S8 & S9.

Extended List S8. Accounts followed, words used, and hashtag used by 10% with the highest probability of being male according to the embedding space of recommender system (i.e., having low values along $\mathbf{w}_{\text{gender}}$).

- **Popular followed accounts:** @RMCsport, @beinsports_FR, @Romain_Molina, @PSG_inside, @lequipe, @AntoGriezmann, @equipedefrance, @KMbappe, @X, @NetflixFR, @AastierOff, @NASA, @Michel_Goya, @realDonaldTrump, @vincentglad
- **Popular words:** genesio, mourinho, zerbi, borussia, lacazette, bundesliga, kolo muani, rabiote, pogacar, clasico, away section, juventus, scorers, muani, goalkeeper,
- **Popular hashtags:** #uel (Europa League), #ligue1 (French Football League), #tdf2023 (Tour de France 2023), #om (Olympique Marseille), #ucl (Champions League), #teamom (Team Olympique Marseille), #cycling, #ompsg (Olympique Marseille vs Paris Saint-Germain), #top14 (French Rugby League), #defense, #frenchxv (French Rugby Team), #bitcoin, #pgw (Paris Games Week), #tdf2024 (Tour de France 2024)

Users projecting low onto $\mathbf{w}_{\text{gender}}$, i.e., having high probability of being male, demonstrate strong engagement with sports content across their following patterns and published content, frequently mentioning football players and associated hashtags.

Extended List S9. Accounts followed, words used, and hashtag used by 10% with the highest probability of being female according to the embedding space of recommender system (i.e., having high values along $\mathbf{w}_{\text{gender}}$).

- **Popular followed accounts:** @ONUFemmes, @Ovidieofficiel, @osezlefeminisme, @NousToutesOrg, @BeaulieuBap, @PepiteSexiste, @PenelopeB, @alicecoffin, @MaiaMazaurette, @monachollet, @Klaire, @LyesLouffok, @faizazrl, @misspress, @ATrapenard
- **Popular words:** Stéphane Plaza (TV host under investigation), masculinists, PPDA (Patrick Poivre d’Arvor - former news anchor accused of SGBV), Godrèche (Judith - actress speaking against SGBV), Haenel (Adèle - actress speaking against SGBV), plaintiffs (multiple accusers in various cases), ableism, intersex, male domination, pimping, Judith Godrèche (leading voice against SGBV), incest, femicides, Jacquot (Benoît - film director accused of SGBV)

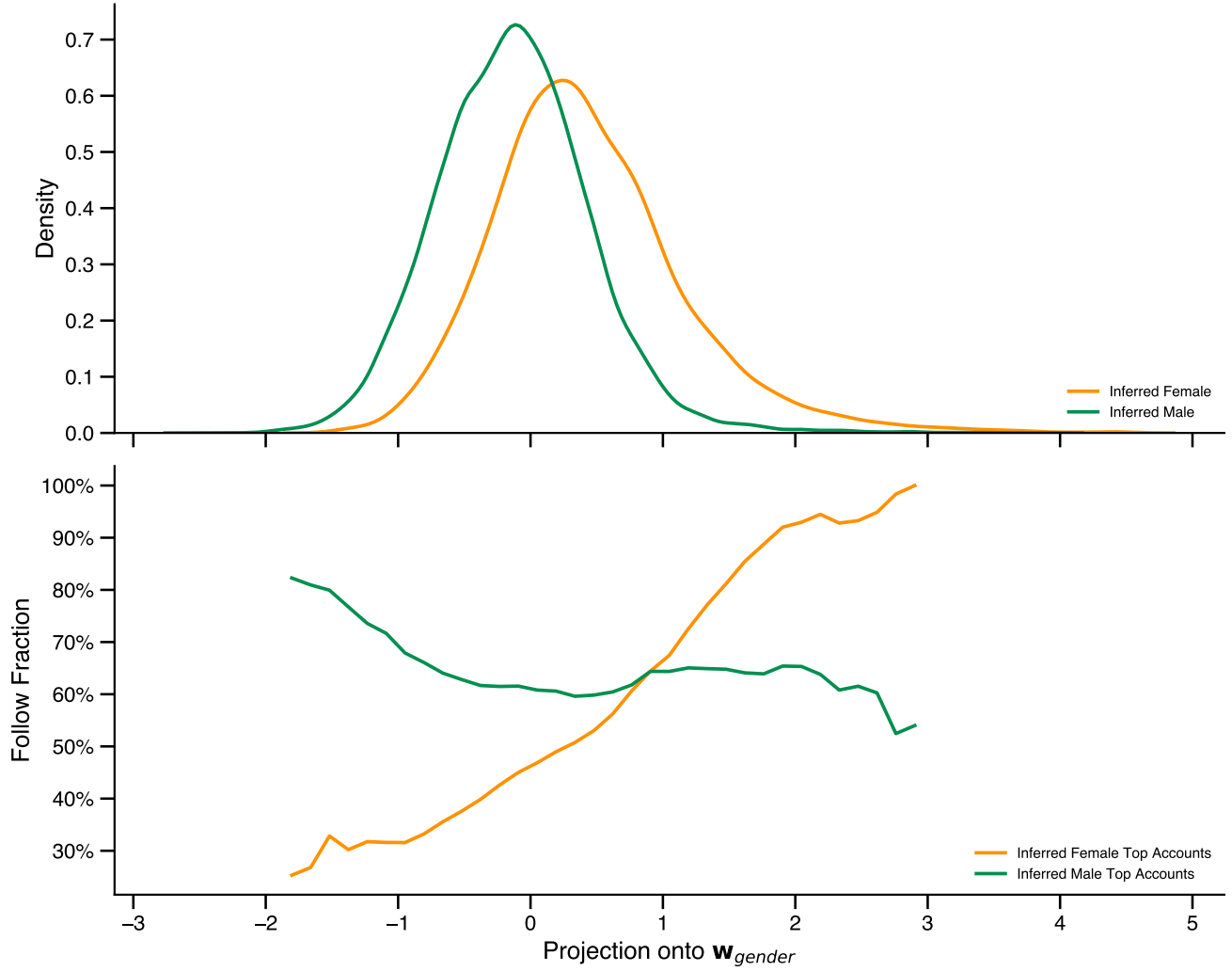


Figure S21: Positioning of users by the recommender along direction $\mathbf{w}_{gender}$ and validation. (Top figure) Distribution of inferred-male/female users according to their projection onto $\mathbf{w}_{gender}$. (Bottom figure) Fraction of users following at least one of the popular accounts, as function of their projection onto $\mathbf{w}_{gender}$.

- **Popular hashtags:** #rape, #ppda (Accused of SGBV), #rapes, #metoo, #pelicot (Survivor of SGBV), #sexism, #mazan, #departieu (Accused of SGBV), #feminism, #insociety, #feminism, #sexualviolence, #giselepelicot (Survivor of SGBV), #bedos (Accused of SGBV), #womensrights

Users projecting high onto $\mathbf{w}_{gender}$, i.e., having high probability of being female, demonstrate significant engagement in French public discourse surrounding sexual violence, industry abuse, and feminist movements. Their content frequently references individuals involved in Sexual and Gender Based Violence cases, including both alleged perpetrators and survivors, alongside associated advocacy hashtags.

F.4 Inferred age direction $\mathbf{w}_{age}$

We denote by $\mathbf{w}_{age}$ the direction in the embedding space that achieves the strongest correlation with users' estimated age. Because M3 provides categorical values (≤ 18 , 19-29, 30-39, and ≥ 40), we computed, for each user, estimated age as a continuous variable by considering the middle points of each category (15.5 years for the ≤ 18 bin, 24.5 years old for the 18-29 bin, 34.5 years old for the 30-39, and 45 years for the ≥ 40 bin), weighted by the probability of belonging to the age category as given by the M3 model. We then compared these continuous age values with positions along $\mathbf{w}_{age}$, achieving a Pearson correlation of $\rho = .562$, $p < 10^{-4}$.

When framed as a multi-categorical classification task, projections onto $\mathbf{w}_{age}$ predict user M3 age bins with an AUC-ROC of $.692 \pm .030$ (standard deviation, 10-fold cross validation). Extended Figure S22 displays the distribution of projections onto $\mathbf{w}_{age}$ for each age category inferred by M3.

A Mann–Whitney U test reveals that the distribution of projection onto $\mathbf{w}_{age}$ of accounts classified as ≤ 18 is stochastically lesser than the distribution underlying accounts classified as 19 – 29, similarly the distribution associated with 19 – 29 accounts is stochastically lesser than the one associated with 30 – 39, finally the distribution of projection of account classified as ≥ 40 is stochastically greater than the one of 30 – 39; with $p \ll 10^{-4}$.

For validation purposes, we identify the direction in the embedding that achieves the strongest correlation with volunteers' self-declared age. The direction identified for this narrow set of users aligns with the direction previously identified considering the M3-inferred population with a cosine similarity of 0.63 and as observed with the M3-inferred direction poorly aligns with the left-right political direction with a cosine similarity of 0.16.

Following the previously outlined methodology, we report the accounts, words and hashtags overrepresented in users' following networks and content production at opposite ends of the $\mathbf{w}_{age}$ spectrum further validate this interpretation.

Extended Figure S22 shows the fraction of users following at least one of the 15 popular accounts reported below, as a function of their projection onto $\mathbf{w}_{age}$. The fraction of users following inferred young-popular accounts significantly decreases along $\mathbf{w}_{age}$, from being followed by 86% of users projecting low onto $\mathbf{w}_{age}$ to being followed by only 5% of those scoring high along this same direction. Conversely, the fraction of users following inferred old-popular accounts increases, yet to a lesser extent, along the direction. As before, we report the accounts followed, and words and hashtags used by users at the extremes of direction $\mathbf{w}_{age}$ in Extended Lists S10 & S11.

Extended List S10. Accounts followed, words used, and hashtag used by 10% with the highest probability of being young according to the embedding space of recommender system (i.e., having low values along $\mathbf{w}_{age}$).

- **Popular followed accounts:** @Theorus_, @idkbvtimleila, @KoteiZousa, @EcrisioFX, @KennyStream, @JoelPostbad, @Grimkujow_, @LRB_LoL, @DephaseuR, @Chap_GG, @mastuu_, @Doigby, @Inoxtag, @KarmineCorp, @Kammeto
- **Popular words:** karmine corp, jcrois, faker, mdrrrrrrr, jpense, ptdrrrr, sdm, jveux, ninho, oomf, tq, wsh, zevent, ptdrrrrr
- **Popular hashtags:** #azurlane (Mobile game), #bluearchive (Mobile game), #worlds2024 (League of Legends Championship), #lec (European League Championship), #kcorp (French esports team), #zevent2024 (French charity stream event), #nintendoswitch (gaming console), #pgw (Paris Games Week), #onepiece (manga/anime), #advertising (ad disclosure), #rt (retweet), #ad (advertisement), #illustration (art content), #staracademy (French TV show)

Users projecting low onto $\mathbf{w}_{age}$, i.e., inferred young, demonstrate strong engagement with gaming and entertainment content in their following patterns and published content, frequently referencing video games, e-sport events, and their associated hashtags.

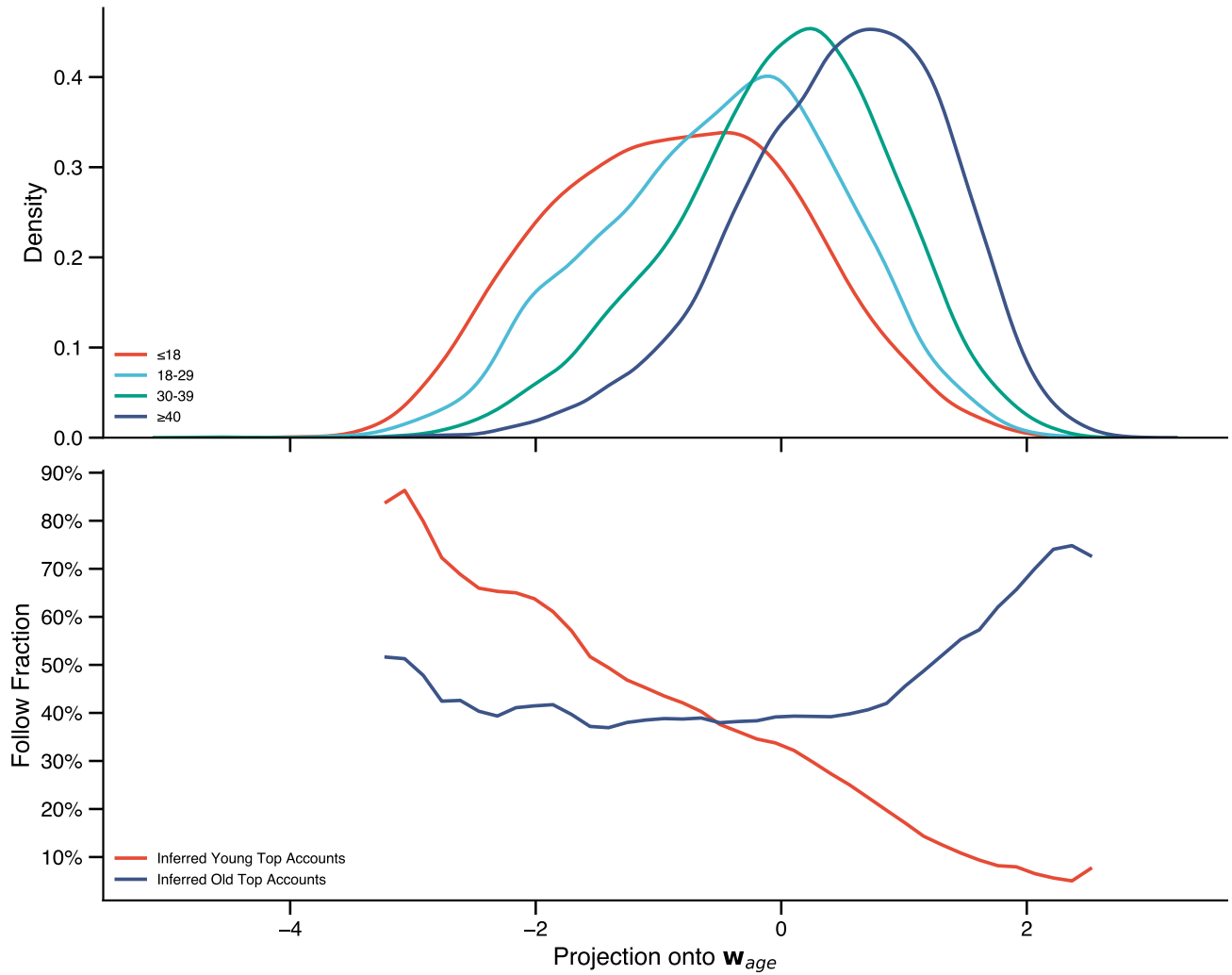


Figure S22: Positioning of users by the recommender along direction $\mathbf{w}_{age}$ and validation. (Top figure) Distribution of users projection onto $\mathbf{w}_{age}$. (Bottom figure) Fraction of users following at least one of the age-popular accounts, as function of their projection onto $\mathbf{w}_{age}$.

Extended List S11. *Accounts followed, words used, and hashtag used by 10% with the highest probability of being old according to the embedding space of recommender system (i.e., having high values along $\mathbf{w}_{age}$).*

- **Popular followed accounts:** @DarmonMichael, @frebsamen, @medef, @JLBorloo, @LaurentFabius, @chevenement, @pierremscovici, @ericwoerth, @MarisolTouraine, @jwaintraub, @JBLemoyne, @danycohnwendit, @VenturaAlba, @jccambadelis, @jeanmarcayrault
- **Popular words:** competitiveness, Jacques Delors (former European Commission President), industrial sovereignty, decentralization, social cohesion, Mario Draghi (former ECB head), ministerial, Michel Barnier (EU Brexit negotiator, Former French Prime Minister), de-escalation, Slava Ukraini
- **Popular hashtags:** #zan (France’s “zero net artificialization” environmental law), #decarbonization, #unga79 (UN General Assembly), #cop16, #bonjourchezvous (morning TV show), #SMEs (Small/Medium Enterprises), #sustainabledevelopment, #localauthorities, #carbon, #france2030 (national investment plan), #cop29, #proximity, #choosefrance (investment attraction program)

Users projecting high onto $\mathbf{w}_{age}$, i.e., inferred old, demonstrate significant engagement with French national and international political discourse.

F.5 Inferred directions in the embedding space of the recommender system and that capture interest in different topics as defined by (37)

F.5.1 News $\mathbf{w}_{News}$

We plot in Extended Figure S23 the projection of users onto $\mathbf{w}_{News}$ versus the fraction of tweets published or relayed by users related to “News & Social Concern”, showing a strong Pearson correlation; $\rho = .851$, $p < 10^{-4}$.

We also show, as function of users projection onto $\mathbf{w}_{News}$, the fraction of users following at least one account from the 15 selected accounts that are popular among news-enthusiast users, according to our previously described methodology. We report the accounts followed by users at the extreme positions in direction $\mathbf{w}_{News}$ in Extended List S12.

Extended List S12. *15 selected accounts that are popular among news-enthusiast users according to their positing along the direction of the embedding that induces the highest correlation with interest in news.*

@JosepBorrellF, @KenRoth, @AgnesCallamard, @haaretzcom, @EP_President, @SecBlinken, @amnesty, @UNHumanRights, @maxseddon, @coe, @L_Insoumission, @hrw, @Europarl_EN, @Refugees, @EUCouncil

F.5.2 Sports $\mathbf{w}_{Sports}$

We plot in Extended Figure S24 the projection of users onto $\mathbf{w}_{Sports}$ versus the fraction of tweets published or relayed by users related to “Sports”, showing a strong Pearson correlation; $\rho = .809$, $p < 10^{-4}$.

Additionally, we show, as function of users projection onto $\mathbf{w}_{Sports}$, the fraction of users following at least one account from the 15 selected accounts that are popular among sports-enthusiast users, according to our previously described methodology. We report the accounts followed by users at the extreme of direction $\mathbf{w}_{Sports}$ in Extended List S13.

Extended List S13. *15 selected accounts that are popular among sports-enthusiast users according to their positing along the direction of the embedding that induces the highest correlation with interest in sports.*

@UNFP, @EuropaLeague, @AS_Monaco, @FIFACOM, @LFPfr, @premierleague, @ChampionsLeague, @PVSportFR, @losclive, @ballondor, @Eurosport_FR, @Ligue1, @CanalplusFoot, @Omar_daFonseca, @lachainelequipe

F.5.3 Science & Technology $\mathbf{w}_{SciencesTech}$

We plot in Extended Figure S24 the projection of users onto $\mathbf{w}_{SciencesTech}$ versus the fraction of tweets published or relayed by users related to “Science & Technology”, showing a strong Pearson correlation; $\rho = .758$, $p < 10^{-4}$.

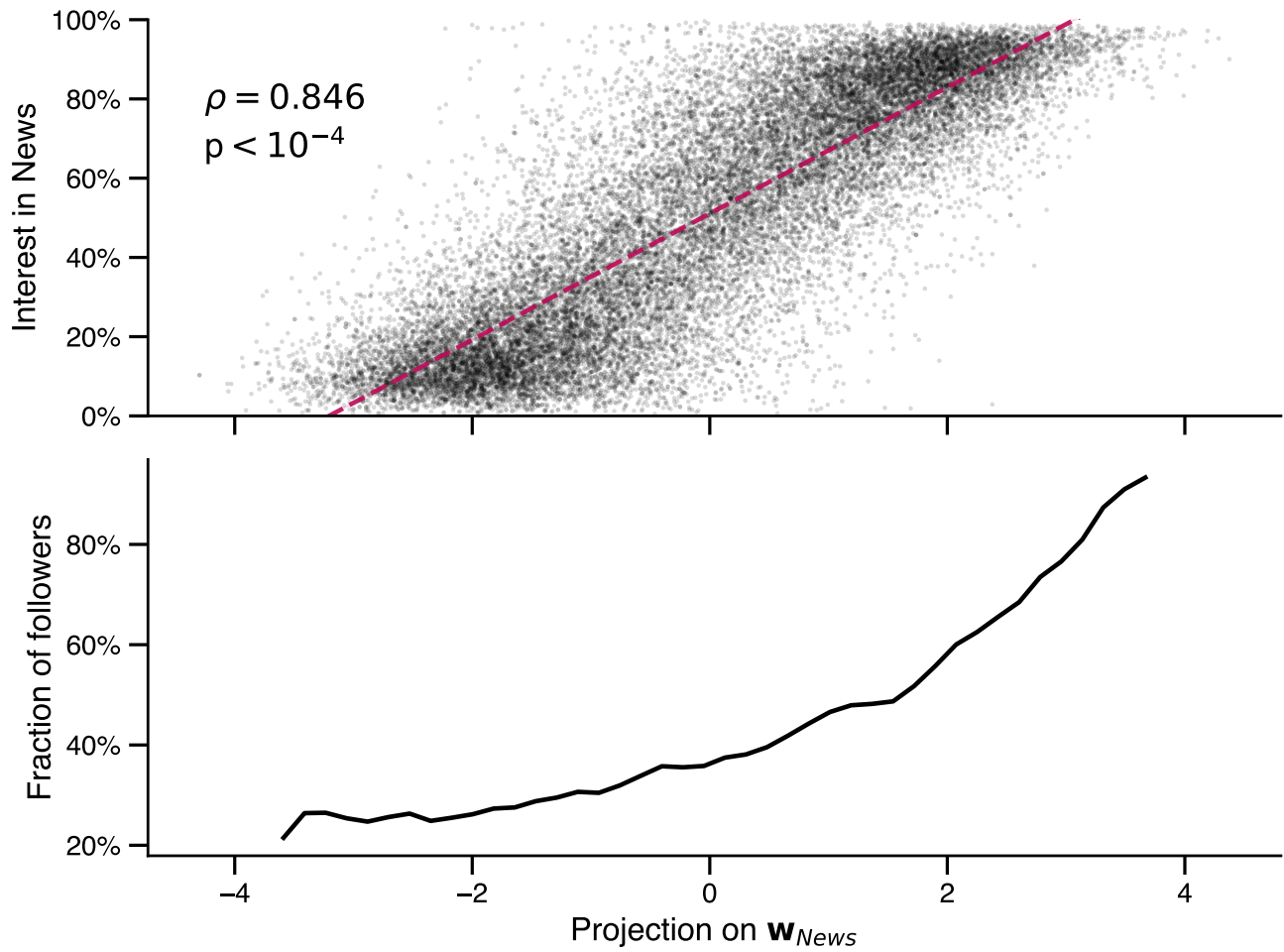


Figure S23: Positioning of users by the recommender along direction $\mathbf{w}_{News}$ and validation. (Top figure) Scatter plot of user projections onto embedding direction $\mathbf{w}_{News}$ versus their fraction of news-related tweets. (Bottom figure) Fraction of users following at least one account popular among news enthusiasts.

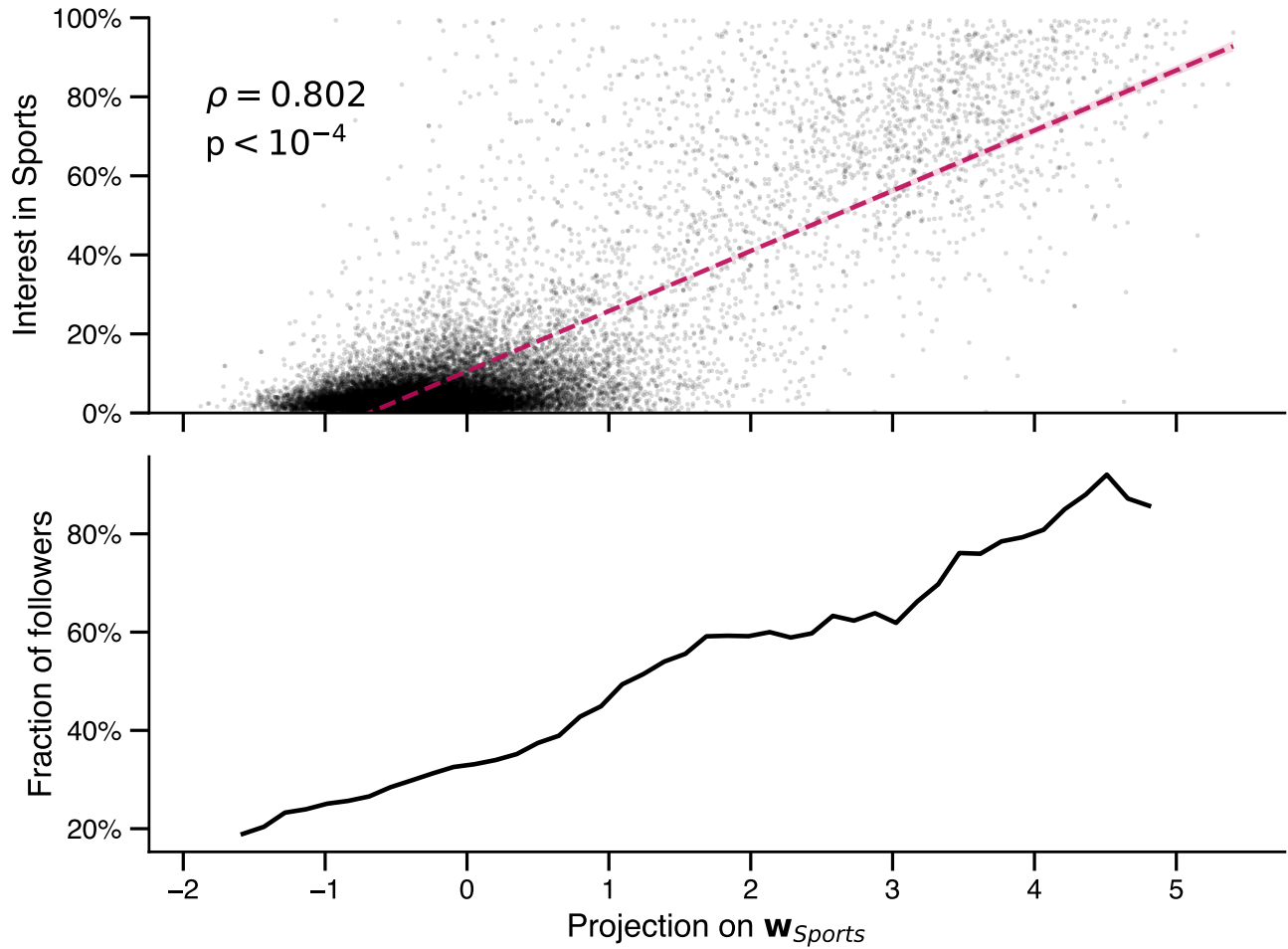


Figure S24: Positioning of users by the recommender along direction $\mathbf{w}_{Sports}$ and validation. (Top figure) Scatter plot of user projections onto embedding direction $\mathbf{w}_{Sports}$ versus their fraction of sports-related tweets. (Bottom figure) Fraction of users following at least one account popular among sports enthusiasts.

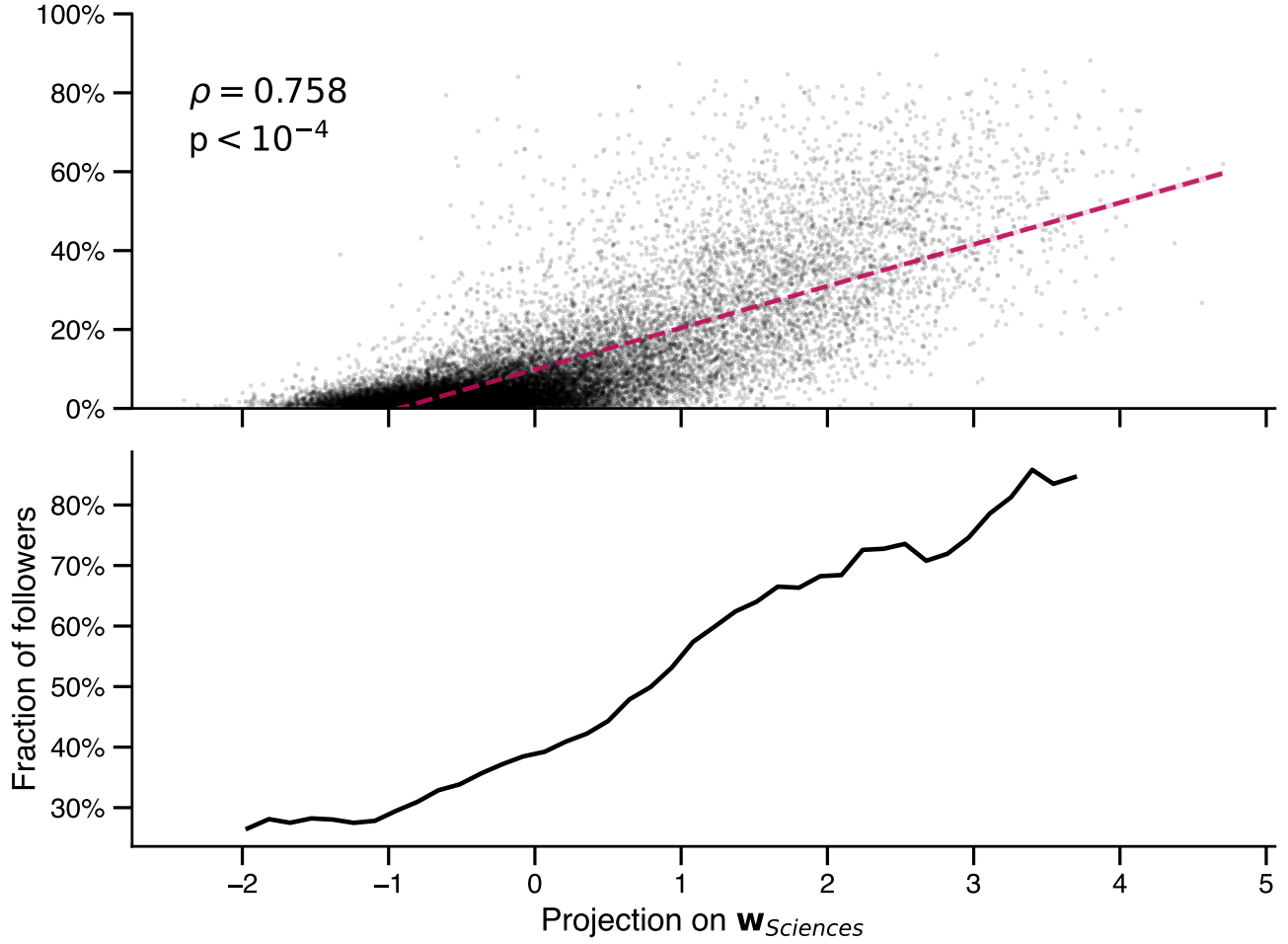


Figure S25: Positioning of users by the recommender along direction $\mathbf{w}_{\text{SciencesTech}}$ and validation. (Top figure) Scatter plot of user projections onto embedding direction $\mathbf{w}_{\text{SciencesTech}}$ versus their fraction of sciences-and-technology-related tweets. (Bottom figure) Fraction of users following at least one account popular among sciences and technology enthusiasts.

Additionally, we show, as function of users projection onto $\mathbf{w}_{\text{SciencesTech}}$, the fraction of users following at least one account from the 15 selected accounts that are popular among science and technology-enthusiast users, according to our previously described methodology. We report the accounts followed by users at the extreme of direction $\mathbf{w}_{\text{SciencesTech}}$ in Extended List S14.

Extended List S14. 15 selected accounts that are popular among sciences and technology-enthusiast users according to their positing along the direction of the embedding that induces the highest correlation with interest in sciences and technology.

@karpathy, @ylecun, @techreview, @OpenAI, @sama, @NaturePortfolio, @Nature, @newscientist, @ScienceMagazine, @TechCrunch, @esa, @UNFCCC, @tim_cook, @IPCC_CH, @Google

F.6 Inferred directions in the embedding space of the recommender system and that the lexical complexity $\mathbf{w}_{\text{complexity}}$

We plot in Extended Figure S26 users' Lexical Complexity —measured as the average word length— against their projection onto $\mathbf{w}_{\text{complexity}}$. This projection exhibits a Pearson correlation, $\rho = 0.704$, $p < 10^{-4}$, with users' Lexical Complexity.

Following the previously outlined methodology, we report in Extended Lists S15 & S16 the accounts, words and hashtags overrepresented in users' following networks and content production at opposite ends of the $\mathbf{w}_{\text{complexity}}$ spectrum further

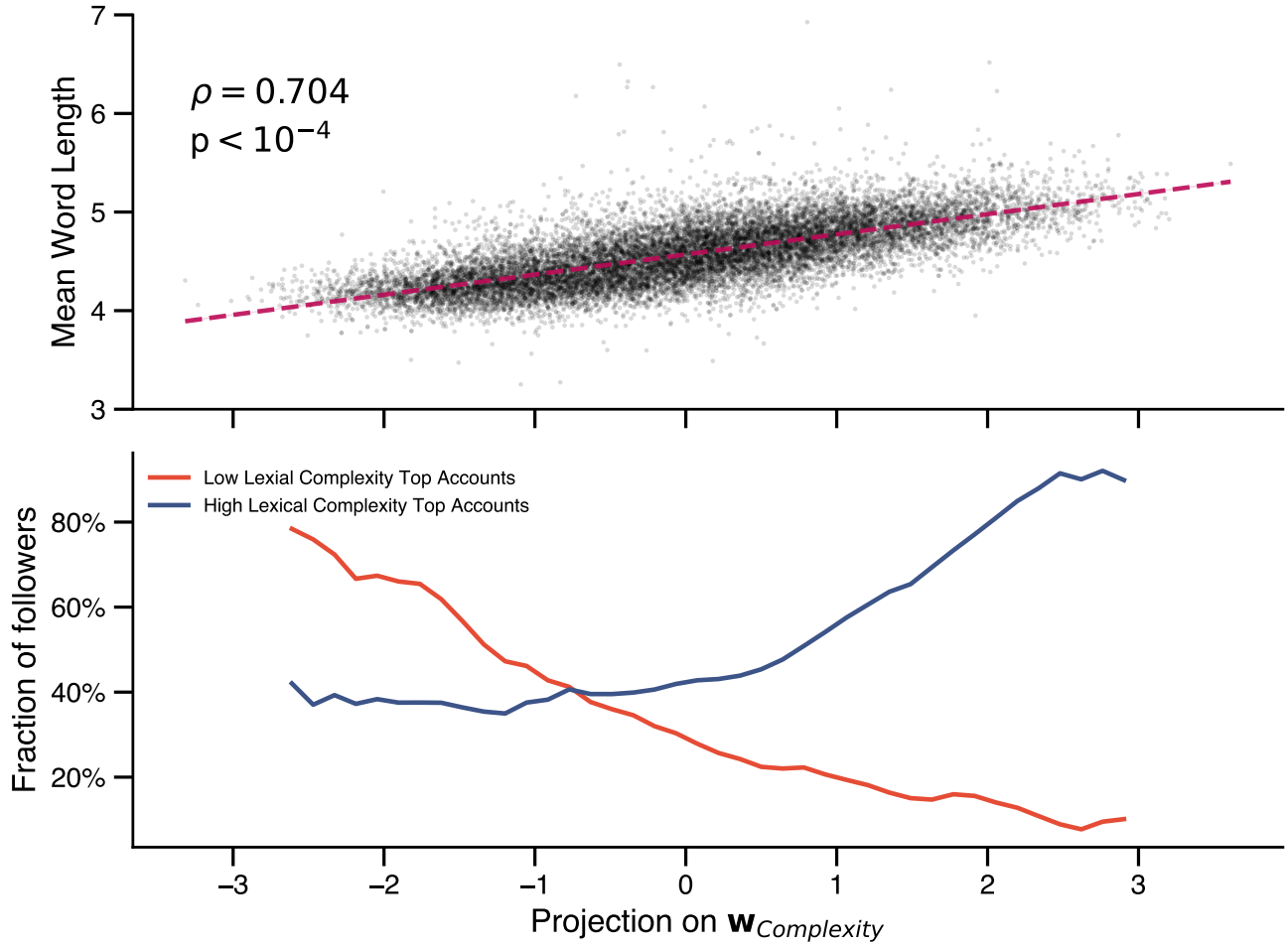


Figure S26: Positioning of users by the recommender along direction $\mathbf{w}_{complexity}$ and validation. (Top figure) Scatter plot showing users' Lexical Complexity against their projection onto $\mathbf{w}_{complexity}$, the embedding direction maximizing feature correlation; $\rho = .704$, $p < 10^{-4}$. (Bottom figure) Fraction of users following at least one of the low-complexity/high-complexity -popular accounts, as function of their projection onto $\mathbf{w}_{complexity}$.

Table S26: Pearson correlations induced by CCA directions in the embedding space of the recommender for different attributes of users. CCA directions are those that maximize the correlation between the attributes and the projection of users in the direction. All p-value are $< 10^{-4}$.

Attributes	Correlation	Attributes	Correlation
Left-Right Leaning	0.89	Interest in Diaries & Daily Life	0.63
Anti-Elite Rhetoric	0.86	Verified Status	0.62
Interest in News & Social Concern	0.85	Interest in Music	0.60
Interest in Sports	0.80	Age	0.56
Interest in Learning & Education	0.76	Interest in Travel & Adventure	0.54
Interest in Science & Technology	0.75	Interest in Other Hobbies	0.53
Followers Count (Percentile Rank)	0.73	Interest in Relationships	0.50
Followers Count (Log)	0.73	Interest in Family	0.41
Average Word Length	0.70	Daily Likes	0.40
Youth & Student Life	0.70	Male-Female	0.38
Arts & Culture	0.67	Followers Count (Raw value)	0.37
Interest in Film, TV & Video	0.67	Friend/Follower Ratio	0.35
Interest in Gaming	0.64	Fashion & Style	0.34
Interest in Fitness & Health	0.64	Interest in Food & Dining	0.32
Interest in Business & Entrepreneurs	0.63	Daily Posts	0.26
Interest in Celebrity & Pop Culture	0.63		

validate this interpretation.

Extended List S15. Accounts followed, words used, and hashtag used by 10% with the highest probability of displaying low linguistic complexity according to the embedding space of recommender system (i.e., having low values along $\mathbf{w}_{Complexity}$).

- **Popular accounts:** @Theorus_, @_ronare, @Grimkujuow_, @EcrisioFX, @Pandrezz, @maximebiaggi, @DephaseuR, @le_egar, @Feldup_official, @Mousstery, @KoteiZousa, @DylanKujo2, @kronomuzik, @Chap_GG, @KarmineCorp
- **Popular words:** jcrois, frf, oomf, nn, stv, jss, jpense, jme, jveux, oe, wsh, jte, karmine corp, bz, jvais,
- **Popular hashtags:** #lec, #kcorp, #worlds2024, #j2s, #zevent2024, #businesscollaboration, #staracademy, #pgw, #new-profilephoto, #rt, #kohlanta, #twitch, #competition, #rapes, #medalalert

Extended List S16. Accounts followed, words used, and hashtag used by 10% with the highest probability of displaying high linguistic complexity according to the embedding space of recommender system (i.e., having high values along $\mathbf{w}_{Complexity}$).

- **Popular accounts:** @INRAE_France, @Edu_Num, @LUsineDigitale, @sup_recherche, @Polytechnique, @usinenouvelle, @Inserm, @palaisdecouvert, @SorbonneParisI, @UNESCO_fr, @citedessciences, @Travail_Gouv, @iledefrance, @AFD_France, @Bpifrance
- **Popular words:** european week, activity report, apply, pollinators, aap, labeled, entrepreneurship, guided tours, artificial intelligence, european heritage, adopt, rse, call for applications, signing agreement, apply
- **Popular hashtags:** #france2030, #sustainabledevelopment, #orientation, #fds2024, #sme, #jobs, #ecologicaltransition, #carbon, #event, #waste, #innovations, #decarbonation, #ecologicaltransition, #sustainable, #territory

F.7 Summary of correlations of user attributes in the embedding spaces of the recommender system

We display in Extended Figure S29 the Spearman correlation between the projection of population $\mathcal{E}$ onto direction $\mathbf{w}_a$ and its projection onto $\mathbf{w}_b$. We observe that, contrary to the feature space where users' age negatively correlates with the probability of being "female" (i.e., younger women and older men), the projections onto the recommender system's learned directions $\mathbf{w}_{age}$ and $\mathbf{w}_{gender}$ show no significant correlation. However, the correlations between age and anti-elite rhetoric, as well as interest in news, are amplified in the projected space compared to the feature space. We display in Extended Figure S30 the projections of users embedding onto latent directions $\mathbf{w}_{LR}$, $\mathbf{w}_{gender}$ and $\mathbf{w}_{age}$, showing low correlations.

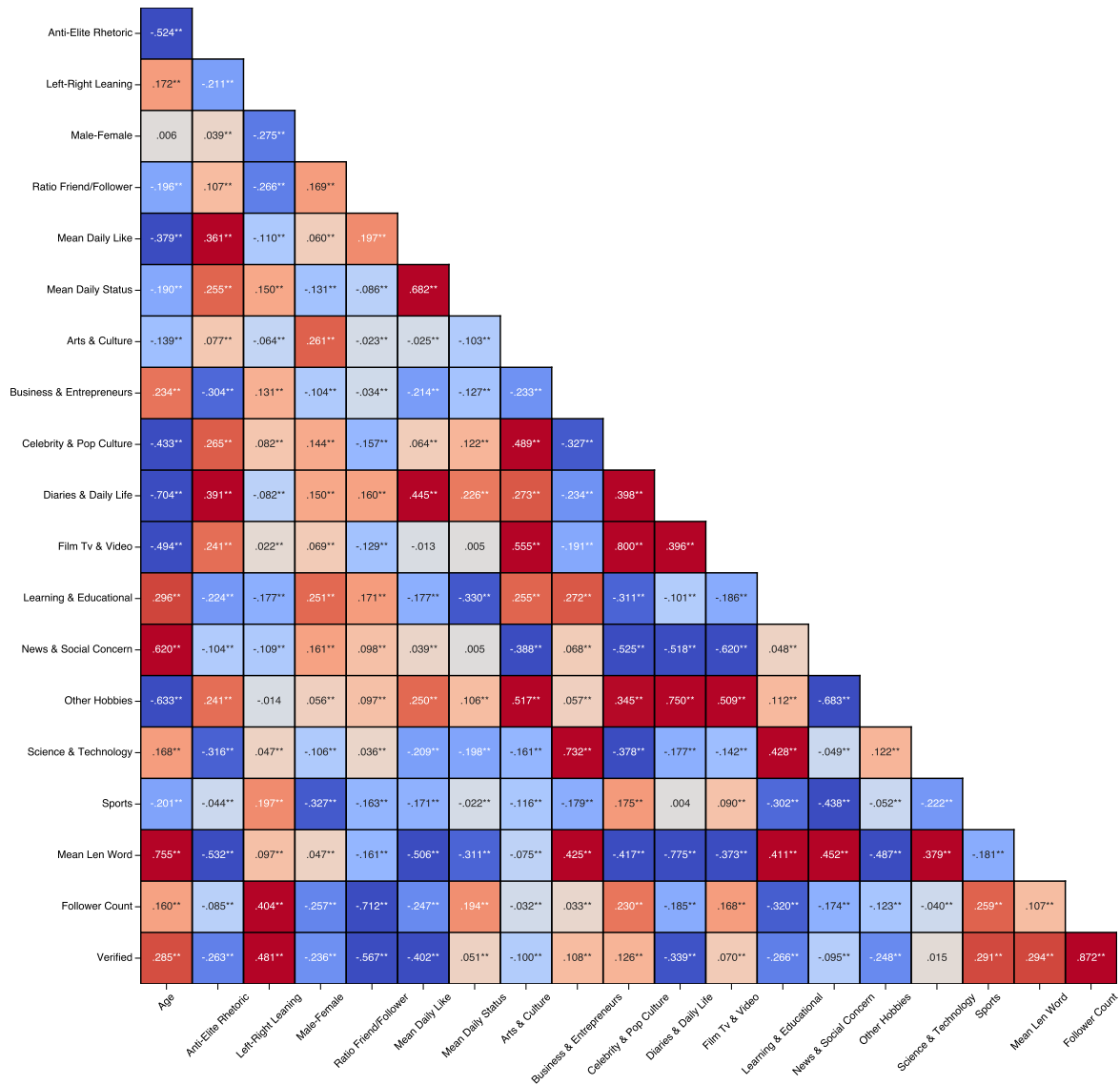


Figure S27: Spearman correlation between projection of users in the study onto the CCA directions of the embedding corresponding to different attributes of users. (* indicates p-value < 0.01, ** < 0.001)

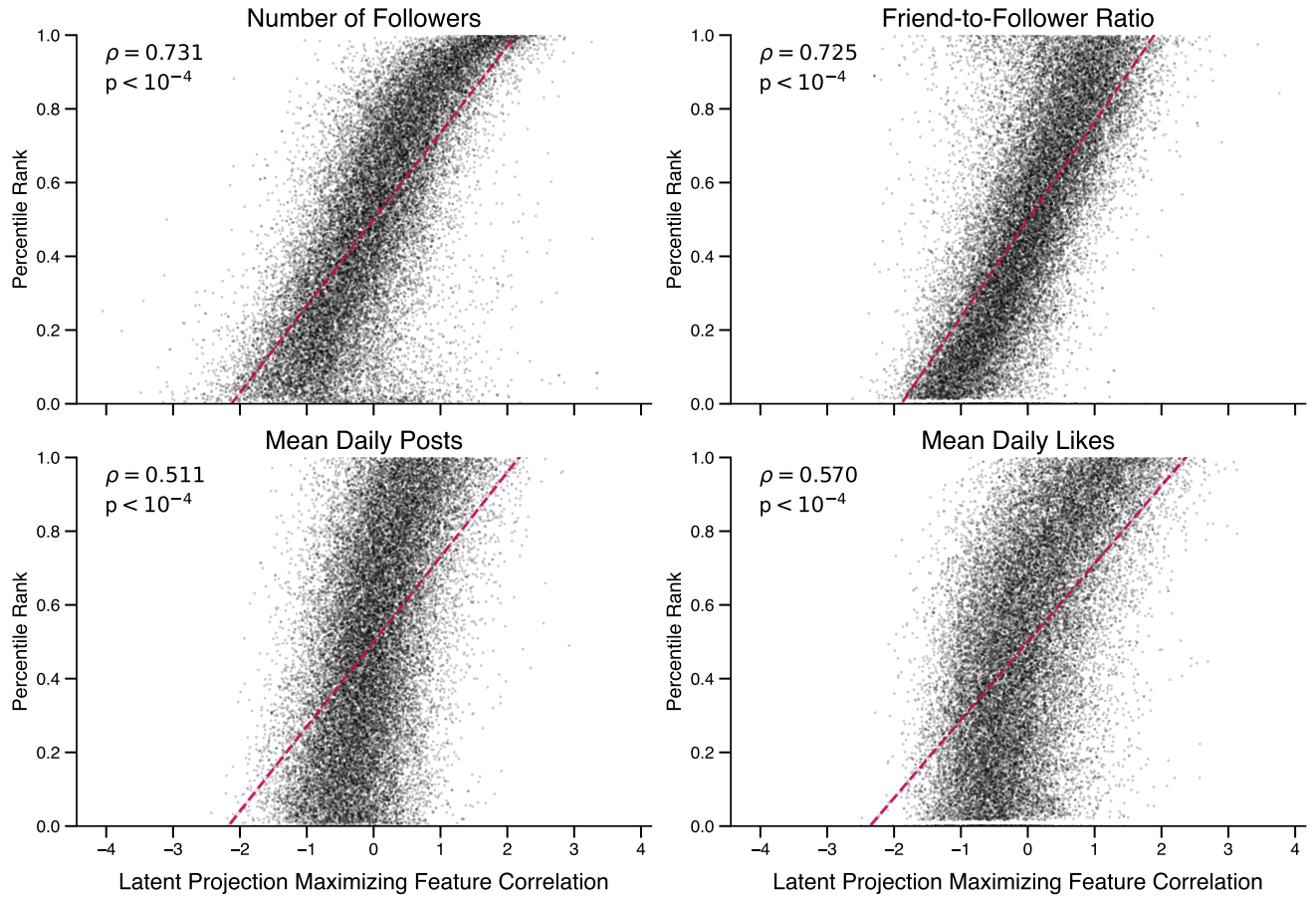


Figure S28: Scatter plot of user features versus projections onto embedding directions that maximally correlate with user characteristics: number of followers, friend-to-follower ratio, mean daily posts, and mean daily likes (percentile rank).

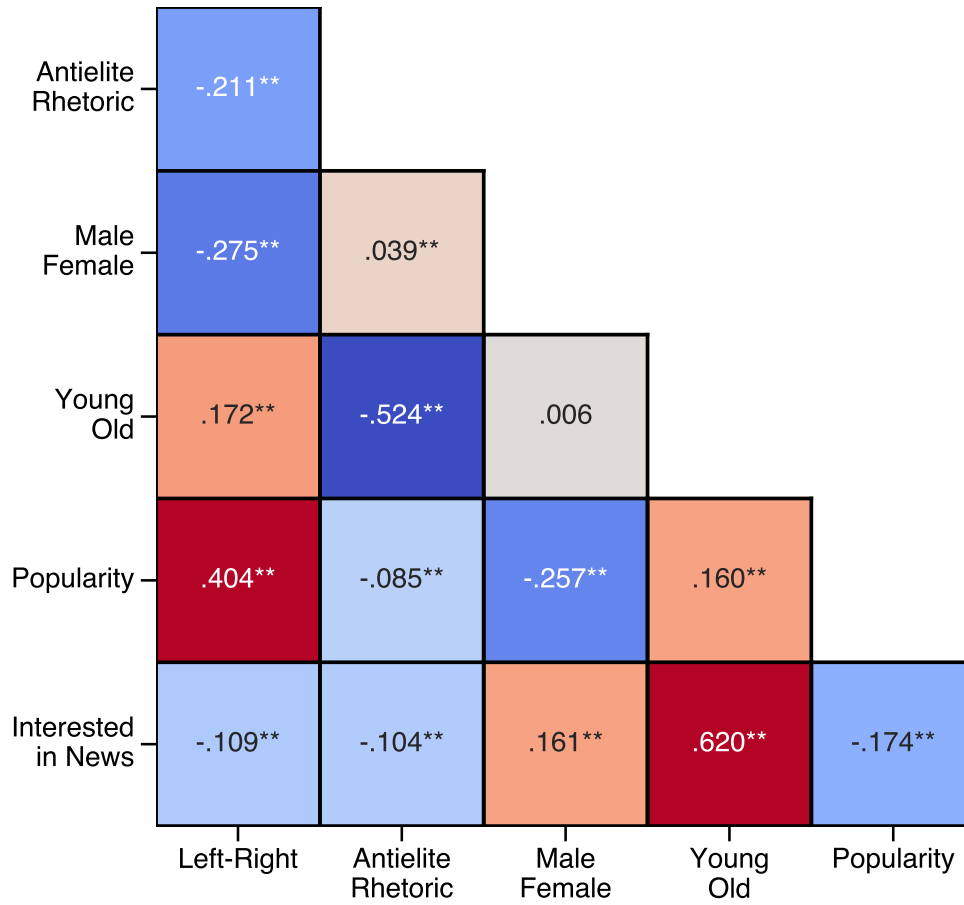


Figure S29: Spearman correlation between projection of users in the study onto the CCA directions of the embedding corresponding to different attributes of users. (* indicates p-value < 0.01, ** < 0.001)

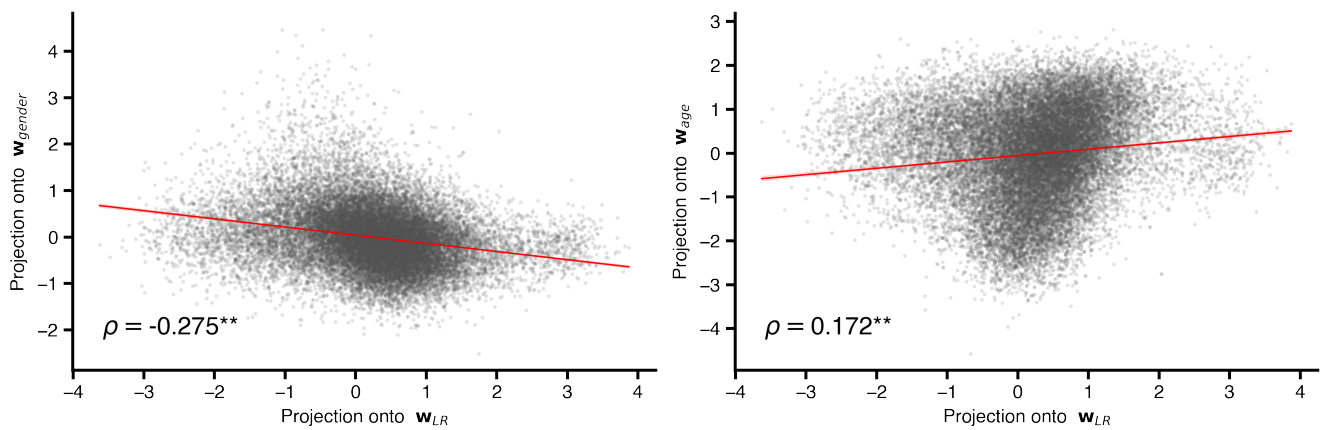


Figure S30: Scatter plots between projection of users in the study onto the CCA directions of the embedding corresponding to Left-Right leaning w_{LR} , gender, w_{gender} and age w_{age} . Low correlation values show the degree of independence of these attributes as learned by the recommender system.

G Constraining political position information encoded in the recommender system

In this section we propose a method for limiting the information encoded in identifiable directions $\mathbf{w}$ associated with user features, amounting to constrained learning in that the information encoded in this directions is rendered unavailable for the downstream recommendation tasks under the hypothesis of linear encoding. In Section G.1 we first specified the proposed procedure for constraining learned attributes. In Section G.2 we then specify a protocol to assess the loss in accuracy and changes in diversity of recommendations resulting from this constraints. Finally, in Section G.3 we show the results for different strategies and at different levels of aggregation on users based on their features.

G.1 Iterative projections onto subspace orthogonal to feature directions $\mathbf{w}_a$ in the embedding space Φ

Having identified directions in the user embedding space inducing the highest correlations with user attributes, we now seek to constrain the use of such attributes when formulating recommendations. This task methodologically relates to “debiasing” language models, similar to Bolukbasi et al.’s identification of the “gender subspace” in word embeddings (54), which allows zeroing out the components of neutral words in the direction of the “gender subspace” to mitigate gender stereotypes in word embeddings.

We adapt this methodology to protect user attributes in recommender systems. Consider a user attribute $\mathbf{a}$. Using Canonical Correlation Analysis, we first identify the direction $\mathbf{w}_a$ where the user embeddings projection $\Phi \mathbf{w}_a$ maximally correlates with $\mathbf{a}$. We then project embedded positions of users Φ onto $\mathbf{w}_a^\perp$ the subspace orthogonal to $\mathbf{w}_a$, effectively removing information about $\mathbf{a}$ along this direction.

However, as shown by Gonen and Goldberg (55) and Ravfogel et al (56), sensitive attributes may be encoded along multiple orthogonal directions in the embedding space. A single projection therefore only masks the most prominent direction $\mathbf{w}_a$ identified by the CCA, leaving the attribute recoverable through other directions. To address this limitation, we adopt, as Ravfogel et al (56), an iterative approach. At each step, we identify and project out the direction most correlated with attribute $\mathbf{a}$.

Relying on the null distribution of correlations obtained through permutation tests in Section E, we continue the projections until we can no longer find a statistically significant, with a p-value at 1%, correlation between directions in the embedding space and $\mathbf{a}$.

Importantly, this procedure only removes linearly-encoded information, meaning that attribute $\mathbf{a}$ might still be recovered from the protected embeddings Φ' using non-linear regression models. However, since profile recommendations are formulated through inner products between (translated) user embeddings, our procedure effectively constrains recommendations from relating to attribute $\mathbf{a}$.

In principle, any feature direction can be targeted to be constrained in the embeddings space through the described procedure. In this section, we illustrate its consequences in accuracy and diversity of recommendations using the Left-Right political dimensions. We choose this dimension because it is the leading political dimension structuring national politics, and because it is the one that more clearly falls into the definition of regulation pertaining to data privacy and profiling on sensitive categories such as political opinion and preferences.

G.2 Protocol for the evaluation of accuracy loss of protected embeddings Φ'

We evaluate the protected embeddings Φ' through both attribute correlation and recommendation accuracy. First, we assess the “Who to Follow” predictive accuracy of the protected embeddings Φ' on a 10% held-out test dataset. Second, we analyze the political diversity of recommendations by measuring homophily in the top-k recommended profiles.

For each user $s \in \mathcal{E}$, we score every non-friend user t using the TransE scoring function from Eq. (6): $f(s, w_{tf}, t) = (\phi_s + \varphi_{w_{tf}})^T \phi_t$, where $\varphi_{w_{tf}}$ represents the “Who to Follow” relation embedding. We then select the k highest-scoring users as recommendations. The political diversity of these recommendations is measured as the standard deviation of the Left-Right political leaning users in the top-k recommendations.

Additionally, we seek to assess whether recommendations align with users’ interest in “News & Social Concern” from the topic ontology (37) used in our study, while being constrained to not reflect their political leanings. We measure this alignment as the difference between user s ’s interest in “News & Social Concern” and that of recommended users, where interest is quantified as the fraction of tweets published or shared that were classified as related to “News & Social Concern” by Antypas et al (37) cardiffnlp/tweet-topic-21-multi model. We choose to evaluate the impact on the topic “News &

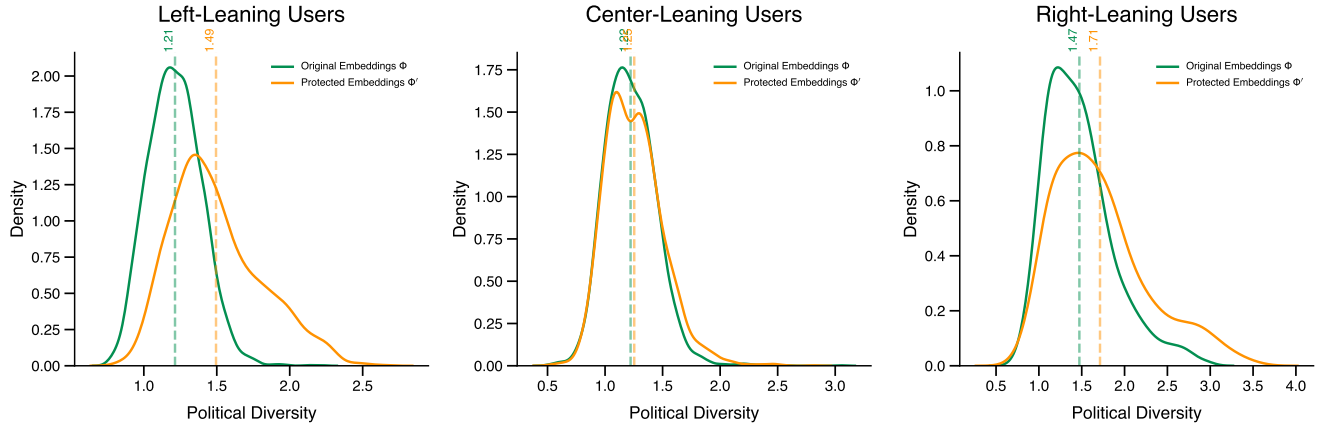


Figure S31: Distribution of political diversity in profile recommendations induced by original unprotected embeddings Φ (in green) and protected embeddings Φ' (in orange), segmented by users' Left-Right leaning. Average values are displayed.

Social Concern” because of its potential alignment with politics. By erasing information about political preferences encoded in the embedding, the risk is that we also erase information about the interest users have in news, thus limiting one the major lever of relevance for contents served to users. In other words, if interest in news and salient Left-Right positions were highly aligned, constraining recommendations to remove the effect of political preferences might also erase information about interests that users have in news, and thus serve recommendations that do not rely on politics, but that are less relevant. In our subsequent analysis we show that this is not the case, and that it is possible to erase information about political preferences for recommendations without affecting the relevance and the weight that interest in news (on the part of users) may have in these recommendations.

To estimate the French Twitter population-level effect size of the intervention, we bootstrapped our embedded population $\mathcal{E}$ to match the Left-Right Leaning distribution observed in the French Twitter population from the dataset of politically-positioned users taken from the dataset (41) (as described in Section D.4.1); displayed in Extended Figure S16. We randomly sampled 250 times while maintaining the sample size to derive 95% confidence intervals for our estimates.

G.3 Results: assessing the effects on accuracy and diversity of recommendations

For the top- $k = 50$ recommended users in the French Twitter population, maximal removal of linearly encoded political information from embedding (intervention strength 1) produced an effect size of $d = 0.477$, 95% CI [0.465, 0.490], relative to the original embedding Φ . In other words, completely removing linearly-encoded Left-Right information increased the Left-Right diversity of recommended users by .477 standard deviations.

Notably, removal of linearly-encoded political information from embeddings demonstrated no significant alteration in the alignment between users' interest in the news and recommended users, relative to intervention-free recommendations. With an effect size of $d = 1.11$, 95% CI [1.06, 1.16], we observe the intervention produces significantly greater increases in political diversity for profile recommendations served to Left-leaning users compared to those served to Center- and Right-leaning users.

This situation mirrors the original distribution of political diversity, where, as displayed in Extended Figure S31, profile recommendations served to Left-leaning users are significantly less diverse than those served to Center- and Right-leaning users. The distribution of political diversity in recommendations induced by the protected embeddings Φ' (intervention strength 1) highlights that, irrespective of users leaning, political diversity in recommendations increases after removing the political information linearly encoded in user embeddings.

We emphasize that the above procedure solely removes linearly encoded Left-Right political leanings from the embeddings Φ . Additionally, we display in Figure, the distribution of inferred gender and age in the audience served with the recommendations of accounts listed in Extended List S9 and Extended List S11, considering the original and political protected embeddings.

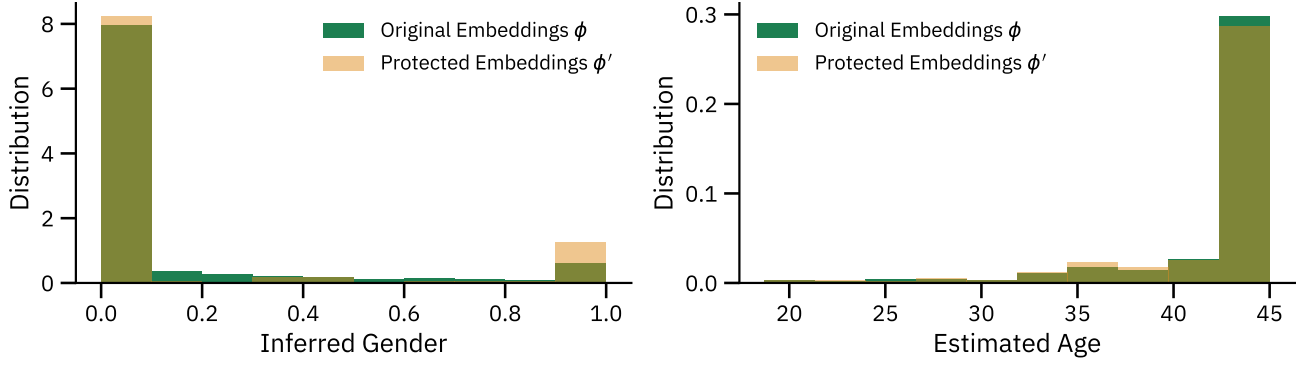


Figure S32: Distribution of gender and age in the population served with the recommendations of listed accounts under the original and political-protected embeddings.

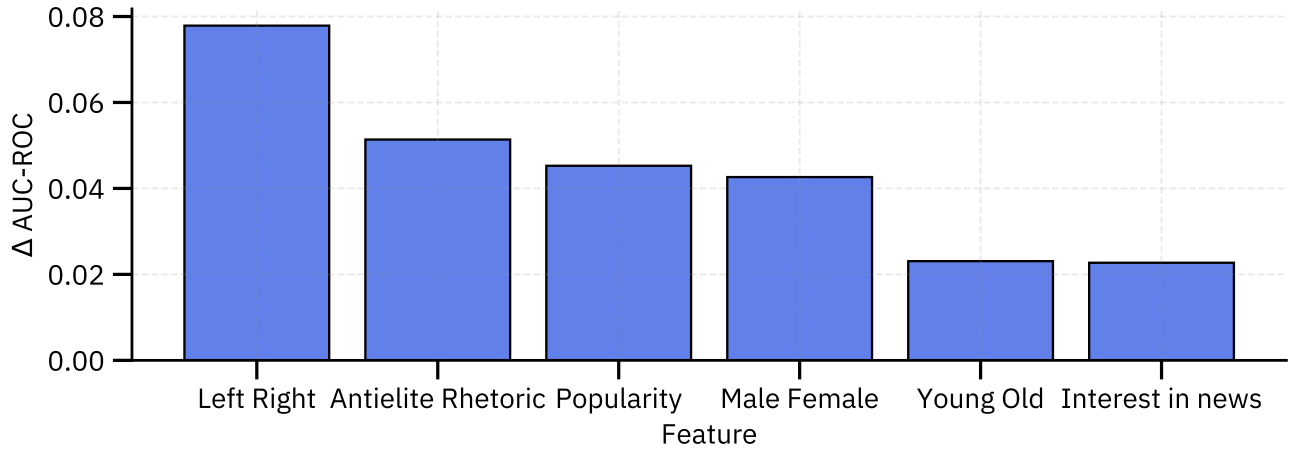


Figure S33: Drop in AUC-ROC in the prediction of “Who To Follow” edges via the protected embedding as a function of the protected user attributes.

G.4 Constraining other attributes information encoded in the recommender system

To assess the robustness and generalizability of the protection protocol, we extend it beyond political information to gender and age attributes. Following the same procedure, we generate two additional sets of embeddings: one where gender cannot be linearly retrieved beyond chance levels, and another for age.

We evaluate these protections by examining the distribution of user attributes among those receiving stereotypical recommendations. Specifically, we identify accounts over-represented in the follows of users inferred as female (Extended List S9) and determine which users receive recommendations for these accounts under different embeddings. We measure disparity using the Wasserstein distance between the gender distribution in the overall population and in the population served these recommendations.

The Wasserstein distance decreases by a factor of 10.4 when using gender-protected embeddings compared to unprotected embeddings, indicating significantly reduced gender disparity in recommendations. Similarly, the Wasserstein distance between the age distribution in the overall population and in the audience served with recommendations for the accounts listed in Extended List S11 is 26.2% smaller for age-protected embeddings compared to original ones.

Finally, we report in Extended Figure S33 the drop in predictive performance of the embedding as a function of which attribute is protected. We observe that Left-Right political leaning emerges as the most important attribute among those tested, with constraining information encoding it leading to the highest performance drop.

References

1. Twitter, Discovering Who To Follow, https://blog.x.com/official/en_us/a/2010/discovering-who-to-follow.html, accessed: 2025-01-06.
2. A. Goel, *et al.*, The who-to-follow system at Twitter: strategy, algorithms, and revenue impact. *Interfaces* **45** (1), 98–107 (2015).
3. P. Gupta, *et al.*, Wtf: The who to follow service at twitter, in *Proceedings of the 22nd international conference on World Wide Web* (2013), pp. 505–514.
4. M. McPherson, L. Smith-Lovin, J. M. Cook, Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology* **27** (1), 415–444 (2001), doi:10.1146/annurev.soc.27.1.415, <http://dx.doi.org/10.1146/annurev.soc.27.1.415>.
5. D. Fogaras, B. Rácz, K. Csalogány, T. Sarlós, Towards Scaling Fully Personalized PageRank: Algorithms, Lower Bounds, and Experiments. *Internet Mathematics* **2** (3), 333–358 (2005), doi:10.1080/15427951.2005.10129104, <http://dx.doi.org/10.1080/15427951.2005.10129104>.
6. R. Lempel, S. Moran, SALSA: the stochastic approach for link-structure analysis. *ACM Transactions on Information Systems* **19** (2), 131–160 (2001), doi:10.1145/382979.383041, <http://dx.doi.org/10.1145/382979.383041>.
7. Twitter, Account Recommendations, <https://help.x.com/en/resources/recommender-systems/account-recommendations> (2023).
8. Twitter, Twitter/The-Algorithm/Twitter’s Recommendation Algorithm, <https://github.com/twitter/the-algorithm/tree/main> (2023).
9. Twitter, Twitter/The-Algorithm/Follow Recommendations Service, <https://github.com/twitter/the-algorithm/tree/72eda9a24f815f6d566818cbf8518138e29d83e9/follow-recommendations-service> (2023).
10. V. Satuluri, *et al.*, Simclusters: Community-based representations for heterogeneous recommendations at twitter, in *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining* (2020), pp. 3183–3193.
11. Twitter, Twitter/The-Algorithm/./Two-hop Random Walk, <https://github.com/twitter/the-algorithm/tree/main/follow-recommendations-service/common/src/main/scala/com/twitter/follow%5Frecommendations/common/candidate%5Fsources/two%5Fhop%5Frandom%5Fwalk> (2023).
12. Twitter, Twitter/The-Algorithm/Follow Recommendations Service/./feature_hydration, https://github.com/twitter/the-algorithm/tree/72eda9a24f815f6d566818cbf8518138e29d83e9/follow-recommendations-service/common/src/main/scala/com/twitter/follow_recommendations/common/feature_hydration (2023).
13. A. El-Kishky, *et al.*, Twihin: Embedding the twitter heterogeneous information network for personalized recommendation, in *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining* (2022), pp. 2842–2850.
14. X. Zhang, *et al.*, Twihin-bert: A socially-enriched pre-trained language model for multilingual tweet representations at twitter, in *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining* (2023), pp. 5597–5607.
15. A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko, Translating Embeddings for Modeling Multi-relational Data, in *Advances in Neural Information Processing Systems*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Weinberger, Eds. (Curran Associates, Inc.), vol. 26 (2013).
16. S. Jolly, *et al.*, Chapel Hill Expert Survey trend file, 1999–2019. *Electoral Studies* **75**, 102420 (2022), doi:10.1016/j.electstud.2021.102420, <http://dx.doi.org/10.1016/j.electstud.2021.102420>.
17. M. Nickel, K. Murphy, V. Tresp, E. Gabrilovich, A Review of Relational Machine Learning for Knowledge Graphs. *Proceedings of the IEEE* **104** (1), 11–33 (2016), doi:10.1109/JPROC.2015.2483592.

18. T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space (2013), <https://arxiv.org/abs/1301.3781>.
19. A. Lerer, *et al.*, PyTorch-BigGraph: A Large-scale Graph Embedding System (2019), <https://arxiv.org/abs/1903.12287>.
20. E. Colleoni, A. Rozza, A. Arvidsson, Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of communication* **64** (2), 317–332 (2014).
21. J.-H. Kang, K. Lerman, Using lists to measure homophily on twitter, in *Workshops at the twenty-sixth AAAI conference on artificial intelligence* (2012).
22. Y. Halberstam, B. Knight, Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter. *Journal of public economics* **143**, 73–88 (2016).
23. Twitter, Twitter/The-Algorithm/./Sims Candidate Source, <https://github.com/twitter/the-algorithm/tree/main/follow-recommendations-service/common/src/main/scala/com/twitter/follow%5Frecommendations/common/candidate%5Fsources/sims> (2023).
24. A. Grover, J. Leskovec, node2vec: Scalable Feature Learning for Networks, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16 (Association for Computing Machinery, New York, NY, USA) (2016), p. 855–864, doi:10.1145/2939672.2939754, <https://doi.org/10.1145/2939672.2939754>.
25. Twitter, Twitter/The-Algorithm/./Triangular Loops Candidate Source, <https://github.com/twitter/the-algorithm/tree/main/follow-recommendations-service/common/src/main/scala/com/twitter/follow%5Frecommendations/common/candidate%5Fsources/triangular%5Floops> (2023).
26. X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, Y. W. Teh, M. Titterton, Eds. (PMLR, Chia Laguna Resort, Sardinia, Italy), vol. 9 of *Proceedings of Machine Learning Research* (2010), pp. 249–256, <https://proceedings.mlr.press/v9/glorot10a.html>.
27. J. Duchi, E. Hazan, Y. Singer, Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research* **12** (61), 2121–2159 (2011), <http://jmlr.org/papers/v12/duchi11a.html>.
28. W. Hu, *et al.*, Open graph benchmark: datasets for machine learning on graphs, in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20 (Curran Associates Inc., Red Hook, NY, USA) (2020).
29. A. Mathai, S. Provost, H. Haubold, *Chapter 10: Canonical Correlation Analysis* (Springer International Publishing), p. 641–677 (2022), doi:10.1007/978-3-030-95864-0_10.
30. X. Corporation, Legacy Verification Policy, <https://help.x.com/en/managing-your-account/legacy-verification-policy> (2023), accessed: 2024-12-02.
31. Z. Wang, *et al.*, Demographic inference and representative population estimates from multilingual social media data, in *The world wide web conference* (2019), pp. 2056–2067.
32. K. Karkkainen, J. Joo, FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation, in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)* (IEEE) (2021), p. 1547–1557, doi:10.1109/wacv48630.2021.00159.
33. I. D. Raji, J. Buolamwini, Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products, in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19 (ACM) (2019), p. 429–435, doi:10.1145/3306618.3314244, <http://dx.doi.org/10.1145/3306618.3314244>.
34. J. Buolamwini, T. Gebru, Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, S. A. Friedler, C. Wilson, Eds. (PMLR), vol. 81 of *Proceedings of Machine Learning Research* (2018), pp. 77–91, <https://proceedings.mlr.press/v81/buolamwini18a.html>.

35. E. Fosch-Villaronga, A. Poulsen, R. Søråa, B. Custers, A little bird told me your gender: Gender inferences in social media. *Information Processing & Management* **58** (3), 102541 (2021), doi:10.1016/j.ipm.2021.102541.
36. S. Volant, Un premier enfant a 28,5 ans en 2015 : 4.5 ans plus tard qu'en 1974. *Insee Premiere* **1642** (2017), <https://www.insee.fr/fr/statistiques/2668280>.
37. D. Antypas, *et al.*, Twitter Topic Classification, in *Proceedings of the 29th International Conference on Computational Linguistics* (International Committee on Computational Linguistics, Gyeongju, Republic of Korea) (2022), pp. 3386–3400.
38. J. Tiedemann, S. Thottingal, OPUS-MT — Building open translation services for the World, in *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT)* (Lisbon, Portugal) (2020).
39. J. Tiedemann, *et al.*, Democratizing neural machine translation with OPUS-MT. *Language Resources and Evaluation* (58), 713–755 (2023), doi:10.1007/s10579-023-09704-w.
40. P. Ramaciotti, J. Royo-Letelier, J.-P. Cointet, A. Pournaki, Attitudinally-positioned European sample dataset, <https://doi.org/10.6084/m9.figshare.25288210.v1> (2024).
41. P. Ramaciotti, J. Royo-Letelier, J.-P. Cointet, A. Pournaki, Attitudinally-positioned European sample dataset. *Social Media for Democracy, Sciences Po Paris* p. 18 p. (2024).
42. J. Clinton, S. Jackman, D. Rivers, The statistical analysis of roll call data. *American Political Science Review* **98** (2), 355–370 (2004).
43. K. Imai, J. Lo, J. Olmsted, Fast estimation of ideal points with massive data. *American Political Science Review* **110** (4), 631–656 (2016).
44. K. T. Poole, H. Rosenthal, A spatial model for legislative roll call analysis. *American journal of political science* pp. 357–384 (1985).
45. P. Barberá, Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political analysis* **23** (1), 76–91 (2015).
46. P. Ramaciotti, *et al.*, Inferring attitudinal spaces in social networks. *Social Network Analysis and Mining* **13** (1), 14 (2022).
47. R. Bakker, S. Jolly, J. Polk, Complexity in the European party space: Exploring dimensionality with experts. *European Union Politics* **13** (2), 219–245 (2012).
48. P. Barberá, J. T. Jost, J. Nagler, J. A. Tucker, R. Bonneau, Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science* **26** (10), 1531–1542 (2015).
49. P. Barberá, G. Rivero, Understanding the political representativeness of Twitter users. *Social Science Computer Review* **33** (6), 712–729 (2015).
50. A. Campbell, P. E. Converse, W. E. Miller, D. E. Stokes, *The american voter* (Wiley) (1960).
51. P. E. Converse, The nature of belief systems in mass publics. *Ideology and Discontent*, pp. 206–261 (1964).
52. P. Ramaciotti, J.-P. Cointet, G. M. Zolotoochin, Unfolding the dimensionality structure of social networks in ideological embeddings, in *Proceedings of the 2021 IEEE/ACM international conference on advances in social networks analysis and mining* (2021), pp. 333–338.
53. L. Tunstall, *et al.*, Zephyr: Direct Distillation of LM Alignment (2023).
54. T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, A. T. Kalai, Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings, in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett, Eds. (Curran Associates, Inc.), vol. 29 (2016).
55. H. Gonen, Y. Goldberg, Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them, in *Proceedings of the 2019 Workshop on Widening NLP*, A. Axelrod, D. Yang, R. Cunha, S. Shaikh, Z. Waseem, Eds. (Association for Computational Linguistics, Florence, Italy) (2019), pp. 60–63, <https://aclanthology.org/W19-3621>.

56. S. Ravfogel, Y. Elazar, H. Gonen, M. Twiton, Y. Goldberg, Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, J. Tetreault, Eds. (Association for Computational Linguistics, Online) (2020), pp. 7237–7256, doi:10.18653/v1/2020.acl-main.647.