



Political audience diversity and news reliability in algorithmic ranking

Saumya Bhadani¹, Shun Yamaya², Alessandro Flammini³, Filippo Menczer³,
Giovanni Luca Ciampaglia¹✉ and Brendan Nyhan⁴

Newsfeed algorithms frequently amplify misinformation and other low-quality content. How can social media platforms more effectively promote reliable information? Existing approaches are difficult to scale and vulnerable to manipulation. In this paper, we propose using the political diversity of a website's audience as a quality signal. Using news source reliability ratings from domain experts and web browsing data from a diverse sample of 6,890 US residents, we first show that websites with more extreme and less politically diverse audiences have lower journalistic standards. We then incorporate audience diversity into a standard collaborative filtering framework and show that our improved algorithm increases the trustworthiness of websites suggested to users—especially those who most frequently consume misinformation—while keeping recommendations relevant. These findings suggest that partisan audience diversity is a valuable signal of higher journalistic standards that should be incorporated into algorithmic ranking decisions.

Concerns continue to grow about the prevalence of misinformation on social media platforms^{1,2}, including during the recent coronavirus disease 2019 pandemic³. These types of content often exploit people's tendency to prefer pro-attitudinal information⁴, which can be exacerbated by platform content recommendations^{5,6}. In this paper, we explore a possible algorithmic approach to mitigate the spread of misinformation and promote content with higher journalistic standards online.

Social media platform recommendation algorithms frequently amplify bias in human consumption decisions. Though the information diets of Americans are less slanted in practice than many assume, the people who consume the most political news are most affected by the tendency towards selective exposure⁷. As a result, the news audience is far more polarized than the public as a whole^{8,9}. Although the prevalence of so-called fake news online is rather limited and concentrated amongst relatively narrow audiences^{7,10–14}, content that generally appeals to these tendencies—which does include low-quality or false news—may generate high levels of readership or engagement², prompting algorithms that seek to maximize engagement to distribute them more widely.

Prior research indicates that existing recommendation algorithms tend to promote items that have already achieved popularity^{15,16}. This bias may have several effects on the consumption of low-quality and false news. First, sorting the news by (either predicted or achieved) engagement can exacerbate polarization by increasing in-group bias and discouraging consumption amongst out-group members¹⁷. Second, it may contribute to information cascades, amplifying differences in rankings from small variations or random fluctuations and degrading the overall quality of information consumed by users^{18–22}. Third, exposure to engagement metrics makes users more likely to share and less likely to fact-check highly engaging content from low-credibility sources, increasing vulnerability to misinformation²³. Finally, popularity bias in recommendation systems can create socio-algorithmic vulnerabilities to threats such as automated amplifiers, which exploit algorithmic

content rankings to spread low-quality and inflammatory content to like-minded audiences^{24,25}.

Given the speed and scale of social media, assessing directly the quality of every piece of content or the behaviour of each user is infeasible. Online platforms are instead seeking to include signals about news quality in their content recommendation algorithms^{26,27}, for example, by extracting information from trusted publishers²⁸ or by means of linguistic patterns analysis^{29,30}. More generally, a vast literature examines how to assess the credibility of online sources^{31,32} and the reputations of individual online users^{33,34}, which could in principle bypass the problem of checking each individual piece of content. Unfortunately, many of these methods are hard to scale to large groups and/or depend upon context-specific information about the type of content being generated. For example, methods for assessing the credibility of content on Wikipedia often assume that content is organized as a wiki. As a result, they are not easily applied to news content recommendations on social media platforms.

Another approach is to try to evaluate the quality of articles directly³⁵, but scaling such an approach would likely be costly and cause lags in the evaluation of novel content. Similarly, while crowd-sourced website evaluations have been shown to be generally reliable in distinguishing between high- and low-quality news sources³⁶, the robustness of such signals to manipulation is yet to be demonstrated.

Building on the literature about the benefits of diversity at the group level^{37,38}, we propose using the partisan diversity of the audience of a news source as a signal of its quality. This approach has two key advantages. First, audience partisan diversity can be computed at scale given that information about the partisanship of users is available or can be inferred in a reliable manner. Second, because diversity is a property of the audience and not of its level of engagement, it is less susceptible to manipulation if one can detect inauthentic partisan accounts^{39–42}. These two conditions (inferring partisanship reliably and preventing abuse by automated amplification/deception) could easily be met by the major social media

¹Department of Computer Science and Engineering, University of South Florida, Tampa, FL, USA. ²Department of Political Science, Stanford University, Stanford, CA, USA. ³Observatory on Social Media, Indiana University, Bloomington, IN, USA. ⁴Department of Government, Dartmouth College, Hanover, NH, USA. ✉e-mail: gciampaglia@gmail.com

platforms, which have routine access to a wealth of signals about their users and their authenticity.

We evaluate the merits of our proposed approach using data from two sources: a comprehensive data set of web traffic history from 6,890 US residents, collected along with surveys of self-reported partisan information from respondents in the YouGov Pulse survey panel, and a dataset of 3,765 news source reliability scores compiled by trained experts in journalism and provided by NewsGuard⁴³. We first establish that domain pageviews are not associated with overall news reliability, highlighting the potential problem with algorithmic recommendation systems that rely on popularity and related metrics of engagement. We next define measures of audience partisan diversity and show that these measures correlate with news reliability better than popularity does. Finally, we study the effect of incorporating audience partisan diversity into algorithmic ranking decisions. When we create a variant of the standard collaborative filtering algorithm that explicitly takes audience partisan diversity into account, our new algorithm provides more trustworthy recommendations than the standard approach, with only a small loss of relevance, suggesting that reliable sources can be recommended without the risk of jeopardizing user experience.

These results demonstrate that diversity in audience partisanship can serve as a useful signal of news reliability at the domain level, a finding that has important implications for the design of content recommendation algorithms used by online platforms. Although the news recommendation technologies deployed by platforms are more sophisticated than the approach tested here, our results highlight a fundamental weakness of algorithmic ranking methods that prioritize content that generates engagement and suggest a new metric that could help improve the reliability of the recommendations that are provided to users.

Results

Popularity does not predict news reliability. To motivate our study, we first demonstrate that the popular news content that algorithmic recommendations often highlight is not necessarily reliable. To do so, we assess the relationship between source popularity and news reliability. We measure source popularity using YouGov Pulse traffic data. Due to skew in audience size amongst domains, we transform these data to a logarithmic scale. In practice, we measure the popularity of a source in two ways: as (the log of) the number of users, and as (the log of) the number of visits, or pageviews. News reliability is instead measured using NewsGuard scores (Data). Figure 1 shows that the popularity of a news source is at best weakly associated with its reliability (see Supplementary Table 2 for a full regression summary). At the user level (Fig. 1, left), the overall Pearson correlation is $r(n=1,024)=0.03$ (two-sided $P=0.36$, 95% confidence interval (CI) -0.01 to 0.09). At the pageview level (Fig. 1, right), $r(n=1,024)=0.05$ (two-sided $P=0.12$, 95% CI -0.01 to 0.11). Bootstrapped equivalence tests at the 0.05 significance level indicate that we can reject Pearson correlation coefficients larger than 0.096 at the visitor level and 0.094 at the pageview level.

The association between the two variables remains weak even if we divide sources based on their partisanship. When measuring popularity at the user level, websites that have a predominantly Democratic audience have a significant positive association ($r(n=783)=0.09$, two-sided $P=0.02$, 95% CI 0.02 to 0.16), but for websites with a Republican audience, the correlation is negative and not significant at conventional standards ($r(n=237)=-0.12$, two-sided $P=0.06$, 95% CI -0.25 to 0.005). A similar pattern holds at the pageview level, with a weak, positive and insignificant association for websites with predominantly Democratic audiences ($r(n=702)=0.07$, two-sided $P=0.07$, 95% CI -0.01 to 0.14) and a weak, negative and non-significant association for those with predominantly Republican audiences ($r(n=322)=-0.01$, two-sided $P=0.90$, 95% CI -0.10 to 0.12). Bootstrapped equivalence tests at the

0.05 significance level for websites with predominantly Democratic audiences reject correlation coefficients larger than 0.127. Similarly, for websites with a predominantly Republican audiences, we can calculate equivalence bounds of $(-0.222, 0)$ and $(-0.078, 0.092)$ at the visitor and pageview level, respectively. Overall, these results suggest that the strength of association between the two variables is quite weak even after taking into account for the partisan traffic of a website.

Audience partisan diversity is signal of reliable news. In contrast, we observe that sites with greater audience partisan diversity tend to have higher NewsGuard scores while those with lower levels of diversity, and correspondingly more homogeneous partisan audiences, tend to have lower reliability scores. As our primary metric of diversity, we selected from a range of alternative definitions (Definition of audience partisan diversity section) the variance of the partisanship distribution. Figure 2 shows how NewsGuard scores vary with both mean audience partisanship and the variance in audience partisanship.

As Fig. 2 indicates, unreliable websites with very low NewsGuard scores are concentrated in the tails of the distribution, where partisanship is most extreme and audience partisan diversity is, by necessity, very low. This relationship is not symmetrical. Low-reliability websites (whose markers are darker shades of blue in the figure) are especially concentrated in the right tail, which corresponds to websites with largely Republican audiences. The data in Fig. 2 also suggest that the reliability of a website may be associated not just with the variance of the distribution of audience partisanship slants but also with its mean. To account for this, we first compute the coefficient of partial correlation between NewsGuard reliability scores and the variance of audience partisanship given the mean audience partisanship of each website. Compared with popularity, we find a stronger (and significant) correlation regardless of whether mean partisanship and audience partisan diversity are calculated by weighting individual audience members equally (user level, left panel: partial correlation $r(n=1,024)=0.38$, two-sided $P<10^{-4}$, 95% CI 0.32 to 0.43) or by how often they visited a given site (pageview level, right panel: partial correlation $r(n=1,024)=0.22$, two-sided $P<10^{-4}$, 95% CI 0.16 to 0.28).

Aside from mean partisanship, a related, but potentially distinct, confounding factor is the extremity of the partisanship slants distribution (that is, the distance of the average partisanship of a website visitor on a 1–7 scale from the midpoint of 4, which represents a true independent). We thus computed partial correlation coefficients again but instead keep the ideological extremity of website audiences constant instead of the mean. Our results are consistent using this approach (user level: $r(n=1,024)=0.26$, $P<10^{-4}$, 95% CI 0.20 to 0.31 ; pageview level: $r(n=1,024)=0.15$, $P<10^{-4}$, 95% CI 0.08 to 0.21 ; both tests are two-sided).

Finally, we test whether bimodal distributions of audience partisanship are associated with quality. This test is important to conduct because unimodal and bimodal distributions may have the same variance. In Supplementary Information Sect. S6, we define a metric for bipolarity and find that audience bipolarity is a much weaker signal of quality than partisan diversity as measured by the variance.

We study the diversity–reliability relationship in more detail in Fig. 3, which differentiates between websites with audiences that are mostly Republican and those with audiences that are mostly Democratic. Consistent with what we report above, Fig. 3 shows that audience partisan diversity is positively associated with news reliability (see Supplementary Table 3 for a full regression summary). Again, this relationship holds both when individual audience members are weighted equally (user level, left panel) and when they are weighted by their number of accesses (pageview level, right panel). In addition, we find that the relationship is stronger for sites whose average visitor identifies as a Republican (standardized

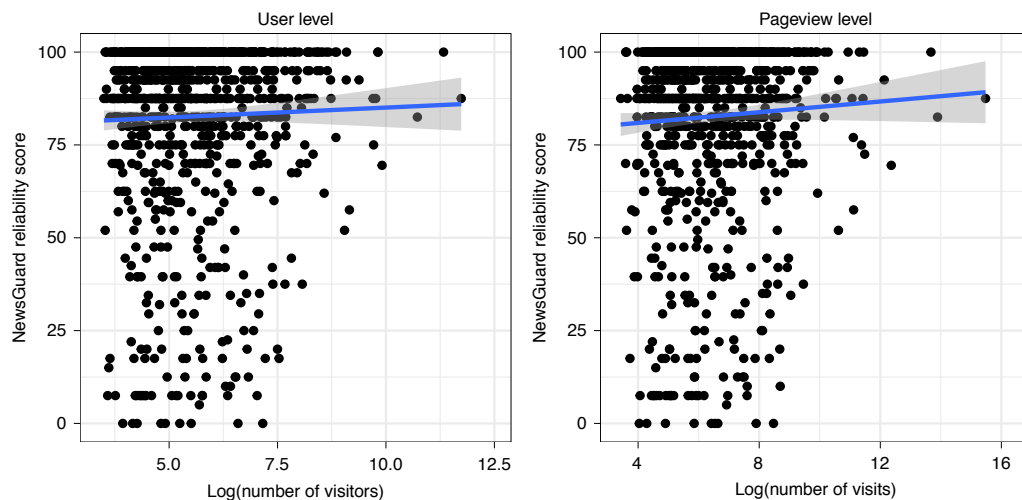


Fig. 1 | Relationship between (log-transformed) audience size and news reliability by domain (solid blue line, $N = 1,024$ domains). Audience size as number of individual visitors (left) and visits (right), with 95% CI (shaded area). Note that the two panels use different scales on the x axis and do not start at zero. Reliability scores provided by NewsGuard⁴³. See Supplementary Table 2 for full regression results.

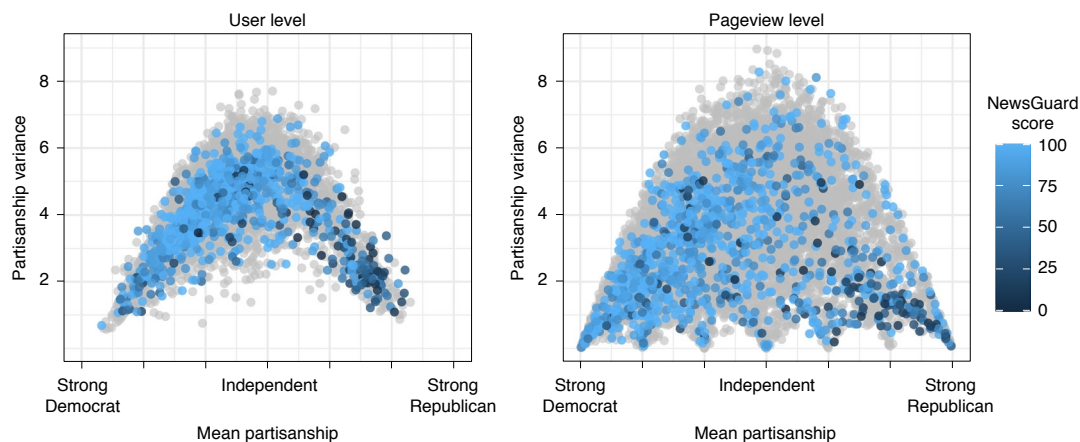


Fig. 2 | Average audience partisanship versus variance ($N = 11,793$ domains). Audience size as number of individual visitors (left) and visits (right). Domains for which we have NewsGuard reliability scores⁴³ are shaded in blue (where darker shades equal lower scores). Domains with no available score are plotted in grey.

ordinary least squares coefficient of Republican domains: $\beta = 10.6$ (0.94) at user level; $\beta = 8.80$ (1.05) at pageview level) versus those whose average visitor identifies as a Democrat (standardized ordinary least squares coefficient of Democrat domains: $\beta = 2.93$ (0.66) at user level; $\beta = 0.82$ (0.86) at pageview level), which is consistent with Fig. 2 (the partisan slope difference is 7.71 at user level, $P < 10^{-4}$, 95% CI 5.46 to 9.97; 7.97 at pageview level, $P < 10^{-4}$, 95% CI 5.32 to 10.62).

These results are not affected by popularity. Partisan diversity is weakly correlated with popularity, regardless of the operational definition of either measure (Supplementary Table 4). In fact, the association between diversity and NewsGuard reliability scores is consistent even when controlling for popularity (user level: $r(n = 1,024) = 0.34$, two-sided $P < 10^{-4}$, 95% CI 0.29 to 0.40; pageview level: $r(n = 1,024) = 0.17$, two-sided $P < 10^{-4}$, 95% CI 0.11 to 0.23), suggesting that diversity could contribute to detecting quality over and above the more typical popularity metrics used by social media algorithms. However, the previous analysis of Fig. 3 shows that the overall relationship masks significant heterogeneity

between websites with mostly Republican or Democratic audiences. To tease apart the contributions of popularity from those of partisanship, we estimate a full multivariate regression model. After controlling for both popularity and political orientation, we find qualitatively similar results. Full regression summaries can be found in Supplementary Tables 2 and 3.

As mentioned above, variance in audience partisanship is not the only possible way to define audience partisan diversity. Alternative definitions (for example, entropy) can be used (Definition of audience partisan diversity section). As a robustness check, we therefore consider a range of alternative definitions of audience partisan diversity and obtain results that are qualitatively similar to the ones presented here, though results are strongest for variance (Supplementary Table 1).

Audience diversity produces trustworthy relevant rankings. To understand the potential effects of incorporating audience partisan diversity into algorithmic recommendations, we next consider how recommendations from a standard user-based collaborative

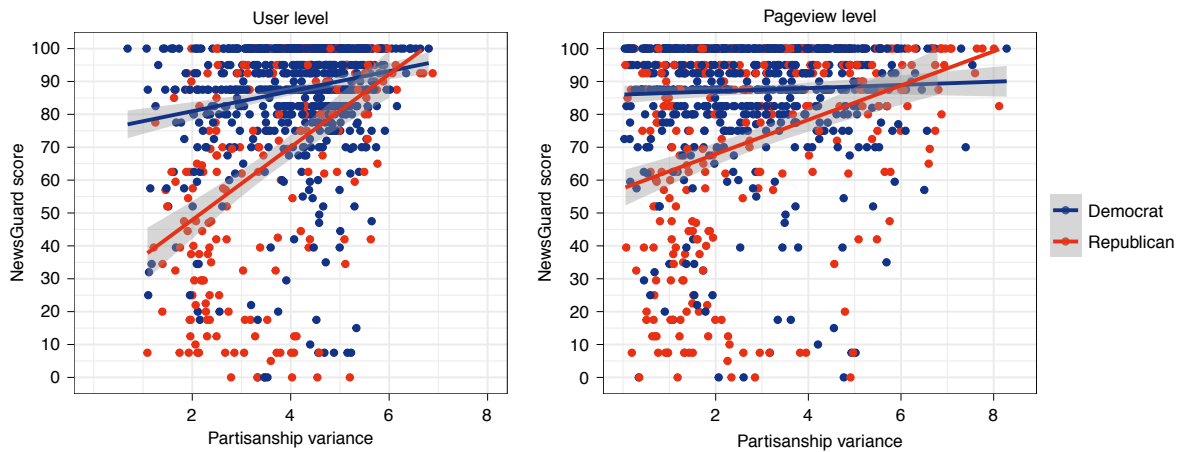


Fig. 3 | Relationship between audience partisan diversity and news reliability for websites whose average visitor is a Democrat (solid blue line) or a Republican (solid red line). Variance computed at user (left, $N=1,020$ domains) or pageview level (right, $N=1,024$ domains), with 95% CI (shaded area). News reliability scores from NewsGuard⁴³. See Supplementary Table 2 for full regression results.

filtering (CF) algorithm^{44,45} change if we include audience partisan diversity as an additional signal. We call this modified version of the algorithm collaborative filtering + diversity (CF+D) (see Audience diversity and collaborative filtering section for a formal definition).

In classic CF, users are presented with recommendations drawn from a set of items (in this case, web domains) that have been ‘rated’ highly by those other users whose tastes are most similar to theirs. Lacking explicit data about how a user would ‘rate’ a given web domain, we use a quantity derived from the number of user pageviews to a domain (based on term frequency-inverse document frequency (TF-IDF); see also Audience diversity and collaborative filtering section) as the rating.

To evaluate our method, we follow a standard supervised learning workflow. We first divide web traffic data for each user in the YouGov Pulse panel into training and testing sets by domain (Supervised learning evaluation workflow section). We then compute similarities in traffic patterns between users for all domains in the training set (not just news websites) and use the computed similarities to predict the aforementioned domain-level pageviews metric on the test set. The domains that receive the highest predicted ratings (that is, expected TF-IDF-transformed pageviews) are then selected as recommendations. As a robustness check, we obtain consistent results if we split the data longitudinally instead of randomly (that is, as a forecasting exercise; see Supplementary Figs. 7 and 8 for details).

Note that, if a user has not visited a domain, then the number of visits for that domain will be zero. In general, due to the long tail in user interests¹⁵, we cannot infer that the user has a negative preference towards a website just because they have not visited it. The user may simply be unaware of the site. We therefore follow standard practice in the machine learning literature in only evaluating recommendations for content for which we have ratings (that is, visits in the test set), though in practice actual newsfeed algorithms rank items from a broader set of inputs, which typically includes content that the user may not have seen (for example, content shared by friends⁵).

To produce recommendations for a given user, we consider all the domains visited by the user in the test set for which ratings are available from one or more respondents in a neighbourhood of most similar users (domains with no neighbourhood rating are discarded since neither CF nor CF+D can make a prediction for them; Audience diversity and collaborative filtering section) and for which we have a NewsGuard score (that is, a reliability score). We then rank those domains by their rating computed using either

CF or CF+D. This process produces a ranked list of news domains and reliability scores from both the standard CF algorithm and the CF+D algorithm, which has been modified to incorporate the audience partisan diversity signal. We evaluate these lists using two different measures of trustworthiness which are computed for the top k domains in each list: the mean score (a number in the 0–100 range) and the proportion of domains with a score of 60 or higher, which NewsGuard classifies as indicating that a site ‘generally adheres to basic standards of credibility and transparency’^{7,43} (Trustworthiness metrics section).

By varying the number of top domains k , we can evaluate how trustworthiness changes as the length of the list of recommendations increases. In Fig. 4, we plot the trustworthiness of the recommended domains as a function of k . We restrict values of k to 1–28, the values for which there are at least 100 users in each bin (see Supplementary Fig. 2 for the plot spanning the full range). Each panel compares the average trustworthiness of domains ranked by CF and CF+D with two baselines. The first is the trustworthiness of websites that users visited in the test set, ranked by their TF-IDF-transformed number of visits (that is, pageviews). This baseline captures the trustworthiness of the websites that users in the YouGov Pulse panel actually visited after adjusting for the fact that more popular websites tend to attract more visits in general. The second baseline is the trustworthiness of recommendations produced according to the overall popularity of domains. This baseline does not include any local information about user–user similarities and thus can be seen as a ‘global’ measure of popularity with no contribution due to user personalization (Recommendation based on global popularity section).

We observe in Fig. 4 that the trustworthiness of recommendations produced by CF+D is significantly better than standard CF recommendations, global popularity recommendations and baseline statistics from user behaviour. In particular, CF produces less trustworthy rankings than both the recommendations based on global popularity and those based on user visits (for small values of k , the difference is within the margin of error). In contrast, CF+D produces rankings that are more trustworthy than CF and either baseline (global popularity or actual visits) across different levels of k . These results suggest that audience partisan diversity can provide a valuable signal to improve the reliability of algorithmic recommendations.

Of course, the above exercise would be meaningless if our proposed algorithm recommended websites that do not interest users. Because CF+D alters the set of recommended domains to prioritize those visited by more diverse partisan audiences, it may be

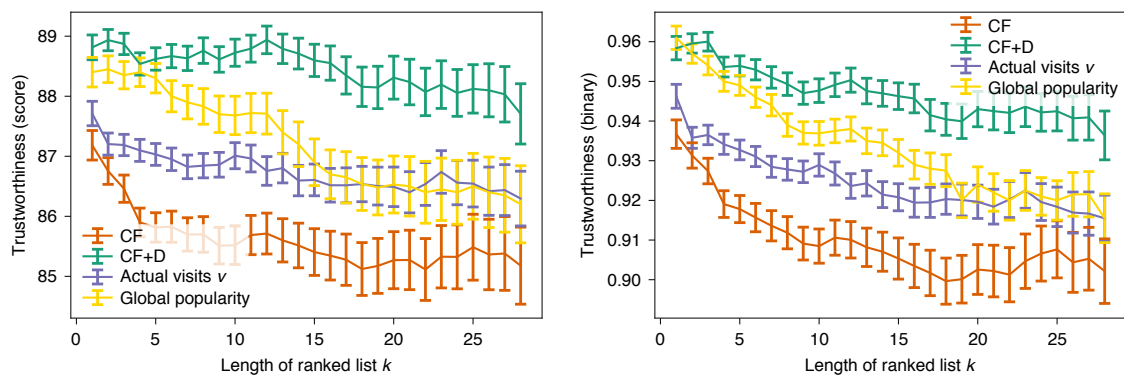


Fig. 4 | Trustworthiness of recommended domains by length of ranked list k ($N_k = 28$ list lengths). Trustworthiness based on scores from NewsGuard (left)⁴³. Proportion of domains labelled as ‘trustworthy’, also by NewsGuard (right). Actual visits v are normalized using TF-IDF (Audience diversity and collaborative filtering section). Global popularity is overall domain popularity (Recommendation based on global popularity section). Each bin represents the average computed on the top k recommendations for all users in the YouGov panel with $\geq k$ recommendations in their test sets. Bars represent the s.e.m. The values of k are capped so that each bin has ≥ 100 users in it (see Supplementary Fig. 2 for plot with all values of k). In this figure, both CF and CF+D compute the similarity between users using the Kendall τ correlation coefficient (Audience diversity and collaborative filtering section). We obtain qualitatively similar results using the Pearson correlation coefficient (Supplementary Fig. 4).

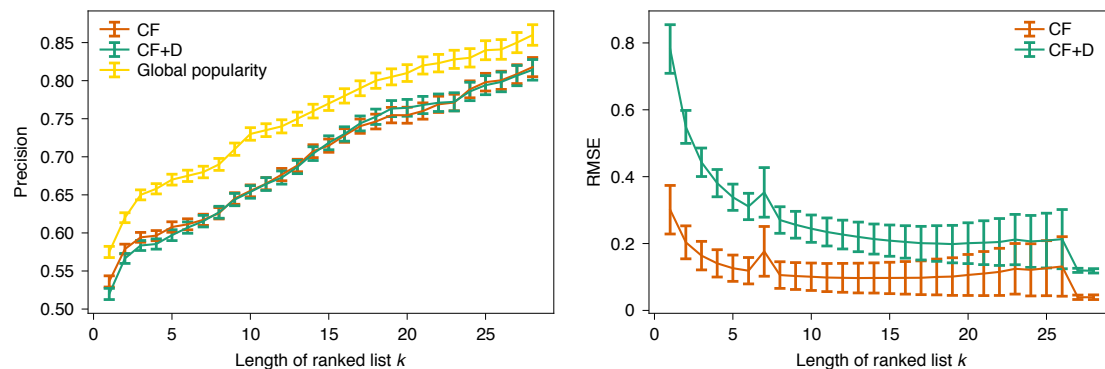


Fig. 5 | Accuracy of domain recommendations by length of ranked list k ($N_k = 28$ list lengths). Left: Precision (proportion of correctly ranked sites) by length of ranked list k (higher is better). Right: RMSE of predicted pageviews for top k ranked domains by length of ranked list k (lower is better). Each bin represents the average computed on the top k recommendations of all users with $\geq k$ recommendations in their test sets. Bars represent the s.e.m. The values of k are capped so that each bin has ≥ 100 users in it (see Supplementary Fig. 3 for plot with all values of k). In this figure, both CF and CF+D compute the similarity between users using the Kendall τ correlation coefficient (Audience diversity and collaborative filtering section). We obtain qualitatively similar results using the Pearson correlation coefficient (Supplementary Fig. 5).

suggesting sources that offer counter-attitudinal information or that users do not find relevant. In this sense, CF+D could represent an audience-based analogue of the topic diversification strategy from the recommender systems literature⁴⁶. If so, a loss of predictive ability would be expected.

Figure 5 compares the accuracy of CF+D in predicting user visits to domains in the test set with that of CF. To evaluate accuracy, we compute both the fraction of correctly predicted domains (precision) and the root-mean-squared error (RMSE) as a function of the number of recommended domains k (see Accuracy metrics section for definitions). Note that precision improves with k (left panel) by definition—as k grows, we are comparing an increasingly large set of recommendations with a list of fixed size. Because each bin averages over users with at least k domains in their test set, when k reaches the maximum size of the recommendation list that we can make, the precision necessarily becomes 100%. Note that the plots in Fig. 5 do not reach this level (they include only bins with at least 100 users in them) but trend upward with k (see Supplementary Fig. 3 for results for all values of k).

As with precision, RMSE declines with k (Fig. 5, right) since we focus progressively on users with longer lists and thus more training data. Like in the left panel, each bin in the right panel averages over users with at least k domains in their test set. Unlike precision, however, RMSE is more prone to producing outliers because it does not depend on the relative ranking of item ratings but instead on their magnitude. This difference is reflected in the sudden drop in the error bars for the RMSE at $k=27$ due to the presence of a single user with a maximum list length of 26 domains in testing. We manually checked the data of this user and found that the training set included only domains visited infrequently, leading to large errors. Removing this outlier eliminated the observed change.

To provide intuition about the contribution of popularity in recommendations, Fig. 5 (left) also shows the precision of the naïve baseline obtained by ranking items by their global popularity. This baseline outperforms CF and CF+D but at the price of providing the same set of recommendations to all users (that is, the results are not personalized) and of providing recommendations of lower trustworthiness (Fig. 4). Note that the RMSE cannot be computed

for this baseline because this metric requires knowledge of the rating of a domain, not just of its relative ranking.

Our results are generally encouraging. In both cases, precision is low and RMSE is high for low values of k , but error levels start to stabilize around $k=10$, which suggests that making correct recommendations for shorter lists (that is, $k < 10$) is more challenging than for longer ones. Moreover, when we compare CF+D with CF, accuracy declines slightly for CF+D relative to CF, but the difference is not statistically significant for all but small values of k , suggesting that CF+D is still capable of producing relevant recommendations.

Re-ranking items by diversity has minimal effects on predictive accuracy, but how does it affect user satisfaction? The recommendations produced by CF+D would be useless if users did not find them engaging. Unfortunately, we lack data about user satisfaction in the YouGov panel. Our primary metric (log number of website visits) cannot be interpreted as a pure measure of satisfaction because other factors of course shape the decision by users in the YouGov panel to visit a website, including social media recommendations themselves.

However, it is possible that more accurate recommendations will result in higher user satisfaction. To quantify the significance of the observed drop in accuracy due to re-ranking by diversity, we simulated the sampling distribution of the precision of recommendations obtained after re-ranking. We do so by re-shuffling domain labels in the ranked list produced by CF+D while maintaining the sequence of predicted ratings fixed. We then compute precision on this reshuffled list. Repeated multiple times, this procedure allows us to calculate the probability, due to random chance alone, of a drop in precision (relative to CF) as small as the observed one. Compared with this null model, we find that our results lead to significantly higher precision. Most random re-rankings of the same magnitude as the one produced by CF+D would result in lower precision than what we observe. We report the results of this additional analysis in Supplementary Fig. 9.

Audience diversity and misinformation exposure. The results above demonstrate that incorporating audience partisan diversity can increase the trustworthiness of recommended domains while still providing users with relevant recommendations. However, we know that exposure to unreliable news outlets varies dramatically across the population. For instance, exposure to untrustworthy content is highly concentrated amongst a narrow subset of highly active news consumers with heavily slanted information diets^{11,14}. We therefore take advantage of the survey and behavioural data available on participants in the Pulse panel to consider how CF+D effects vary by individual partisanship (self-reported via survey), behavioural measures such as volume of news consumption activity and information diet slant, and contextual factors that are relevant to algorithm performance such as similarity to other users.

In this section, we again produce recommendations using either CF or CF+D and measure their difference in trustworthiness with respect to a baseline based on user visits (specifically the ranking by TF-IDF-normalized number of visits v ; Audience diversity and collaborative filtering section). However, we analyse the results differently than those reported above. Rather than considering recommendations for lists of varying length k , we create recommendations for different subgroups based on the factors of interest and compare how the effects of the CF+D approach vary between those groups.

To facilitate comparisons in performance between subgroups that do not depend on list length k , we define a new metric to summarize the overall trustworthiness of the ranked lists obtained with CF and CF+D over all possible values of k . Since users tend to pay less attention to items ranked lower in the list⁴⁷, it is reasonable to assume that lower-ranked items ought to contribute less to the overall trustworthiness of a given ranking.

Let us now consider probabilistic selections from two different rankings, represented by random variables X and X' , where X

is the random variable of the ranking produced by one of the two recommendation algorithms (either CF or CF+D) and X' is the selection from the baseline ranking based on user visits. Using a probabilistic discounting method (equation (8)), we compute the expected change in trustworthiness Q from switching the selection from X' to X as

$$\Delta Q = E[Q(X)] - E[Q(X')], \quad (1)$$

where the expectations of $Q(X)$ and $Q(X')$ are taken with regard to the respective rankings (Discounting via ranking section). A value of $\Delta Q > 0$ indicates that algorithmic recommendations are more trustworthy than what users actually accessed. If $\Delta Q < 0$, the trustworthiness of a ranked list is lower than the baseline from user visits. (To ensure that the results below are not affected by the discounting method we employ, we report qualitatively similar results obtained without any discounting for a selection of values of k in Supplementary Figs. 10–16.)

Applying equation (1), we find that CF+D substantially increases trustworthiness for users who tend to visit sources that lean conservative (Fig. 6a) and for those who have the most polarized information diets (in either direction; Fig. 6c), two segments of users who are especially likely to be exposed to unreliable information^{10,11,14}. In both cases, CF+D achieves the greatest improvement amongst the groups where CF reduces the trustworthiness of recommendations the most, which highlights the pitfalls of algorithmic recommendations for vulnerable audiences and the benefits of prioritizing sources with diverse audiences in making recommendations to those users.

Note that, even though the YouGov sample includes self-reported information on both party identification and partisanship of respondents, we use only the former variable (Fig. 6b) for stratification to avoid circularity given the definition of CF+D, which relies on the latter. In Fig. 6a,c, we instead stratify on an external measure of news diet slant (calculated from a large sample of social media users; Stratification analysis section).

We also observe that CF+D has strong positive effects for users who identify as Republicans or lean Republican (Fig. 6b) and for those who are the most active news consumers in terms of both total consumption (Fig. 6d) and number of distinct sources (Fig. 6e). Furthermore, since the two recommendation schemes considered here (CF and CF+D) are predicated on identifying similar users according to their tastes and behaviours, we also segment the users of the YouGov sample according to the degree of similarity with their nearest neighbours (identified based on Kendall's rank correlation coefficient between user vectors; Audience diversity and collaborative filtering section). Stratifying on the average of nearest-neighbour similarities, we find that CF+D results in improvements for the users whose browsing behaviour is most similar to others in their neighbourhood and who might thus be most at risk of 'echo chamber' effects (Fig. 6f). Finally, when we group users by the trustworthiness of the domains they visit, we find that the greatest improvements from the CF+D algorithm occur for users who are exposed to the least trustworthy information (Fig. 6g). By contrast, the standard CF algorithm often recommends websites that are less trustworthy than those that respondents actually visit ($\Delta Q < 0$).

Discussion

The findings presented here suggest that the ideological diversity of the audience of a news source is a reliable indicator of its journalistic quality. To obtain these findings, we combined source reliability ratings compiled by expert journalists with traffic data from the YouGov Pulse panel. Of course, we are not the first to study the information diets of internet users. Prior work has leveraged web traffic data to pursue related topics such as identifying potential dimensions of bias of news sources^{16,48}, designing methods to

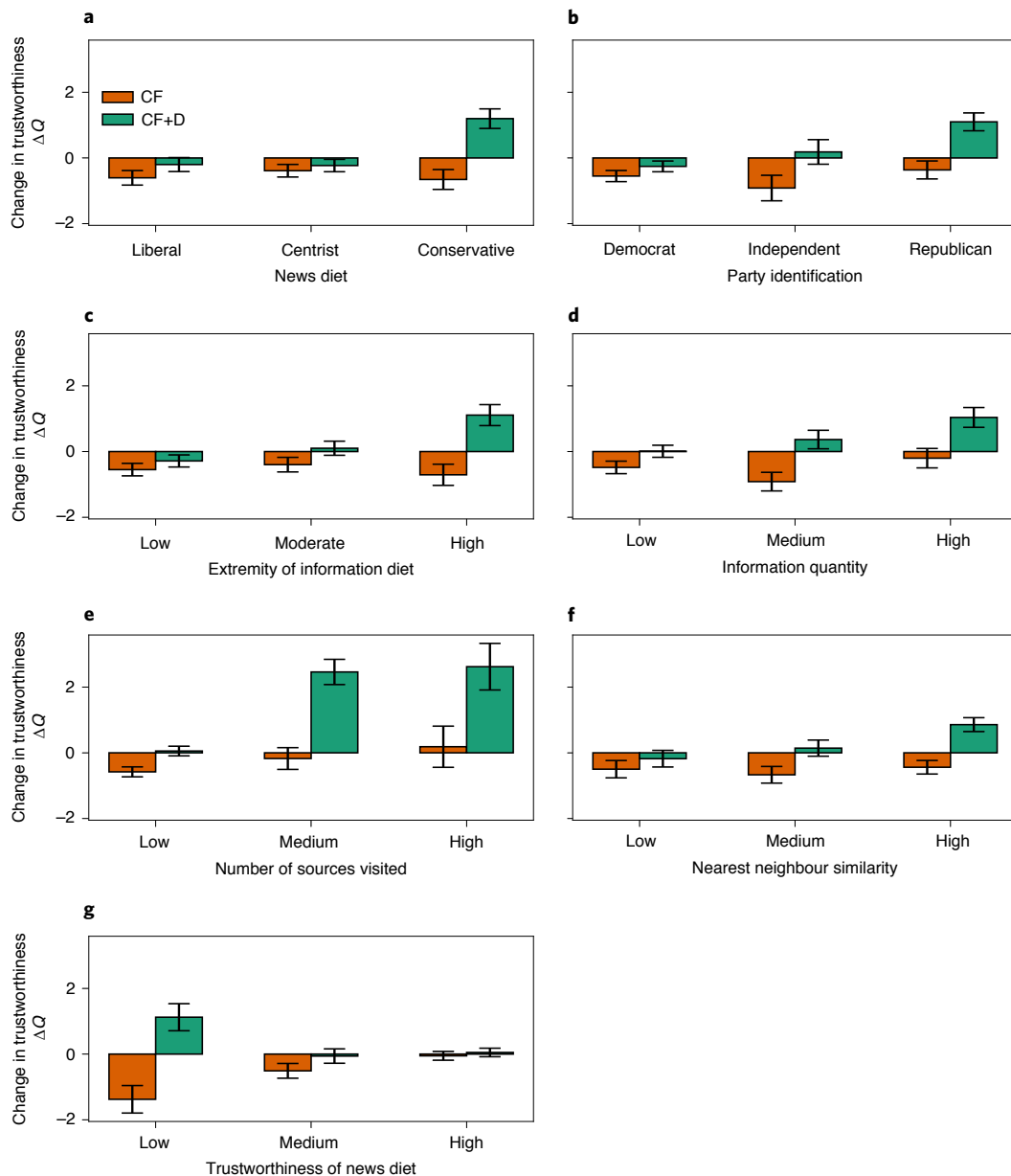


Fig. 6 | Effect of CF and CF+D versus actual visits baseline on trustworthiness by user characteristics and behaviour. **a**, Ideological slant of visited domains (terciles using scores from Bakshy et al. ⁵). **b**, Self-reported party identification from YouGov Pulse responses as measured on a 7-point scale (1-3, Democrats including people who lean Democrat but do not identify as Democrats; 4, independents; 5-7, Republicans including people who lean Republican but do not identify as Republicans). **c**, Absolute slant of visited domains (terciles using scores from Bakshy et al. ⁵). **d**, Total online activity (TF-IDF-transformed pageviews; terciles). **e**, Distinct number of domains visited (terciles). **f**, Average user-user similarity with nearest $n=10$ neighbours in training set (terciles). **g**, Trustworthiness of domains visited by users (in training set; terciles). Bars represent the s.e.m. of each stratum. Change in trustworthiness ΔQ based on scores from NewsGuard⁴³.

present diverse political opinions^{49,50} and measuring the prevalence of filter bubbles⁹. Unlike these studies, however, we focus on how to promote exposure to trustworthy information rather than seeking to quantify or reduce different sources of bias.

A number of limitations must be acknowledged. First, our current methodology, which is based on reliability ratings compiled at the level of individual sources, does not allow us to evaluate the quality of specific articles that participants saw. However, even a coarse signal about source quality could still be useful for ranking a newsfeed given that information about reliability is more widely available at the publisher level than the article level. Another limitation is that our data lack information about actual engagement.

Though we show that our re-ranking procedure is associated with a minimal loss in predictive accuracy, it remains an open question whether diversity-based rankings lead not just to higher exposure to trustworthy content but also to more engagement with it. Our analysis seems to suggest a trade-off between ranking accuracy and trustworthiness, but the results are specific to one algorithm (user-based collaborative filtering). Different ranking schemes might make better use of the diversity signal. In general, more research is needed to tease apart the causal link between political attitudes, readership, engagement and information quality.

Our work has a number of implications for the integrity of the online information ecosystem. First, our findings suggest that

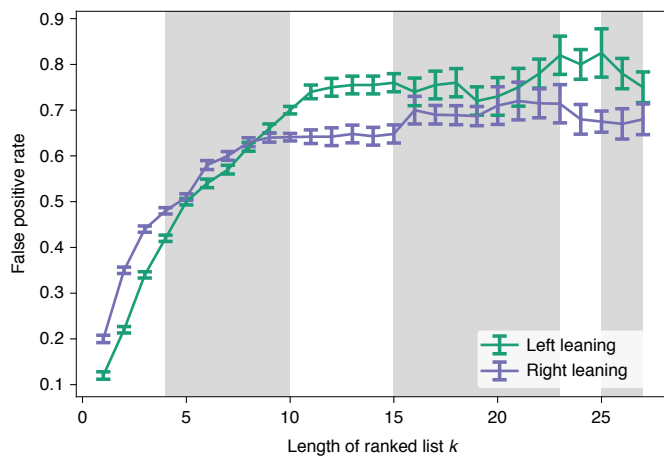


Fig. 7 | Probability that a trustworthy domain (NewsGuard score ≥ 60) is not recommended by CF+D but is recommended by CF for left- and right-leaning domains as a function of list length k ($N_k = 28$ list lengths). Each point is the average over a sample of users. Error bars represent the s.e.m. The shaded regions represent the values of k for which the difference is not statistically significant at standard levels ($\alpha = 0.05$, Welch's t tests with Bonferroni correction for $n = 28$; all tests are two-sided, see Supplementary Table 7 for a full summary).

search engines and social media platforms should consider including audience diversity into their existing set of news quality signals. Such a change could be especially valuable for domains for which we lack other quality signals such as source reliability ratings compiled by experts. Media ratings systems such as NewsGuard could also benefit from adopting our diversity metric, for example, to help screen and prioritize domains for manual evaluation. Likewise, designers of recommendation algorithms should consider measuring the trustworthiness of rankings as an additional measure of performance of their systems.

Critics may raise concerns that such a change in ranking criteria would result in unfair outcomes, for example, by reducing exposure to content by certain partisan groups but not others. To see whether ranking by diversity leads to any differential treatment for different partisan news sources, we compute the rate of false positives due to re-ranking by diversity. Here, the false positive rate is defined as the conditional probability that CF+D does not rank a trustworthy domain amongst the top k recommendations while CF does, for both left- and right-leaning domains. To determine whether a domain is trustworthy, we rely on the classification provided by NewsGuard (that is, the domain has a reliability score ≥ 60). Figure 7 shows the rate of false positives as a function of k of both left- and right-leaning domains averaged over all users. Despite some small differences, especially for low values of k , we find no consistent evidence that this change would produce systematically differential treatment across partisan groups.

Another concern is the possibility of abuse. For example, an attacker could employ a number of automated accounts to collectively engage with an ideologically diverse set of sources. This inauthentic, ideologically diverse audience could then be used to push specific content that the attacker wants to promote atop the rankings of a recommender system. Similarly, an attacker who wanted to demote a particular content could craft an inauthentic audience with low diversity. Fortunately, there is a vast literature on the topic of how to defend recommender systems against such 'shilling' attacks^{31,52} and platforms already collect a wealth of signals to detect and remove inauthentic coordinated behaviour of this kind. Future work should investigate the feasibility of creating trusted social media audiences that are modelled on existing efforts in

marketing research using panels of consumers. We hope that our result stimulates further research in this area.

Methods

This study complies with all relevant ethical regulations and was reviewed by the institutional review board under protocols #HUM00161944 (University of Michigan) and #STUDY000433 (University of South Florida).

Data. Our analysis combines two sources of data. The first is the NewsGuard News Website Reliability Index⁴³, a list of web domain reliability ratings compiled by a team of professional journalists and news editors. The data that we licensed for research purposes include scores for 3,765 web domains on a 100-point scale based on a number of journalistic criteria such as editorial responsibility, accountability and financial transparency. These data were current as of 12 November 2019 and do not reflect subsequent updates (see Data availability section for more information). NewsGuard categorizes web domains into four main groups: 'green' domains, which have a score of 60 or more points and are considered reliable; 'red' domains, which score less than 60 points and are considered unreliable; 'satire' domains, which should not be regarded as news sources regardless of their score; and 'platform' domains such as Facebook or YouTube that primarily host content generated by users. The mean reliability score for domains in the data is 69.6. The distribution of scores is shown in Supplementary Fig. 1.

The second data source is the YouGov Pulse panel, a sample of US-based internet users whose web traffic was collected in anonymized form with their prior consent. This traffic data were collected during seven periods between October 2016 and March 2019 (Supplementary Table 6). A total of 6,890 participants provided data. Overall, this group is diverse and resembles the US population on key demographic and political dimensions (47.9% male, 29.0% with a four-year college degree, 67.9% white, median age 55 years, 37.8% identifying as Democrats and 26.3% identifying as Republicans; see Supplementary Table 6 for a full summary by sample collection period). Note that, to be eligible for the study, participants in the YouGov Pulse panel had to be 18+ years of age, so the reported dimensions should be interpreted as being conditional on this extra eligibility criterion.

We perform a number of pre-processing steps on this data. We combine all waves into a single sample. We pool web traffic for each domain that received 30 or more unique visitors. Finally, we use the self-reported partisanship of the visitors (on a seven-point scale from an online survey) to estimate mean audience partisanship and audience partisan diversity, which we estimate using different measures described next. These different measures are compared in Supplementary Table 1.

Definition of audience partisan diversity. To measure audience partisan diversity, first define N_j as the count of participants who visited a web domain and reported their political affiliation to be equal to j for $j = 1, \dots, 7$ (where 1 indicates strong Democrat and 7 indicates strong Republican). The total number of participants who visited the domain is thus $N = \sum_j N_j$, and the fraction of participants with a partisanship value of j is $p_j = N_j/N$. Denote the partisanship of the i th individual as s_i . We calculate the following metrics to measure audience partisan diversity:

Variance: $\sigma^2 = N^{-1} \sum (s_i - \bar{s})^2$, where \bar{s} is average partisanship.

Shannon's entropy: $S = - \sum p(j) \log p(j)$, where $p(j)$ is estimated in the following three different ways: (1) $p(j) = p_j$ (maximum likelihood), (2) $p(j) = \frac{N_j + \alpha}{N + 7\alpha}$ (the mean of the posterior distribution of the Dirichlet prior with $\alpha = 1$) and (3) the Nemenman-Shafee-Bialek (NSB) method in ref. ⁵³, which uses a mixture of Dirichlet priors.

Complementary maximum probability: $1 - \max_j \{p_j\}$;

Complementary Gini: $1 - G$, where G is the Gini coefficient of the count distribution $\{N_j\}_{j=1, \dots, 7}$.

The above metrics all capture the idea that the partisan diversity of the audience of a web domain should be reflected in the distribution of its traffic across different partisan groups. Each weights the contribution of each individual person who visits the domain equally. They can thus be regarded as user-level measures of audience partisan diversity. However, the volume and content of web browsing activity are highly heterogeneous across internet users^{8,54}, with different users recording different numbers of pageviews to the same website. To account for this imbalance, we also compute the pageview-level, weighted variants of the above audience partisan diversity metrics where, instead of treating all visitors equally, each individual visitor is weighted by the number of pageviews they made to any given domain.

As a robustness check, we compare the strength of association of each of these metrics to news reliability in Supplementary Table 1. We find that all variants correlate with news reliability, but the relationship is strongest for variance.

Audience diversity and collaborative filtering. In general, a recommendation algorithm takes a set of users \mathcal{U} and a set of items \mathcal{D} and learns a function $f: \mathcal{U} \times \mathcal{D} \rightarrow \mathbb{R}$ that assigns a real value to each user-item pair (u, d) representing the interest of user u in item d . This value denotes the estimated rating that user u will give to item d . In the context of the present study, \mathcal{D} is a set of news sources

identified by their web domains (for example, nytimes.com, wsj.com), so from now on we will refer to $d \in \mathcal{D}$ interchangeably as either a web domain or a generic item.

Collaborative filtering is a classic recommendation algorithm in which some ratings are provided as input and unknown ratings are predicted based on those known input ratings. In particular, the user-based CF algorithm, which we employ here, seeks to provide the best recommendations for users by learning from others with similar preferences. CF therefore requires a user–domain matrix where each entry is either known or needs to be predicted by the algorithm. Once the ratings are predicted, the algorithm creates a ranked list of domains for each user that are sorted in descending order by their predicted ratings.

To test the standard CF algorithm and our modified CF+D algorithm, we first construct a user–domain matrix V from the YouGov Pulse panel. The YouGov Pulse dataset does not provide user ratings of domains, so we instead count the number of times $\pi_{u,d} \in \mathbb{Z}^+$ a user u has visited a domain d (that is, pageviews) and use this variable as a proxy³⁷. Because this quantity is known to follow a very skewed distribution, we compute the rating as the TF-IDF of the pageview counts:

$$v_{u,d} = \frac{\pi_{u,d}}{\sum_h \pi_{u,h}} \log \left(\frac{\pi}{\sum_u \pi_{u,d}} \right), \quad (2)$$

where $\pi = \sum_u \sum_d \pi_{u,d}$ is the total number of visits. Note that, if a user has never visited a particular domain, then $v_{u,d} = 0$. Therefore, if we arrange all the ratings into a user–domain matrix $V \in \mathbb{R}^{|\mathcal{U}| \times |\mathcal{D}|}$, such that $(V)_{u,d} = v_{u,d}$, we will obtain a sparse matrix. The goal of any recommendation task is to complete the user–domain matrix by predicting the missing ratings, which in turn allows us to recommend new web domains to users that may not have seen them. In this case, however, we lack data on completely unseen domains. To test the validity of our methods, we therefore follow the customary practice in machine learning of setting aside some data to be used purely for testing (Supervised learning evaluation workflow section).

Having defined V , the next step of the algorithm is to estimate the similarity between each pair of users. To do so, we use either the Pearson correlation coefficient or the Kendall rank correlation of their user vectors, that is, their corresponding row vectors in V (that is, with zeroes included). For example, if $\tau(\cdot, \cdot) \in [-1, 1]$ denotes the Kendall rank correlation coefficient between two sets of observations, then the corresponding coefficient of similarity between $u \in \mathcal{U}$ and $u' \in \mathcal{U}$ can be defined as

$$\text{sim}(u, u') = \frac{\tau(V_u, V_{u'}) + 1}{2}, \quad (3)$$

where $V_u, V_{u'} \in \mathbb{R}^{1 \times |\mathcal{U}|}$ are the row vectors of u and u' , respectively. A similar definition can be used for Pearson's correlation coefficient in place of τ .

These similarity coefficients are in turn used to calculate the predicted ratings. In the standard user-based CF, the predicted rating of a user u for a domain d is calculated as

$$\hat{v}_{u,d}^{\text{CF}} = \bar{v}_u + \frac{\sum_{u' \in N_{u,d}} \text{sim}(u, u') (v_{u',d} - \bar{v}_{u'})}{\sum_{u' \in N_{u,d}} \text{sim}(u, u')}, \quad (4)$$

where $N_{u,d} \subseteq \mathcal{U}$ is the set of the $n=10$ users most similar to u who have also rated d (that is, the neighbours of u), $v_{u',d}$ is the observed rating (computed with equation (2)) that neighbouring user u' has given to domain d , \bar{v}_u and $\bar{v}_{u'}$ are the average ratings of u and u' across all domains they visited, respectively, and $\text{sim}(u, u')$ is the similarity coefficient (computed with equation (3)) between users u and u' based on either the Pearson or the Kendall correlation coefficient.

Having defined the standard CF in equation (4), we now define our variant CF+D, which incorporates audience partisan diversity of domain $d \in \mathcal{D}$ as a re-ranking signal in the following way:

$$\hat{v}_{u,d}^{\text{CF+D}} = \hat{v}_{u,d}^{\text{CF}} + g(\delta_d), \quad (5)$$

where $g(\delta_d)$ is the re-ranking term of domain d , obtained by plugging the audience partisan diversity δ_d (for example, we use the variance of the distribution of self-reported partisan slants of its visitors, $\delta_d = \sigma_d^2$) into a standard logistic function:

$$g(\delta) = \frac{a}{1 + \exp(-(\delta - t)/\psi)}. \quad (6)$$

In equation (6), parameters a , ψ and t generalize the upper asymptote, inverse growth rate and location of the standard logistic function, respectively. For the results reported in this study, we empirically estimate the location as $t = \bar{\delta}$, the average audience partisan diversity across all domains, which corresponds to the value of $\bar{\delta} = 4.25$ since we measure diversity as the variance of the distribution of self-reported partisan slants. For the remaining parameters, we choose $a = 1$ and $\psi = 1$. As a robustness check, we re-ran all analyses with a larger value of a and obtained qualitatively similar results (available upon reasonable request).

Supervised learning evaluation workflow. To evaluate both recommendation algorithms, we follow a standard supervised learning workflow. We use precision and RMSE, two standard metrics used to measure the relevance and accuracy of predicted ratings in supervised learning settings. We define these two metrics in Accuracy metrics section. Here, we instead describe the workflow we followed to evaluate the recommendation methods. Since our approach is based on supervision, we need to designate some of the user ratings (that is, the number of visits to each domain, which are computed using equation (2)) as ground truth to compute performance metrics.

For each user, we randomly split the domains they visited into a training set (70%) and a testing set (30%). This splitting varies by user, so the same domain could be included in the training set of a user and in the testing set of another. Then, given any two users, their training set ratings are used to compute user–user similarities using equation (3) (which is based on Kendall's rank correlation coefficient; a similar formula can be defined using Pearson's correlation). If, in computing user–user similarities with equation (3), a domain is present for one user but not the other, the latter rating is assumed to be zero regardless of whether the domain is present in testing or not. This assumption, which follows standard practice in collaborative filtering algorithm, ensures that there is no leaking of information between the test and training sets.

Finally, using either equation (4) or (5), we predict ratings for domains in the test set and compare them with the TF-IDF of the actual visit counts in the data.

Recommendation based on global popularity. We also generate ranked lists for users based on global domain popularity (user level) as an additional baseline recommendation technique. All the domains are initially assigned a rank (global popularity rank) according to their user-level popularity, which is calculated from the training set views. Then, the domains in the test set of each user are ranked according to their global popularity ranks to generate the recommendations. This method does not include any personalization as the rank of a domain for a particular user does not depend on other similar users but depends on the whole population. In particular, if two users share the same two domains in testing, their relative ranking is preserved, even if the two users visited different domains in training.

Trustworthiness metrics. In addition to standard metrics of accuracy (precision and RMSE; Accuracy metrics section), we define a new metric called 'trustworthiness' to measure the news reliability of the recommended domains. It is calculated using NewsGuard scores in two ways: either using the numerical scores or the set of binary indicators for whether a site meets or exceeds the threshold score of 60 defined by NewsGuard as indicating that a site is generally trustworthy⁴³. Let d_1, d_2, \dots, d_k be a ranked list of domains. Using numerical scores, the trustworthiness is the average:

$$\frac{1}{k} \sum_{r=1}^k Q(d_r), \quad (7)$$

where $Q(d) \in [0, 100]$ denotes the NewsGuard reliability score of $d \in \mathcal{D}$.

If instead we use the binary indicator of trustworthiness provided by NewsGuard, then the trustworthiness of domains in a list is defined as the fraction of domains that meet or exceed the threshold score. Note that, unlike precision and RMSE, the trustworthiness of a list of recommendations does not use information on the actual ratings $v_{u,d}$. Instead, using equation (7), we compute the trustworthiness of the domains in the test set ranked in decreasing order of user visits $v_{u,d}$. We then compare the trustworthiness of the rankings obtained with either CF or CF+D against the trustworthiness of this baseline.

Accuracy metrics. Given a user u , let us consider a set \mathcal{D} of web domains for which $|\mathcal{D}| = D$. For each domain $d \in \mathcal{D}$, we have three pieces of information: the two predicted ratings $\hat{v}_{u,d}^{\text{CF}}$ and $\hat{v}_{u,d}^{\text{CF+D}}$ produced by CF and CF+D and the actual rating $v_{u,d}$ (defined in Audience diversity and collaborative filtering section). In the following, we omit the subscript u of the user, which is fixed throughout, and the CF/CF+D superscript unless it is not obvious from context.

Let us consider a given recommendation method (CF or CF+D) and denote by $r(d)$ (respectively, $r'(d)$) the rank of d when the domains are sorted by decreasing order of recommendation and actual ratings, respectively. Given a recommendation list length $0 < k \leq D$, let us define the set of predicted domains as

$$P_k = \{d \in \mathcal{D} : r(d) \leq k\}$$

and the set of actual domains as

$$A_k = \{d \in \mathcal{D} : r'(d) \leq k\}.$$

Then the precision for a given value of k is given by the fraction of correctly predicted domains:

$$\text{Precision} = \frac{|P_k \cap A_k|}{|P_k|}.$$

Similarly, the RMSE for a given value of k between the two ranked lists of ratings is computed as

$$\text{RMSE} = \sqrt{\frac{1}{k} \sum_{r=1}^k (\hat{v}_{\rho(r)} - v_{\rho'(r)})^2},$$

where $\rho : [D] \mapsto \mathcal{D}$ (respectively ρ') is the inverse function of $r(\cdot)$ (respectively, $r'(\cdot)$), that is, the function that maps ranks back to their domain by the recommendation method (respectively, by actual visits). Note that, in the summation, $\rho(r)$ and $\rho'(r)$ do not generally refer to the same web domain: the averaging is over the two ranked lists of ratings, not over the set of domains in common between the two lists.

Discounting via ranking. To measure the effect of CF+D on the trustworthiness of rankings, we must select a particular list length k . Although Fig. 4 shows improvements for all values of k , one potential problem when stratifying on different groups of users is that the results could depend on the particular choice of k . To avoid dependence on k , we consider a probabilistic model of a hypothetical user visiting web domains from a ranked list of recommendations and define overall trustworthiness as the expected value of the trustworthiness of domains selected from that list (that is, discounted by probability of selection).

Let us consider a universe of domains \mathcal{D} as the set of items to rank. Inspired by prior approaches on stochastic processes based on ranking⁵⁵, we consider a discounting method that posits that the probability of selecting domain $d \in \mathcal{D}$ from a given ranked recommendation list decays as a power law of its rank in the list:

$$\Pr\{X = d\} = \frac{r_d^{-\alpha}}{\sum_h r_h^{-\alpha}} \quad (8)$$

where $X \in \mathcal{D}$ is a random variable denoting the probabilistic outcome of the selection from the ranked list, $r_d \in \mathbb{N}$ is the rank of a generic $d \in \mathcal{D}$ and $\alpha \geq 0$ is the exponent of power-law decay (when $\alpha = 0$, all domains are equally likely; when $\alpha > 0$, top-ranked domains are more likely to be selected).

This procedure allows us to compute, for any given user, the effect of a recommendation method (CF or CF+D) simply as the difference between its expected trustworthiness and the trustworthiness of the ranking obtained by sorting the domains visited by the user in decreasing order of pageviews (equation (1)).

In practice, to compute equation (1), let d_1, d_2, \dots, d_k and d'_1, d'_2, \dots, d'_k be two ranked lists of domains, $d_r, d'_r \in \mathcal{D} \forall r = 1, \dots, k$, generated by a recommendation algorithm and by actual user pageviews, respectively, and let us denote by $Q(d)$ the NewsGuard reliability score of $d \in \mathcal{D}$ (Trustworthiness metrics section). Recall that equation (8) specifies the probability of selecting a given domain $d \in \mathcal{D}$ from a particular ranked list as a function of its rank. Even though any pair of equally ranked domains will be different across these two lists (that is, $d_r \neq d'_r$ in general), their probability will be the same because equation (8) only depends on r . We can thus calculate the expected improvement in trustworthiness as

$$\Delta Q = \sum_{r=1}^k P(r) (Q(d_r) - Q(d'_r)), \quad (9)$$

where $P(r)$ is the probability of selecting a domain with rank r from equation (8), which we computed setting $\alpha = 1$.

Stratification analysis. Recall that we use the self-reported partisanship of respondents in the YouGov Pulse panel as the basis for our diversity signal (Definition of audience partisan diversity section). To avoid the circular reasoning in stratifying on the same source of data, Fig. 6a,c groups these users according to the slant of their actual news consumption, which may not necessarily reflect their self-reported partisanship (for example, a self-reported Democrat might access mostly conservative-leaning websites). We determined this latter metric using an external classification originally proposed by Bakshy et al.⁵, who estimated the slant of 500 web domains focused on hard news topics. In practice, Bakshy et al.⁵ based their classification on how hard news from those domains was shared on Facebook by users who self-identified as liberal or conservative in their profile. For almost all domains, Bakshy et al.⁵ reported a value $s \in [-1, 1]$ with a value of $s = +1$ for domains that are shared almost exclusively by conservatives and a value of $s = -1$ for those shared almost exclusively by liberals. (These values could technically vary over $[-2, 2]$ but only 1% of domains fell outside $[-1, 1]$ when using the measurement approach described by Bakshy et al.⁵)

In Fig. 6c, respondents are grouped according to the absolute slant $|s|$ of the visited domains, where a value of $|s| = 0$ denotes domains with a perfectly centrist slant and a value of $|s| = 1$ indicates domains with extreme liberal or conservative slants (that is, they are almost exclusively shared by one group and not the other).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Data necessary to reproduce the findings in the main manuscript text and in the Supplementary Information are available, in aggregated and anonymized format, at <https://github.com/gliampaglia/InfoDiversity/>. The raw data that support the findings of this study are available from NewsGuard Technology, Inc., but restrictions apply to the availability of these data, which were used under licence for the current study and thus cannot be made publicly available. However, data are available from the authors upon reasonable request subject to licensing from NewsGuard. The data used in this study were current as of 12 November 2019 and do not reflect NewsGuard's regular updates of the data.

Code availability

Code necessary to reproduce the findings in the main text and in the Supplementary Information are available at <https://github.com/gliampaglia/InfoDiversity/>.

Received: 15 July 2020; Accepted: 9 December 2021;

Published online: 3 February 2022

References

- Lazer, D. The science of fake news. *Science* **359**, 1094–1096 (2018).
- Vosoughi, S., Roy, D. & Aral, S. The spread of true and false news online. *Science* **359**, 1146–1151 (2018).
- Hui, P.-M., Yang, K.-C., Torres-Lugo, C. & Menczer, F. BotSlayer: DIY real-time influence campaign detection. In *Proc. 14th International AAAI Conference on Web and Social Media (ICWSM '20)* 980–982 (AAAI, 2020).
- Hart, W. Feeling validated versus being correct: a meta-analysis of selective exposure to information. *Psychol. Bull.* **135**, 555 (2009).
- Bakshy, E., Messing, S. & Adamic, L. Exposure to ideologically diverse news and opinion on Facebook. *Science* **348**, 1130–1132 (2015).
- Chen, W., Pacheco, D., Yang, K.-C. & Menczer, F. Neutral bots reveal political bias on social media. *Nat. Commun.* **12**, 5580 (2021).
- Guess, A., Lyons, B., Nyhan, B., & Reifler, J. *Avoiding the Echo Chamber about Echo Chambers: Why Selective Exposure to Like-Minded Political News Is Less Prevalent than You Think*. Technical Report (Knight Foundation, 2018).
- Guess, A. M. (Almost) everything in moderation: new evidence on Americans' online media diets, unpublished manuscript (2018).
- Flaxman, S., Goel, S. & Rao, J. M. Filter bubbles, echo chambers, and online news consumption. *Public Opin. Q.* **80**, 298–320 (2016).
- Allcott, H. & Gentzkow, M. Social media and fake news in the 2016 election. *J. Econ. Perspect.* **31**, 211–236 (2017).
- Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B. & Lazer, D. Fake news on Twitter during the 2016 U.S. presidential election. *Science* **363**, 374–378 (2019).
- Guess, A., Nagler, J. & Tucker, J. Less than you think: prevalence and predictors of fake news dissemination on Facebook. *Sci. Adv.* **5**, eaau4586 (2019).
- Allen, J., Howland, B., Mobius, M., Rothschild, D. & Watts, D. J. Evaluating the fake news problem at the scale of the information ecosystem. *Sci. Adv.* **6**, eaay3539 (2020).
- Guess, A. M., Nyhan, B. & Reifler, J. Exposure to untrustworthy websites in the 2016 US election. *Nat. Hum. Behav.* **4**, 472–480 (2020).
- Goel, S., Broder, A., Gabrilovich, E. & Pang, B. Anatomy of the long tail: ordinary people with extraordinary taste. In *Proc. 3rd ACM International Conference on Web Search and Data Mining (WSDM '10)* 201–210 (Association for Computing Machinery, 2010).
- Nikolov, D., Lalmas, M., Flammini, A. & Menczer, F. Quantifying biases in online information exposure. *J. Assoc. Inf. Sci. Technol.* **70**, 218–229 (2019).
- Shmargad, Y. & Klar, S. Sorting the news: how ranking by popularity polarizes our politics. *Polit. Commun.* **37**, 423–446 (2020).
- Salganik, M. J., Dodds, P. S. & Watts, D. J. Experimental study of inequality and unpredictability in an artificial cultural market. *Science* **311**, 854–856 (2006).
- Hogg, T. & Lerman, K. Disentangling the effects of social signals. *Hum. Comput.* **2**, 189–208 (2015).
- Ciampaglia, G. L., Nematzadeh, A., Menczer, F. & Flammini, A. How algorithmic popularity bias hinders or promotes quality. *Sci. Rep.* **8**, 15951– (2018).
- Germano, F., Gómez, V., & Le Mens, G. The few-get-richer: a surprising consequence of popularity-based rankings? In *The World Wide Web Conference (WWW '19)* 2764–2770 (ACM, 2019).
- Macy, M., Deri, S., Ruch, A. & Tong, N. Opinion cascades and the unpredictability of partisan polarization. *Sci. Adv.* **5**, eaax0754 (2019).
- Avram, M., Micallef, N., Patil, S. & Menczer, F. Exposure to social engagement metrics increases vulnerability to misinformation. *Harvard Kennedy School Misinform. Rev.* <https://doi.org/10.37016/mr-2020-033> (2020).
- Shao, C. The spread of low-credibility content by social bots. *Nat. Commun.* **9**, 4787 (2018).

25. Stella, M., Ferrara, E. & De Domenico, M. Bots increase exposure to negative and inflammatory content in online social systems. *Proc. Natl Acad. Sci. USA* **115**, 12435–12440 (2018).
26. Brown, C. & Levin, J. Prioritizing original news reporting on Facebook. *Internet Archive* <https://web.archive.org/web/20210126011953/https://about.fb.com/news/2020/06/prioritizing-original-news-reporting-on-facebook/> (2020).
27. Google. Surfacing useful and relevant content – how news works. *Internet Archive* <https://web.archive.org/web/20201017172355/https://newsinitiative.withgoogle.com/hownewsworks/approach/surfacing-useful-and-relevant-content/> (2020).
28. Jiang, S., Baumgartner, S., Ittycheriah, A. & Yu, C. *Factoring Fact-Checks: Structured Information Extraction from Fact-Checking Articles* (Association for Computing Machinery, 2020).
29. Rashkin, H., Choi, E., Jang, J. Y., Volkova, S. & Choi, Y. Truth of varying shades: analyzing language in fake news and political fact-checking. In *Proc. 2017 Conference on Empirical Methods in Natural Language Processing* 2931–2937 (Association for Computational Linguistics, 2017).
30. Jiang, S. & Wilson, C. Linguistic signals under misinformation and fact-checking: evidence from user comments on social media. In *Proc. ACM on Human-Computer Interaction* <https://doi.org/10.1145/3274351> (2018).
31. Gupta, A., Kumaraguru, P., Castillo, C. & Meier, P. *TweetCred: Real-Time Credibility Assessment of Content on Twitter* 228–243 (Springer International, 2014).
32. Cho, J.-H., Chan, K. & Adali, S. A survey on trust modeling. *ACM Comput. Surv.* **48**, 1–40 (2015).
33. Golbeck, J. A. *Computing and Applying Trust in Web-Based Social Networks*. PhD thesis, Univ. Maryland at College Park (2005).
34. Adler, B. T. & de Alfaro, L. A content-driven reputation system for the Wikipedia. In *Proc. 16th International Conference on World Wide Web (WWW '07)* 261–270 (ACM, 2007).
35. Zhang, A. X. et al. A structured response to misinformation: defining and annotating credibility indicators in news articles. In *Companion Proc. Web Conference 2018 (WWW '18)* 603–612 (International World Wide Web Conference Steering Committee, 2018).
36. Pennycook, G. & Rand, D. G. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proc. Natl Acad. Sci. USA* **116**, 2521–2526 (2019).
37. Hong, L. & Page, S. E. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proc. Natl Acad. Sci. USA* **101**, 16385–16389 (2004).
38. Shi, F., Teplitskiy, M., Duede, E. & Evans, J. A. The wisdom of polarized crowds. *Nat. Hum. Behav.* **3**, 329–336 (2019).
39. Varol, O., Ferrara, E., Davis, C., Menczer, F. & Flammini, A. Online human-bot interactions: detection, estimation, and characterization. In *Proc. 11th International AAAI Conference on Web and Social Media (ICWSM '17)* 280–289 (AAAI, 2017).
40. Yang, K.-C. Arming the public with artificial intelligence to counter social bots. *Hum. Behav. Emerg. Technol.* **1**, 48–61 (2019).
41. Yang, K.-C., Varol, O., Hui, P.-M. & Menczer, F. Scalable and generalizable social bot detection through data selection. In *Proc. 34th AAAI Conference on Artificial Intelligence* 1096–1103 (AAAI, 2020).
42. Sayyadiharikandeh, M., Varol, O., Yang, K.-C., Flammini, A. & Menczer, F. Detection of novel social bots by ensembles of specialized classifiers. In *Proc. 29th ACM International Conference on Information and Knowledge Management* 2725–2732 (ACM, 2020).
43. NewsGuard, Inc. Rating process and criteria. *Internet Archive* <https://web.archive.org/web/20200630151704/https://www.newsguardtech.com/ratings/rating-process-criteria/> (2020).
44. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P. & Riedl, J. GroupLens: an open architecture for collaborative filtering of netnews. In *Proc. 1994 ACM Conference on Computer Supported Cooperative Work (CSCW '94)* 175–186 (Association for Computing Machinery, 1994).
45. Konstan, J. A. GroupLens: applying collaborative filtering to usenet news. *Commun. ACM* **40**, 77–87 (1997).
46. Ziegler, C.-N., McNee, S. M., Konstan, J. A. & Lausen, G. Improving recommendation lists through topic diversification. In *Proc. 14th International Conference on World Wide Web (WWW '05)* 22–32 (ACM, 2005).
47. Joachims, T., Granka, L., Pan, B., Hembrooke, H. & Gay, G. Accurately interpreting clickthrough data as implicit feedback. *SIGIR Forum* **51**, 4–11 (2017).
48. Ribeiro, F. et al. Media bias monitor: quantifying biases of social media news outlets at large-scale. In *Proc. International AAAI Conference on Web and Social Media* 290–299 (AAAI, 2018).
49. Munson, S. A. & Resnick, P. *Presenting Diverse Political Opinions: How and How Much* 1457–1466 (Association for Computing Machinery, 2010).
50. Munson, S., Lee, S., & Resnick, P. Encouraging reading of diverse political viewpoints with a browser widget. In *Proc. International AAAI Conference on Web and Social Media (ICWSM '13)* 419–428 (AAAI, 2013).
51. Lam, S. K. & Riedl, J. Shilling recommender systems for fun and profit. In *Proc. 13th International Conference on World Wide Web (WWW '04)* 393–402 (ACM, 2004).
52. Gunes, I., Kaleli, C., Bilge, A. & Polat, H. Shilling attacks against recommender systems: a comprehensive survey. *Artif. Intell. Rev.* **42**, 767–799 (2014).
53. Nemenman, I., Shafee, F. & Bialek, W. Entropy and inference, revisited. In *Proc. 14th International Conference on Neural Information Processing Systems: Natural and Synthetic (NIPS'01)* 471–478 (MIT Press, 2001).
54. Montgomery, A. L. & Faloutsos, C. Identifying web browsing trends and patterns. *Computer* **34**, 94–95 (2001).
55. Fortunato, S., Flammini, A. & Menczer, F. Scale-free network growth by ranking. *Phys. Rev. Lett.* **96**, 218701 (2006).

Acknowledgements

We thank NewsGuard for licensing the data and A. Guess and J. Reifler, B.N.'s coauthors on the research project that generated the web traffic data used in this study. We also thank the organizers, chairs and participants of the News Quality in the Platform Era workshop (organized by the Social Science Research Council), especially R. Lawrence, P.M. Napoli, K. Munger, J. Dunaway, C. Moon Sehat and J. Shin, for their helpful comments. This work was supported in part by the National Science Foundation under a collaborative award (NSF grant no. 1915833 to G.L.C. and 1949077 to B.N.). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

All authors designed the research. S.B. and S.Y. performed data analysis. All authors wrote, reviewed and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41562-021-01276-5>.

Correspondence and requests for materials should be addressed to Giovanni Luca Ciampaglia.

Peer review information *Nature Human Behaviour* thanks Paul Resnick and Katherine Ognyanova for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2022

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☒ ☐ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☐ ☒ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☐ ☒ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection Data used were already existing; no data collection was performed in this study.

Data analysis Python / R scripts are available online at <https://github.com/glciampaglia/InfoDiversity/>

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data necessary to reproduce the findings in the manuscript are available, in aggregated and anonymized format, at <https://github.com/glciampaglia/InfoDiversity/>

The raw data that support the findings of this study are available from NewsGuard Technology, Inc. but restrictions apply to the availability of these data, which were used under license for the current study and thus cannot be made publicly available. However, data are available from the authors upon reasonable request subject to licensing from NewsGuard. The data used in this study were current as of November 12, 2019 and do not reflect NewsGuard's regular updates of the data.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	<i>Describe how sample size was determined, detailing any statistical methods used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.</i>
Data exclusions	<i>Describe any data exclusions. If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.</i>
Replication	<i>Describe the measures taken to verify the reproducibility of the experimental findings. If all attempts at replication were successful, confirm this OR if there are any findings that were not replicated or cannot be reproduced, note this and describe why.</i>
Randomization	<i>Describe how samples/organisms/participants were allocated into experimental groups. If allocation was not random, describe how covariates were controlled OR if this is not relevant to your study, explain why.</i>
Blinding	<i>Describe whether the investigators were blinded to group allocation during data collection and/or analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.</i>

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This is an observational study of news consumption patterns based on quantitative data of internet browsing behavior, combined with surveys of political information. These data are combined with a high-coverage set of source reliability ratings US-based news sources.
Research sample	The sample used was drawn from the YouGov Pulse panel, a representative sample of Americans, collected along with their internet browsing history, in multiple waves between 2016 and 2019. Informed consent to participate in the study was obtained prior to data collection.
Sampling strategy	Multiple waves were collected with random sampling, with sample size of each wave reported in Table 1. Sample size calculations were not necessary since estimated quantities were not sample properties but properties of the internet traffic produced by participants, like total popularity of each web domain, etc.
Data collection	Data are collected via desktop / mobile tracking apps.
Timing	The data were collected in multiple waves between 2016 and 2019. Table 1 reports duration of data collection for each wave.
Data exclusions	We excluded all traffic to sources that received less than 30 visits across all waves combined.
Non-participation	We do not observe data from survey respondents who declined to make Internet browsing data available. Respondents may voluntarily disable browsing data collection at any time or withdraw from the YouGov Pulse panel entirely.
Randomization	Participants were not allocated to experimental groups.

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	<i>Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.</i>
Research sample	<i>Describe the research sample (e.g. a group of tagged <i>Passer domesticus</i>, all <i>Stenocereus thurberi</i> within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.</i>

Sampling strategy	Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.
Data collection	Describe the data collection procedure, including who recorded the data and how.
Timing and spatial scale	Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken
Data exclusions	If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.
Reproducibility	Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.
Randomization	Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.
Blinding	Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.
Did the study involve field work?	<input type="checkbox"/> Yes <input type="checkbox"/> No

Field work, collection and transport

Field conditions	Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).
Location	State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).
Access & import/export	Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).
Disturbance	Describe any disturbance caused by the study and how it was minimized.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Antibodies

Antibodies used	Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.
Validation	Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	State the source of each cell line used.
Authentication	Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.

Mycoplasma contamination

Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.

Commonly misidentified lines
(See [ICLAC](#) register)

Name any commonly misidentified cell lines used in the study and provide a rationale for their use.

Palaeontology and Archaeology

Specimen provenance

Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information).

Specimen deposition

Indicate where the specimens have been deposited to permit free access by other researchers.

Dating methods

If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight

Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals

For laboratory animals, report species, strain, sex and age OR state that the study did not involve laboratory animals.

Wild animals

Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.

Field-collected samples

For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.

Ethics oversight

Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics

See above.

Recruitment

Participants were recruited among respondents of the YouGov Pulse panel, a representative sample of U.S.-based Internet users. Informed consent to participate in the study was obtained prior to data collection.

Ethics oversight

This study was reviewed by the IRB under protocols #HUM00161944 (University of Michigan) and #STUDY000433 (University of South Florida).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about [clinical studies](#)

All manuscripts should comply with the ICMJE [guidelines for publication of clinical research](#) and a completed [CONSORT checklist](#) must be included with all submissions.

Clinical trial registration

Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.

Study protocol

Note where the full trial protocol can be accessed OR if not available, explain why.

Data collection

Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.

Outcomes

Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.

Dual use research of concern

Policy information about [dual use research of concern](#)

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

- | No | Yes |
|--------------------------|---|
| <input type="checkbox"/> | <input type="checkbox"/> Public health |
| <input type="checkbox"/> | <input type="checkbox"/> National security |
| <input type="checkbox"/> | <input type="checkbox"/> Crops and/or livestock |
| <input type="checkbox"/> | <input type="checkbox"/> Ecosystems |
| <input type="checkbox"/> | <input type="checkbox"/> Any other significant area |

Experiments of concern

Does the work involve any of these experiments of concern:

- | No | Yes |
|-------------------------------------|--|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Demonstrate how to render a vaccine ineffective |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Confer resistance to therapeutically useful antibiotics or antiviral agents |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Enhance the virulence of a pathogen or render a nonpathogen virulent |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Increase transmissibility of a pathogen |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Alter the host range of a pathogen |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Enable evasion of diagnostic/detection modalities |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Enable the weaponization of a biological agent or toxin |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Any other potentially harmful combination of experiments and agents |

ChIP-seq

Data deposition

- ☐ Confirm that both raw and final processed data have been deposited in a public database such as [GEO](#).
- ☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links

May remain private before publication.

For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.

Files in database submission

Provide a list of all files available in the database submission.

Genome browser session

(e.g. [UCSC](#))

Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.

Methodology

Replicates

Describe the experimental replicates, specifying number, type and replicate agreement.

Sequencing depth

Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.

Antibodies

Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.

Peak calling parameters

Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.

Data quality

Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.

Software

Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.

Flow Cytometry

Plots

Confirm that:

- ☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☐ All plots are contour plots with outliers or pseudocolor plots.
- ☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

- Sample preparation *Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*
- Instrument *Identify the instrument used for data collection, specifying make and model number.*
- Software *Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*
- Cell population abundance *Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*
- Gating strategy *Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*
- ☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

- Design type *Indicate task or resting state; event-related or block design.*
- Design specifications *Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*
- Behavioral performance measures *State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

- Imaging type(s) *Specify: functional, structural, diffusion, perfusion.*
- Field strength *Specify in Tesla*
- Sequence & imaging parameters *Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*
- Area of acquisition *State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*
- Diffusion MRI ☐ Used ☐ Not used

Preprocessing

- Preprocessing software *Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*
- Normalization *If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*
- Normalization template *Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*
- Noise and artifact removal *Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*

Volume censoring

Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.

Statistical modeling & inference

Model type and settings

Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).

Effect(s) tested

Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference
(See [Eklund et al. 2016](#))

Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.

Correction

Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).

Models & analysis

n/a | Involved in the study

- ☐ ☐ Functional and/or effective connectivity
- ☐ ☐ Graph analysis
- ☐ ☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).

Graph analysis

Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).

Multivariate modeling and predictive analysis

Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.