Political audience diversity and news reliability in algorithmic ranking

Saumya Bhadani,[1] Shun Yamaya,[2] Alessandro Flammini,[3] Filippo Menczer,[3] Giovanni Luca Ciampaglia,[1,*] and Brendan Nyhan[4]

[1] *Department of Computer Science and Engineering, University of South Florida, Tampa, FL, USA*

[2] *Department of Political Science, Stanford University, Stanford, CA, USA*

[3] *Observatory on Social Media, Indiana University, Bloomington, IN, USA*

[4] *Department of Government, Dartmouth College, Hanover, NH, USA*

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 U.S. citizens, 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.

* Corresponding author. ORCID: https://orcid.org/ 0000-0001-5354-9257

Concerns continue to grow about the prevalence of misinformation on social media platforms [31, 50], including during the recent COVID-19 pandemic [51]. These types of content often exploit people's tendency to prefer pro-attitudinal information [23], 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 toward selective exposure [17]. As a result, the news audience is far more polarized than the public as a whole [10, 19]. Although the prevalence of so-called "fake news" online is rather limited and concentrated among relatively narrow audiences [2, 3, 16–18, 20], content that generally appeals to these tendencies — which does include low-quality or false news — may generate high levels of readership or engagement [50], 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 [13, 38]. This bias may have several effects on the consumption of low-quality and false news. First, sorting the news by engagement (either predicted or achieved) can exacerbate polarization by increasing in-group bias and discouraging consumption among out-group members [47]. 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 [8, 12, 24, 32, 43]. 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 misinformation [4]. 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 [45, 48].

Given the speed and scale of social media, assessing directly the quality of every piece of content or the behavior of each user is infeasible. Online platforms are instead seeking to include signals about news quality in their content recommendation algorithms [9, 15], for example by extracting information from trusted publishers [26] or by means of linguistic patterns analysis [27, 40]. More generally, a vast literature examines how to assess the credibility of online sources [7, 22] and the reputations of individual online users [1, 14], 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 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 directly [54], but scaling such an approach would likely be costly and cause lags in the evaluation of novel content. Similarly, while crowdsourced website evaluations have been shown to be generally reliable in distinguishing between high and low quality news sources [39], 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 [25, 46], 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 [44, 49, 52, 53]. These two conditions (inferring partisanship reliably and preventing abuse by automated amplification/deception) could easily be met by the major social media 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 Americans, collected along with surveys of self-reported partisan information from respondents in the YouGov Pulse survey panel, and a data set of 3,765 news source reliability scores compiled by trained experts in journalism and provided by NewsGuard [37]. 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 the YouGov Pulse traffic data. Due to skew in audience size among 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) number of users, and as the (log of) number of visits, or pageviews. News reliability is instead measured using NewsGuard scores (see Methods A). Figure 1 shows that the popularity of a news source is at best weakly associated with its reliability (a full regression summary can be found in Supplementary Table 2). At the user level (left pane), the overall Pearson correlation is $r(n = 1024) = 0.03$ (two-sided $p = 0.36$, 95% c.i. $= [-0.03, 0.09]$). At the pageview level (right pane), $r(n = 1024) = 0.05$ (two-sided $p = 0.12$, 95% c.i. $= [-0.01, 0.11]$). 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% c.i. $= [0.02, 0.16]$), 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% c.i. $= [-0.25, 0.005]$). A similar pattern holds at the pageview level: a weak, positive and insignificant association for websites with predominantly Democratic audiences ($r(n = 702) = 0.07$, two-sided $p = 0.07$, 95% c.i. $= [-0.01, 0.14]$) 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% c.i.

$= [-0.10, 0.12]$). 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 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 (see Methods B) 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 Figure 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 Figure 2 also suggests 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 $r(n = 1024) = 0.38$, two-sided $p < 10^{-4}$, 95% c.i. $= [0.32, 0.43]$) or by how often they visited a given site (pageview level, right panel: partial correlation $r(n = 1024) = 0.22$, two-sided $p < 10^{-4}$, 95% c.i. $= [0.16, 0.28]$).

Aside from mean partisanship, a related, but potentially distinct, confounding factor is the extremity of the partisanship slants distribution (i.e., 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: $r(n = 1024) = 0.26, p < 10^{-4}$, 95% c.i. $= [0.20, 0.31]$; pageview level: $r(n = 1024) = 0.15, p < 10^{-4}$, 95% c.i. $= [0.08, 0.21]$; 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 the Supplementary Materials (Sec. 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 Figure 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, Figure 3 shows that audience partisan diversity is positively associated with news reliability (full regression summary can be found in Supplementary Table 3). 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 OLS 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 OLS coefficient of Democrat domains: $\beta = 2.93\,(0.66)$ at user level; $\beta = 0.82\,(0.86)$ at pageview level), which is consistent with Figure 2 (the partisan slope difference is 7.71 at user level, $p < 10^{-4}$, 95% c.i. $= [5.46, 9.97]$; 7.97 at pageview level, $p < 10^{-4}$, 95% c.i. $= [5.32, 10.62]$).

These results are not affected by popularity. Partisan diversity is weakly correlated with popularity, regardless of the operational definition of either measure (see Supplementary Table 4). In fact, the association between diversity and Newsguard reliability scores is consistent even when controlling for popularity (user level: $r(n = 1024) = 0.34$, two-sided $p < 10^{-4}$, 95% c.i. $= [0.29, 0.40]$; pageview level: $r(n = 1024) = 0.17$, two-sided $p < 10^{-4}$, 95% c.i. $= [0.11, 0.23]$), 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 Figure 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 Table 2 and Supplementary Table 3.

As mentioned before, variance in audience partisanship is not the only possible way to define audience partisan diversity; alternative definitions can be used (e.g., entropy; see Methods B). 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 (see 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 filtering (CF) algorithm [29, 41] change if we include audience partisan diversity as an additional signal. We call this modified version of the algorithm CF+D, which stands for Collaborative Filtering + Diversity (see Methods C for 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 TF-IDF; see also Methods C) 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 (see Methods D). 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 (i.e., 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 (i.e., as a forecasting exercise; see Supplementary Figures 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 interests [13], we cannot infer that the user has a negative preference toward 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 (i.e., visits in the test set), though in practice actual newsfeed algorithms rank items from a broader set of inputs, which typically includes content the user may not have seen (for example, content shared by friends [5]).

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 neighborhood of most similar users (domains with no neighborhood rating are discarded since neither CF nor CF+D can make a prediction for them; see Methods C) and for which we have a NewsGuard score (i.e., 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" [37] (see Methods F).

By varying the number of top domains k, we can evaluate how trustworthiness changes as the length of the list of recommendations increases. In Figure 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 Figure 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 (i.e., 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 (see Methods E).

We observe in Figure 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 behavior. In particular, CF produces less trustworthy rankings than both the recommendations based on global popularity and 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 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 literature [55]. If so, a loss of predictive ability would be expected.

Figure 5 compares the accuracy of CF+D in predicting user visits to domain in the test set with that of CF. To evaluate accuracy, we compute both the fraction of correctly predicted domains (precision) and root mean squared error (RMSE) as a function of the number of recommended domains k (see Methods G 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 we can make, the precision necessarily becomes 100%. Note that the plots in Figure 5 do not reach this level — they include only bins with at least 100 users in them — but trend upward with k. (Supplementary Figure 3 shows results for all values of k.)

As with precision, RMSE declines with k (right panel) 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, the left panel of Figure 5 also shows the precision of the naïve baseline obtained by ranking items by their global popularity. This baseline outperform CF and CF+D but at the price of providing the same set of recommendations to all users (i.e., the results are not personalized) and of providing recommendations of lower trustworthiness (Figure 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 (i.e., $k < 10$) 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 (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 Figure 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 among a narrow subset of highly active news consumers with heavily slanted information diets [16, 20]. We therefore take advantage of the survey and behavioral data available on participants in the Pulse panel to consider how CF+D effects vary by individual partisanship (self-reported via survey), behavioral measures such as volume of news consumption activity and information diet slant, and contextual factors that are relevant to algorithm performance such as similarity with 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; see Methods C). However, we analyze 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 list [28], 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 (see Eq. 8 in Method H), we compute the expected change in trustworthiness Q from switching the selection from X' to X,

$$\Delta Q = \mathsf{E}\left[Q(X)\right] - \mathsf{E}\left[Q(X')\right] \tag{1}$$

where the expectations of $Q(X)$ and $Q(X')$ are taken with regard to the respective rankings (see Methods H). 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 Figures 10–16.)

Applying Eq. 1, we find that CF+D substantially increases trustworthiness for users who tend to visit sources that lean conservative (Figure 6(a)) and for those who have the most polarized information diets (in either direction; see Figure 6(c)), two segments of users who are especially likely to be exposed to unreliable information [2, 16, 20]. In both cases, CF+D achieves the greatest improvement among 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 ID

and partisanship of respondents, we use only the former variable (Figure 6(b)) for stratification to avoid circularity given the definition of CF+D, which relies on the latter. In Figures 6(a) and 6(c), we instead stratify on an external measure of news diet slant (calculated from a large sample of social media users; see Methods I).

We also observe that CF+D has strong positive effects for users who identify as Republicans or lean Republican (Figure 6(b)) and for those who are the most active news consumers in terms of both total consumption (Figure 6(d)) and number of distinct sources (Figure 6(e)). Furthermore, since the two recommendation schemes considered here (CF and CF+D) are predicated on identifying similar users according to their tastes and behaviors, we also segment the users of the YouGov sample according to the degree of similarity with their nearest neighbors (identified based on Kendall's rank correlation coefficient between user vectors; see Methods C). Stratifying on the average of nearest neighbor similarities, we find that CF+D results in improvements for the users whose browsing behavior is most similar to others in their neighborhood and who might thus be most at risk of "echo chamber" effects (Figure 6(f)). 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 (Figure 6(g)). 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 [38, 42], designing methods to present diverse political opinions [34, 35], and measuring the prevalence of filter bubbles [10]. 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 tradeoff 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 search engines and social media platforms should consider including audience diversity to their existing set of news quality signals. Such a change could be especially valuable for domains for which we lack other quality signals, like 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 among 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 (i.e. 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 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 [21, 30] and platforms already collect a wealth of signals to detect and remove inauthentic coordinated behavior of this kind. Future work should investigate the feasibility of creating trusted social media audiences that are modeled 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 IRB under protocols #HUM00161944 (University of Michigan) and #STUDY000433 (University of South Florida).

A. Data

Our analysis combines two sources of data. The first is the NewsGuard News Website Reliability Index [37], 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 includes scores of 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 November 12, 2019 and do not reflect subsequent updates; see Data Availability 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 like 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 Figure 1.

The second data source is the YouGov Pulse panel, a sample of U.S.-based Internet users whose web traffic was collected in anonymized form with their prior consent. This traffic data was collected during seven periods between October 2016 and March 2019 (see Supplementary Table 6). A total of 6,890 participants provided data. Overall, this group is diverse and resembles the U.S. population on key demographic and political dimensions (47.9% male, 29.0% with a four-year college degree, 67.9% white, median age of 55, 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 thirty 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 the Supplementary Table 1.

B. 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 $j = 1, \ldots, 7$ (where 1 = strong Democrat and 7 = 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: $\sigma^2 = N^{-1} \sum (s_i - \overline{s})^2$, where $\overline{s}$ is average partisanship;

Shannon's entropy: $S = -\sum p(j) \log p(j)$, where $p(j)$ is estimated in the following three different ways: (*i*) $p(j) = p_j$ (maximum likelihood); (*ii*) $p(j) = \frac{N_j + \alpha}{N + 7\alpha}$ (mean of the posterior distribution of Dirichlet prior with $\alpha = 1$); and (*iii*) the method of Nemenman et al. [36], which uses a mixture of Dirichlet priors (NSB prior).

Complementary Maximum Probability: $1 - \max_j \{p_j\}$;

Complementary Gini: $1 - G$ where G is the Gini coefficient of the count distribution $\{N_j\}_{j=1\ldots7}$.

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 weighs 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 is highly heterogeneous across internet users [19, 33], 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 the Supplementary Table 1. We find that all variants correlate with news reliability, but the relationship is strongest for variance.

C. 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 $f : \mathcal{U} \times \mathcal{D} \to \mathbb{R}$ 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 (e.g., `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 (i.e., pageviews) and use this variable as a proxy [28]. Because this quantity is known to follow a very skewed distribution, we compute the rating as the TF-IDF of the pageview counts:

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

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 (see Methods D).

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*; i.e., their corresponding row vectors in V (i.e., 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:

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

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:

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

where $N_{u_d} \subseteq \mathcal{U}$ is the set of the $n = 10$ most similar users to u who have also rated d (i.e., the neighbors of u), $v_{u',d}$ is the observed rating (computed with Eq. 2) that neighboring 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 $\mathrm{sim}(u,u')$ is the similarity coefficient (computed with Eq. 3) between users u and u' based on either the Pearson or the Kendall correlation coefficient.

Having defined the standard CF in Eq. 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:

$$\hat{v}^{\mathrm{CF+D}}_{u,d} = \hat{v}^{\mathrm{CF}}_{u,d} + g\left(\delta_d\right) \tag{5}$$

where $g\left(\delta_d\right)$ 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, $\delta_d = \sigma^2_d$) into a standard logistic function:

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

In Eq. 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 $a = 1$, $\psi = 1$. As a robustness check, we re-ran all analyses with a larger value of a and obtained qualitatively similar results (available upon reasonable request).

D. Supervised learning evaluation workflow

To evaluate both recommendation algorithms, we follow a standard supervised learning workflow. We use precision and root mean squared error (RMSE), two standard metrics used to measure the relevance and accuracy of predicted ratings in supervised learning settings. We define these two metrics elsewhere (see Methods G). 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 (i.e., the number of visits to each domain, which are computed using Eq. 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: 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 Eq. 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 Eq. 3, a domain is present for a user but not for the other, then 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 Eq. 4 or Eq. 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.

E. 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.

F. Trustworthiness metrics

In addition to standard metrics of accuracy (precision and RMSE; see Methods G), 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 trustworthy [37]. Let $d_1, d_2, \ldots, d_k$ be a ranked list of domains. Using numerical scores, the trustworthiness is the average:

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

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 Eq. 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.

G. 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 $\hat{v}_{u,d}^{\mathrm{CF}}$ and $\hat{v}_{u,d}^{\mathrm{CF+D}}$ produced by CF and CF+D and the actual rating $v_{u,d}$ (defined elsewhere; see Methods C). 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 (either CF or CF+D) and denote with $r(d)$ (respectively, $r'(d)$) the rank of d when the domains are sorted by decreasing order of recommen-

dation 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 *root mean squared error* for a given value of k between the two ranked lists of ratings is computed as:

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

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.

H. Discounting via ranking

To measure the effect of CF+D on the trustworthiness of rankings, we must select a particular list length k. Although Figure 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 (i.e., 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 ranking [11], 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:

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

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 (either 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 (see Eq. 1).

In practice, to compute Eq. 1, let $d_1, d_2, \ldots, d_k$ and $d'_1, d'_2, \ldots, d'_k$ be two ranked lists of domains, $d_r, d'_r \in \mathcal{D}\ \forall r = 1, \ldots, k$, generated by a recommendation algorithm and by actual user pageviews, respectively, and let us denote with $Q(d)$ the NewsGuard reliability score of $d \in \mathcal{D}$ (see Methods F). Recall that Eq. 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 Eq. 8 only depends on r. We can thus calculate the expected improvement in trustworthiness as:

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

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

I. Stratification analysis

Recall that we use the self-reported partisanship of respondents in the YouGov Pulse panel as the basis for our diversity signal (see Methods B). To avoid the circular reasoning in stratifying on the same source of data, Figure 6(a) and Figure 6(c) group these users according to the slant of their actual news consumption, which may not necessarily reflect their self-reported partisanship (e.g., 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. based their classification on how hard news from those domains were shared on Facebook by users who self-identified as liberal or conservative in their profile. For almost all domains, Bakshy et al.

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 $[-1, 1]$ using the measurement approach described by Bakshy et al. [5].)

In Figure 6(c), 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 (i.e., they are almost exclusively shared by one group and not the other).

Data Availability

Data necessary to reproduce the findings in the main manuscript text and in the Supplementary Materials 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.

Code Availability

Code necessary to reproduce the findings in the main text and in the Supplementary Materials are available at https://github.com/glciampaglia/InfoDiversity/.

ACKNOWLEDGEMENTS

We thank NewsGuard for licensing the data and acknowledge Andrew Guess and Jason Reifler, Nyhan's coauthors on the research project that generated the web traffic data used in this study. We are also grateful to organizers, chairs and participants of the News Quality in the Platform Era workshop (organized by the Social Science Research Council), especially Regina Lawrence, Philip Michael Napoli, Kevin Munger, Johanna Dunaway, Connie Moon Sehat and Jieun 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 National Science Foundation. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

AUTHOR CONTRIBUTIONS STATEMENT

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

COMPETING INTERESTS STATEMENT

The authors declare no competing interests.

FIGURE LEGENDS/CAPTIONS

FIG. 1. Relationship between audience size (log-transformed) and news reliability by domain (blue solid line; $N = 1024$ domains). Left: audience size as number of individual visitors. Right: audience size as number of visits. The shaded area the 95% confidence interval. Note that the two panels use different scales on the x axis and do not start at zero. Reliability scores provided by NewsGuard [37]. Full regression results in Supplementary Table 2.

FIG. 2. Average audience partisanship versus variance ($N = 11,793$ domains). Left: audience size as number of individual visitors. Right: audience size as number of visits. Domains for which we have NewsGuard reliability scores [37] are shaded in blue (where darker shades equal lower scores). Domains with no available score are plotted in gray.

FIG. 3. Relationship between audience partisan diversity and news reliability for websites whose average visitor is a Democrat (blue solid line) or a Republican (red solid line). Left panel: variance computed at user level ($N = 1020$ domains). Right panel: variance computed at pageview level ($N = 1024$ domains). The shaded area represents 95% confidence intervals. News reliability scores from NewsGuard [37]. Full regression results in Supplementary Table 3.

FIG. 4. Trustworthiness of recommended domains by length of ranked list k ($N_k = 28$ list lengths). Left: Trustworthiness based on scores from NewsGuard [37]. Right: proportion of domains labeled as 'trustworthy', also by NewsGuard. Actual visits v are normalized using TF-IDF (term frequencyinverse document frequency, see Methods C). Global popularity is overall domain popularity (see Methods E). 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 standard error of the mean. The values of k are capped so that each bin has ≥ 100 users in it (see Supplementary Figure 2 for plot with all values of k). In this figure, both CF (collaborative filtering) and CF+D (collaborative filtering + diversity) compute the similarity between users using the Kendall τ correlation coefficient (see Methods C). We obtain qualitatively similar results using the Pearson correlation coefficient (see Supplementary Figure 4).

FIG. 5. Accuracy of domain recommendations by length of ranked list k ($N_k = 28$ list lenghts). Left: Precision (proportion of correctly ranked sites) by length of ranked list k (higher is better). Right: RMSE (root mean squared error) 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 standard error of the mean. The values of k are capped so that each bin has ≥ 100 users in it (see Supplementary Figure 3 for plot with all values of k). In this figure, both CF (collaborative filtering) and CF+D (collaborative filtering + diversity) compute the similarity between users using the Kendall τ correlation coefficient (see Methods C). We obtain qualitatively similar results using the Pearson correlation coefficient (see Supplementary Figure 5).

FIG. 6. Effect of CF (collaborative filtering) and CF+D (collaborative filtering + diversity; versus actual visits baseline) on trustworthiness by user characteristics and behavior. (a) Ideological slant of visited domains (terciles using scores from Bakshy et al. [5]). (b) Self-reported party ID 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.). (d) Total online activity (TF-IDF-transformed pageviews; terciles; TF-IDF is short for term frequencyinverse document frequency). (e) Distinct number of domains visited (terciles). (f) Average user-user similarity with nearest $n = 10$ neighbors in training set (terciles) (g) Trustworthiness of domains visited by users (in training set; terciles). Bars represent the standard error of the mean of each stratum. Change in trustworthiness ΔQ based on scores from NewsGuard [37].

FIG. 7. Probability that a trustworthy domain (NewsGuard score ≥ 60) is not recommended by CF+D (collaborative filtering + diversity) but is recommended by CF (collaborative filtering) 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 standard error of the mean. 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 full summary).

REFERENCES

[1] Adler, B. T. and de Alfaro, L. (2007). A content-driven reputation system for the Wikipedia. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 261–270, New York, NY, USA. ACM.

[2] Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36.

[3] Allen, J., Howland, B., Mobius, M., Rothschild, D., and Watts, D. J. (2020). Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, 6(14).

[4] Avram, M., Micallef, N., Patil, S., and Menczer, F. (2020). Exposure to social engagement metrics increases vulnerability to misinformation. *Harvard Kennedy School Misinformation Review*, 1(5).

[5] Bakshy, E., Messing, S., and Adamic, L. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239):1130–1132.

[6] Chen, W., Pacheco, D., Yang, K.-C., and Menczer, F. (2020). Neutral bots reveal political bias on social media. arXiv e-print 2005.08141, CoRR.

[7] Cho, J.-H., Chan, K., and Adali, S. (2015). A survey on trust modeling. *ACM Comput. Surv.*, 48(2):28:1–28:40.

[8] Ciampaglia, G. L., Nematzadeh, A., Menczer, F., and Flammini, A. (2018). How algorithmic popularity bias hinders or promotes quality. *Scientific Reports*, 8(1):15951–.

[9] FB, I. (2020). Prioritizing original news reporting on facebook. Retrieved from *Internet Archive*: https://web.archive.org/web/20210126011953/https://about.fb.com/news/2020/06/prioritizing-original-news-reporting-on-facebook/.

[10] Flaxman, S., Goel, S., and Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, 80(S1):298–320.

[11] Fortunato, S., Flammini, A., and Menczer, F. (2006). Scale-free network growth by ranking. *Physical Review Letters*, 96(21):218701.

[12] Germano, F., Gómez, V., and Le Mens, G. (2019). The few-get-richer: A surprising consequence of popularity-based rankings? In *The World Wide Web Conference*, WWW '19, pages 2764–2770, New York, NY, USA. ACM.

[13] Goel, S., Broder, A., Gabrilovich, E., and Pang, B. (2010). Anatomy of the long tail: Ordinary people with extraordinary tastes. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM 10, page 201210, New York, NY, USA. Association for Computing Machinery.

[14] Golbeck, J. A. (2005). *Computing and applying trust in web-based social networks*. PhD thesis, University of Maryland at College Park.

[15] Google, I. (2020). Surfacing useful and relevant content – how news works. Retrieved from *Internet Archive* :https://web.archive.org/web/20201017172355/https://newsinitiative.withgoogle.com/hownewsworks/approach/surfacing-useful-and-relevant-content/.

[16] Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., and Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science*, 363(6425):374–378.

[17] Guess, A., Lyons, B., Nyhan, B., and Reifler, J. (2018). 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.

[18] Guess, A., Nagler, J., and Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, 5(1).

[19] Guess, A. M. (2018). (almost) everything in moderation: New evidence on americans' online media diets. Unpublished manuscript.

[20] Guess, A. M., Nyhan, B., and Reifler, J. (2020). Exposure to untrustworthy websites in the 2016 US election. *Nature Human Behaviour*, 4(5):472–480.

[21] Gunes, I., Kaleli, C., Bilge, A., and Polat, H. (2014). Shilling attacks against recommender systems: a comprehensive survey. *Artificial Intelligence Review*, 42(4):767–799.

[22] Gupta, A., Kumaraguru, P., Castillo, C., and Meier, P. (2014). *TweetCred: Real-Time Credibility Assessment of Content on Twitter*, pages 228–243. Springer International Publishing, Cham.

[23] Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., and Merrill, L. (2009). Feeling validated versus being correct: a meta-analysis of selective exposure to information. *Psychological Bulletin*, 135(4):555.

[24] Hogg, T. and Lerman, K. (2015). Disentangling the effects of social signals. *Human Computation*, 2(2):189–208.

[25] Hong, L. and Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46):16385–16389.

[26] Jiang, S., Baumgartner, S., Ittycheriah, A., and Yu, C. (2020). *Factoring Fact-Checks: Structured Information Extraction from Fact-Checking Articles*, pages 1592–1603. Association for Computing Machinery, New York, NY, USA.

[27] Jiang, S. and Wilson, C. (2018). Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW).

[28] Joachims, T., Granka, L., Pan, B., Hembrooke, H., and Gay, G. (2017). Accurately interpreting clickthrough data as implicit feedback. *SIGIR Forum*, 51(1):411.

[29] Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., and Riedl, J. (1997). GroupLens: Applying collaborative filtering to usenet news. *Commun. ACM*, 40(3):7787.

[30] Lam, S. K. and Riedl, J. (2004). Shilling recommender systems for fun and profit. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, pages 393–402, New York, NY, USA. ACM.

[31] Lazer, D., Baum, M., Benkler, Y., Berinsky, A., Greenhill, K., Menczer, F., Metzger, M., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S., Sunstein, C., Thorson, E., Watts, D., and Zittrain, J. (2018). The science of fake news. *Science*, 359(6380):1094–1096.

[32] Macy, M., Deri, S., Ruch, A., and Tong, N. (2019). Opinion cascades and the unpredictability of partisan polarization. *Science Advances*, 5(8).

[33] Montgomery, A. L. and Faloutsos, C. (2001). Identifying web browsing trends and patterns. *Computer*, 34(7):94–95.

[34] Munson, S., Lee, S., and Resnick, P. (2013). Encouraging reading of diverse political viewpoints with a browser widget. In *International AAAI Conference on Web and Social Media*, ICWSM '13, pages 419–428, Palo Alto, Calif., USA. AAAI.

[35] Munson, S. A. and Resnick, P. (2010). *Presenting Diverse Political Opinions: How and How Much*, pages 1457–1466. Association for Computing Machinery, New York, NY, USA.

[36] Nemenman, I., Shafee, F., and Bialek, W. (2001). Entropy and inference, revisited. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, pages 471–478, Cambridge, MA, USA. MIT Press.

[37] NewsGuard, Inc. (2020). Rating process and criteria. Retrieved from *Internet Archive*: https://web.archive.org/web/20200630151704/https://www.newsguardtech.com/ratings/rating-process-criteria/.

[38] Nikolov, D., Lalmas, M., Flammini, A., and Menczer, F. (2019). Quantifying biases in online information exposure. *Journal of the Association for Information Science and Technology*, 70(3):218–229.

[39] Pennycook, G. and Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, 116(7):2521–2526.

[40] Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., and Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

[41] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J. (1994). GroupLens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, CSCW 94, page 175186, New York, NY, USA. Association for Computing Machinery.

[42] Ribeiro, F., Henrique, L., Benevenuto, F., Chakraborty, A., Kulshrestha, J., Babaei, M., and Gummadi, K. (2018). Media bias monitor: Quantifying biases of social media news outlets at large-scale. In

Proceedings of the International AAAI Conference on Web and Social Media, pages 290–299, Palo Alto, CA, USA. AAAI.

[43] Salganik, M. J., Dodds, P. S., and Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762):854–856.

[44] Sayyadiharikandeh, M., Varol, O., Yang, K.-C., Flammini, A., and Menczer, F. (2020). Detection of novel social bots by ensembles of specialized classifiers. In *Proc. 29th ACM International Conference on Information & Knowledge Management (CIKM)*, pages 2725–2732.

[45] Shao, C., Ciampaglia, G. L., Varol, O., Yang, K., Flammini, A., and Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications*, 9(1):4787.

[46] Shi, F., Teplitskiy, M., Duede, E., and Evans, J. A. (2019). The wisdom of polarized crowds. *Nature Human Behaviour*, 3(4):329–336.

[47] Shmargad, Y. and Klar, S. (2020). Sorting the news: How ranking by popularity polarizes our politics. *Political Communication*, 37(3):423–446.

[48] Stella, M., Ferrara, E., and De Domenico, M. (2018). Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences*, 115(49):12435–12440.

[49] Varol, O., Ferrara, E., Davis, C., Menczer, F., and Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterization. In *Proc. Eleventh Intl AAAI Conference on Web and Social Media*, ICWSM '17, pages 280–289, Palo Alto, Calif., USA. AAAI.

[50] Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380):1146–1151.

[51] Yang, K.-C., Torres-Lugo, C., and Menczer, F. (2020a). Prevalence of low-credibility information on Twitter during the COVID-19 outbreak. In *Proc. Fourteenth Intl AAAI Conference on Web and Social Media*, ICWSM '20, Palo Alto, Calif., USA. AAAI.

[52] Yang, K.-C., Varol, O., Davis, C. A., Ferrara, E., Flammini, A., and Menczer, F. (2019). Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, 1(1):48–61.

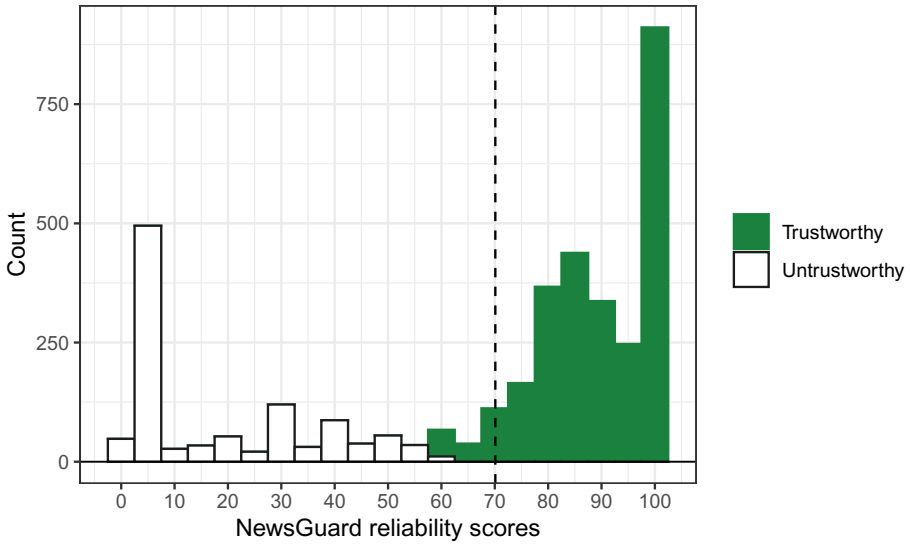
[53] Yang, K.-C., Varol, O., Hui, P.-M., and Menczer, F. (2020b). Scalable and generalizable social bot detection through data selection. In *Proc. 34th AAAI Conf. on Artificial Intelligence (AAAI)*.

[54] Zhang, A. X., Ranganathan, A., Metz, S. E., Appling, S., Sehat, C. M., Gilmore, N., Adams, N. B., Vincent, E., Lee, J., Robbins, M., Bice, E., Hawke, S., Karger, D., and Mina, A. X. (2018). A structured response to misinformation: Defining and annotating credibility indicators in news articles. In *Companion Proc. of The Web Conference 2018*, WWW '18, pages 603–612, Republic and Canton of Geneva, Switzerland. Intl. World Wide Web Conf. Steering Committee.

[55] Ziegler, C.-N., McNee, S. M., Konstan, J. A., and Lausen, G. (2005). Improving recommendation lists through topic diversification. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, pages 22–32, New York, NY, USA. ACM.

Supplementary Materials for Bhadani *et al.*, "*Political audience diversity and news reliability in algorithmic ranking*"

S1. NEWSGUARD DATA



SUPPLEMENTARY FIGURE 1. Distribution of NewsGuard scores ($N = 3{,}726$) by trustworthiness rating. Domains that score below 60 points (i.e., untrustworthy) on the rubric used by NewsGuard [3] are shown in white. Those that score 60 or above are shown in green. The bin width is 5; the bin containing score 60 also includes a few domains with lower scores. The dashed line indicates the average score in the data.

S2. ALTERNATIVE DEFINITIONS OF AUDIENCE DIVERSITY

We repeat the analysis of Fig. 3 for all diversity metrics (see Methods B) and summarize the results in Table 1. For each metric, we estimate the degree of linear association with news quality using the Pearson correlation coefficient. We also report the R^2 coefficient of determination and the two-sided p-value of the F-statistic as a measure of significance of the fit. And finally, we show the partial correlation coefficient by controlling the mean partisanship and the extremity of domains. Each metric is positively correlated with quality at the user level, but we find that the relationship is strongest for variance of audience partisanship. At the pageview level, however, the association disappears for all metrics but variance, which still produces a modest correlation.

SUPPLEMENTARY TABLE 1. Relationship between audience partisan diversity and news quality. (PCC = 'Partial Correlation Coefficient'.)

Diversity metric	**Corr.**	R^2	p	PCC (Mean)		PCC (Extremity)	
($n = 1707$)				**Corr.**	p	**Corr.**	p
USER LEVEL							
Variance	0.32	0.10	$< 10^{-4}$	0.38	$< 10^{-4}$	0.26	$< 10^{-4}$
Entropy (Dir.)	0.21	0.04	$< 10^{-4}$	0.39	$< 10^{-4}$	0.31	$< 10^{-4}$
Entropy (ML)	0.20	0.04	$< 10^{-4}$	0.34	$< 10^{-4}$	0.24	$< 10^{-4}$
Entropy (NSB)	0.22	0.05	$< 10^{-4}$	0.2	$< 10^{-4}$	0.14	$< 10^{-4}$
Compl. Max. Prob.	-0.04	0.00	0.14	0.26	$< 10^{-4}$	0.14	$< 10^{-4}$
Compl. Gini	0.14	0.02	$< 10^{-4}$	0.26	$< 10^{-4}$	0.21	$< 10^{-4}$
PAGEVIEW LEVEL							
Variance	0.14	0.02	$< 10^{-4}$	0.22	$< 10^{-4}$	0.15	$< 10^{-4}$
Entropy (Dir.)	0.03	0.00	0.24	0.044	0.07	0.04	0.09
Entropy (ML)	0.03	0.00	0.19	0.046	0.057	0.042	0.078
Entropy (NSB)	0.03	0.00	0.18	0.048	0.05	0.044	0.07
Compl. Max. Prob.	0.004	0.00	0.86	0.03	0.22	0.019	0.42
Compl. Gini	-0.001	0.00	0.97	0.019	0.43	0.017	0.46

S3. REGRESSION OF NEWSGUARD SCORES ON WEBSITE AUDIENCE VARIANCE

In Fig. 1 in the main text we show the relationship between NewsGuard reliability scores of news domains and traffic, while in Fig. 3 in the main text we show the relationship between NewsGuard reliability scores of news domains and audience partisan diversity, via linear regression. In Tables 2 and 3 we report the associated summary tables for these two regression modeling exercises. Each table provides summary information for fitting a regression model with either user- or pageview-level data, and for different controls.

SUPPLEMENTARY TABLE 2. Relationship between NewsGuard scores and popularity.

	Dependent variable: NewsGuard score			
	Model 1	Model 2	Model 3	Model 4
Logged N visitors	0.533	1.164		
	[−0.598, 1.663]	[0.006, 2.323]		
	$p = 0.356$	$p = 0.049$		
Logged N pageviews			0.720	0.749
			[−0.191, 1.631]	[−0.263, 1.761]
			$p = 0.121$	$p = 0.147$
Conservative indicator		3.840		−11.136
		[−11.177, 18.858]		[−24.230, 1.958]
		$p = 0.616$		$p = 0.095$
N visitor × Conservative		−4.572		
		[−7.252, −1.892]		
		$p = 0.001$		
N pageview × Conservative				−0.594
				[−2.589, 1.400]
				$p = 0.559$
Constant	79.729	81.074	78.045	82.552
	[73.302, 86.156]	[74.448, 87.701]	[72.006, 84.083]	[75.802, 89.303]
	$p < 10^{-4}$	$p < 10^{-4}$	$p < 10^{-4}$	$p < 10^{-4}$
Num.Obs.	1024	1020	1024	1024
R^2	0.001	0.154	0.002	0.089
R^2 Adj.	0.000	0.152	0.001	0.086
F	0.854	61.657	2.406	33.245

Note: 95% confidence intervals are reported in square brackets and exact p values below.

SUPPLEMENTARY TABLE 3. Relationship between NewsGuard scores and diversity.

	Dependent variable: NewsGuard score			
	Model 1	Model 2	Model 3	Model 4
User-level variance	6.662	2.926		
	[5.530, 7.793]	[1.626, 4.226]		
	$p < 10^{-4}$	$p < 10^{-4}$		
Pageview-level variance			3.919	0.822
			[2.532, 5.305]	[−0.861, 2.505]
			$p < 10^{-4}$	$p = 0.338$
Conservative indicator		−11.372		−10.336
		[−15.025, −7.718]		[−13.632, −7.041]
		$p < 10^{-4}$		$p < 10^{-4}$
User variance × Conservative		7.715		
		[5.457, 9.974]		
		$p < 10^{-4}$		
Pageview variance × Conservative				7.974
				[5.321, 10.627]
				$p < 10^{-4}$
Constant	87.174	89.184	85.080	87.906
	[85.613, 88.734]	[87.578, 90.790]	[83.420, 86.740]	[85.979, 89.834]
	$p < 10^{-4}$	$p < 10^{-4}$	$p < 10^{-4}$	$p < 10^{-4}$
Num.Obs.	1024	1020	1024	1024
R^2	0.116	0.253	0.029	0.147
R^2 Adj.	0.115	0.251	0.028	0.145
F	133.545	114.590	30.753	58.661

Note: 95% confidence intervals are reported in square brackets and exact p values below. User-level and pageview-level variance measures are standardized to have zero mean and unit variance.

S4. CORRELATIONS BETWEEN DOMAIN POPULARITY AND AUDIENCE DIVERSITY

In Table 4 we show the Pearson correlation coefficients between the popularity of a domain and its diversity. We operationalize the popularity of website as either its audience size (i.e., number of unique users) or its traffic (number of pageviews). For our diversity measures, we rely on the user-level and pageview-level partisanship variance.

Overall, domain popularity is very weakly correlated with the variance of audience partisanship regardless of how we choose to operationalize each measure. Recall that in our original analysis we show that domain popularity is largely uncorrelated with quality (as proxied by NewsGuard scores). Together, these findings suggest that audience partisan diversity is associated with quality of news independent of the variation caused by domain popularity.

SUPPLEMENTARY TABLE 4. Pearson correlation coefficients between domain diversity and popularity

Variance (rows) / Popularity (columns)	N Unique users	N Pageviews
User-level	0.04 ($p < 10^{-4}$)	0.0093 ($p = 0.31$)
Pageview-level	0.062 ($p < 10^{-4}$)	0.019 ($p = 0.038$)

Furthermore, we estimate multivariate regressions interacting our diversity measures with an indicator of whether the website has a predominantly Democratic or Republican website with the following model:

$$\begin{aligned}\text{Reliability} = \beta_0 + \beta_1(\text{Diversity measure}) + \beta_2(\text{Conservative website dummy}) + \\ \beta_3(\text{Logged audience size}) + \beta_4(\text{Diversity measure} \times \text{Conservative website dummy})\end{aligned}$$

We estimate two separate regression models, using our two operational diversity measures: user-level and pageview-level partisanship variance. At the user level, after controlling for audience size, a one-standard deviation increase in diversity is associated with a 2.91 point increase in NewsGuard reliability for websites with predominantly Democratic audiences, and with a 10.8 point increase for websites with predominantly Republican audiences. Both estimates are statistically significant.

At the pageview level, we find that a one-standard deviation increase is associated with a 0.68 point increase (statistically indistinguishable from zero) in NewsGuard reliability for Democratic websites, and with a 9.24 point increase for Republican websites. In summary, both methods indicate that our diversity measure is a good predictor of journalistic quality, independent of

audience size. This relationship is especially strong for websites with predominantly Republican audiences.

S5. ROBUSTNESS CHECKS

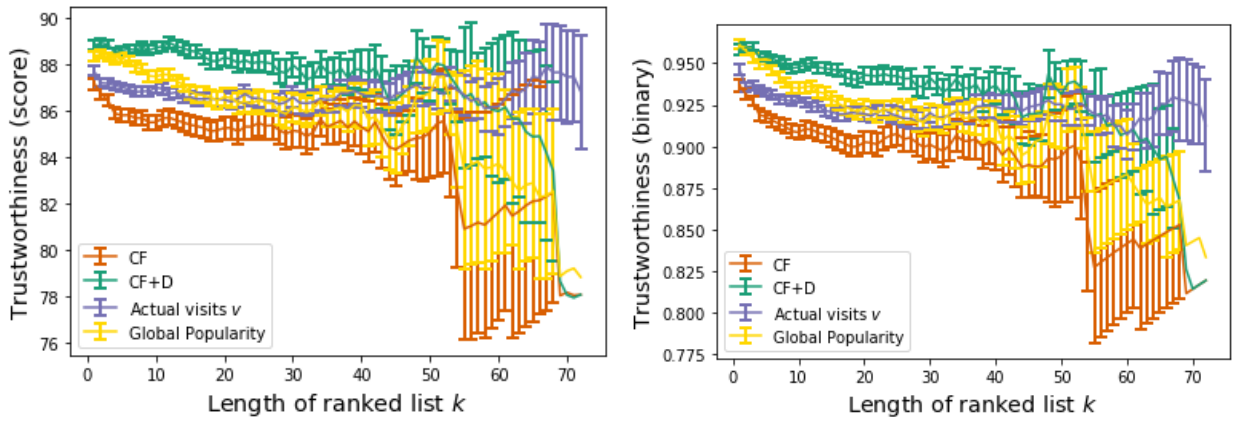
a. No minimum frequency capping. Figs. 2 and 3 are the analogous of Figs. 4 and 5 from the main text, but unlike the plots in the main text, which capped the range of k to include only bins with a minimum frequency, the plots here show all possible values of k.

b. Alternative similarity metric based on rank correlation. Figs. 4 and 5 also show the results of analyses analogous to those in Figs. 4 and 5, but unlike the plots in the main text, which used the Kendall rank correlation coefficient to compute the similarity between users, the plots here show the results obtained using the Pearson correlation coefficient. Moreover, the plots here show all possible values of k, without the aforementioned cap. To get a better sense sense of this difference, Fig. 6 shows the distribution of the number of users as a function of the length of the ranked list k. We observe that Pearson tends to produce smaller recommendation lists than Kendall.

c. Longitudinal analysis. Figs. 7 and 8 show the results of an analysis analogous to those in Figs. 4 and 5, but in which training and testing sets are split longitudinally instead of randomly. In this sense, they represent a true forecasting exercise. Despite a slightly larger loss of precision relative to CF (compare the left panel of Fig. 5 in the main text with the left panel of Fig. 8), our results remain qualitatively consistent with those shown in the main text. For the prior Figs. 4, 5, 2, 3, 4 and 5, the data for each user are randomly split into a training (70%) and testing set (30%), so that, for any given user, there is no overlap between the two sets. Note that each user is split independently of the others, so a given domain can appear in the training set of one user and in the testing set of another. Instead, in Figs. 7 and 8, the data of traffic that took place before a fixed boundary date (which is identical for all users) form the training set, and those that took place after form the testing set. This means that the same domain can occur in both the training and the testing set.

Data collection for the YouGov Pulse panel took place in 7 different time periods (see Table 6), but for simplicity we considered only 3 waves (the first three). Figs. 7 and 8 show the analysis performed on the first wave of data collection, which took place between October 7 and November 14, 2016, and we split the data using November 1, 2016 as boundary. We find qualitatively similar results for the second and third waves. (Data available upon reasonable request to the authors.)

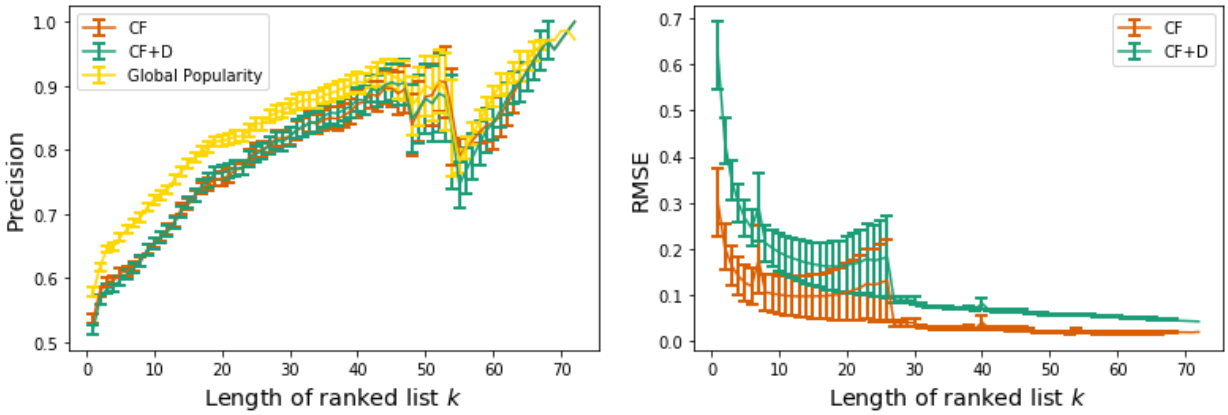
d. Resampling. To estimate the significance of the observed drop in precision of CF+D, we simulate the process of re-ranking a list of items. Recommendations are obtained in this context by sorting items by their predicted rating. Since CF+D simply shifts the rating of each item by adding a term that depends on diversity (see Eq. 5), we simulate this process by simply shuffling



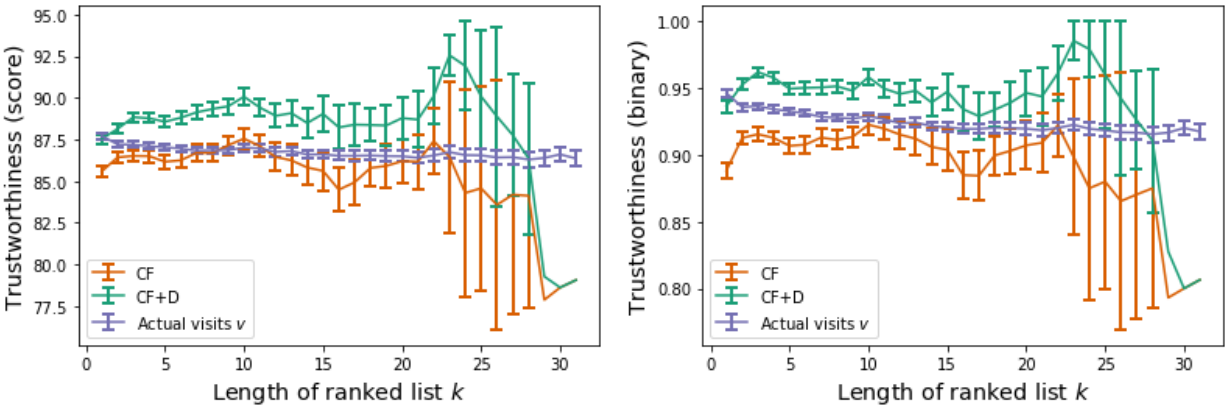
SUPPLEMENTARY FIGURE 2. Trustworthiness of recommended domains by length of ranked list k, for all values of k. Left: Trustworthiness based on scores from NewsGuard [3]. Right: proportion of domains labeled as 'trustworthy,' also by NewsGuard. Actual visits v are normalized using TF-IDF (see Methods C). 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 standard error of the mean. In this figure, both CF and CF+D compute the similarity between users using the Kendall τ correlation coefficient (see Methods C).

the diversity terms among the items before ranking them. This procedure ensures that we consider only lists obtained by shifting the ratings by the same amount of CF+D. Fig. 9 shows the sampling distribution of the precision of re-rankings of the same magnitude as those of CF+D using this process for $k = 1$ and $k = 10$. To sample from this distribution, we rank domains using the ratings computed from Eq. 4. We then compute in a separate labeled vector the diversity term $g(\delta_d)$ obtained using the logistic function (Eq. 6), reshuffle the labels at random, obtaining for each term a new label d', and finally apply the reshuffled term $g(\delta_{d'})$ as in Eq. 5. We then re-rank based on the new ratings and compute the precision of the ranked list. This reshuffling is carried out separately for each user with at least k domains in their testing set. The precision is then averaged over all users. This procedure is repeated 1,000 times to obtain the sampling distribution. Finally, we compute a one-tailed p-value by finding the proportion of samples that have a precision higher than the observed value for CF+D.

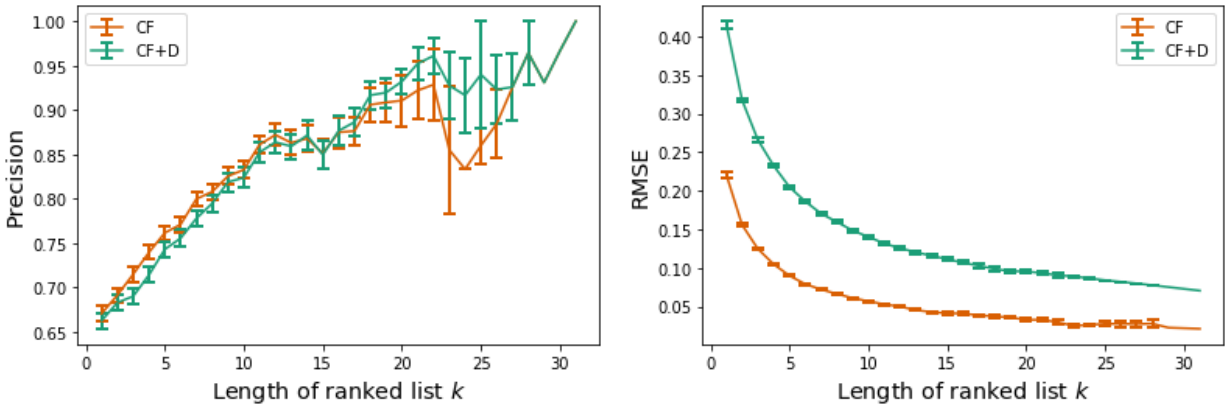
e. Stratification analysis without discounting. Fig. 10–16 show the results of the stratification analysis without using the discounting model.



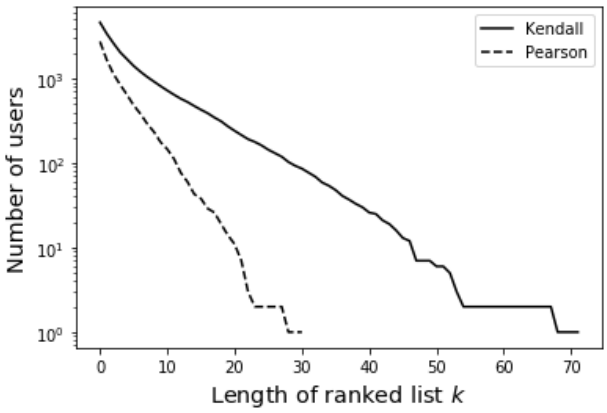
SUPPLEMENTARY FIGURE 3. Accuracy of domain recommendations by length of ranked list k, for all values of k. Left: Precision (proportion of correctly ranked sites) by length of ranked list k (higher is better). Right: RMSE (root mean squared error) of predicted pageviews for top k ranked domains by length 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 standard error of the mean. In the last bin ($k = 73$) precision is 100% for all users. In this figure, both CF and CF+D compute the similarity between users using the Kendall τ correlation coefficient (see Methods C).



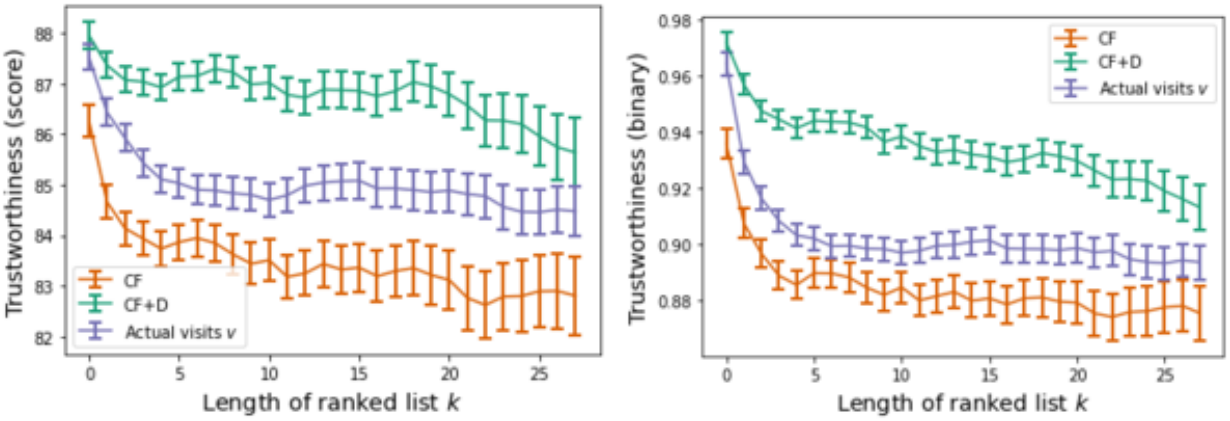
SUPPLEMENTARY FIGURE 4. Trustworthiness of recommended domains by length of ranked list k, for all values of k. Left: Trustworthiness based on scores from NewsGuard [3]. Right: proportion of domains labeled as 'trustworthy,' also by NewsGuard. Actual visits v are normalized using TF-IDF (see Methods C). All results represent averages computed for all users in the YouGov panel. Bars represent the standard error of the mean. In this figure, both CF and CF+D compute the similarity between users using the Pearson correlation coefficient.



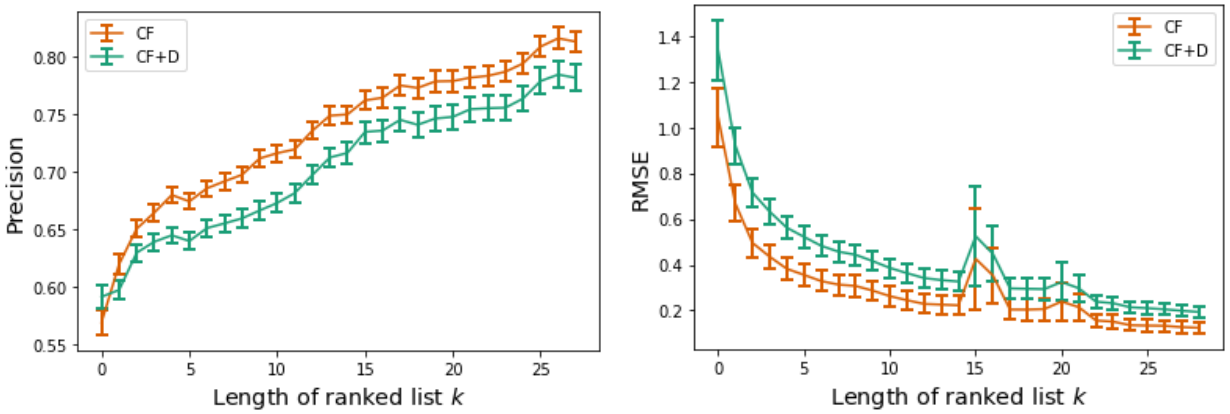
SUPPLEMENTARY FIGURE 5. Accuracy of domain recommendations by length of ranked list, for all values of k. Left: Precision (proportion of correctly ranked sites) by length of ranked list k (higher is better). Right: RMSE (root mean squared error) of predicted pageviews for top k ranked domains by length 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 standard error of the mean. In the last bin ($k = 30$) precision is 100% for all users. Bars represent the standard error of the mean. In this figure, both CF and CF+D compute the similarity between users using the Pearson correlation coefficient.



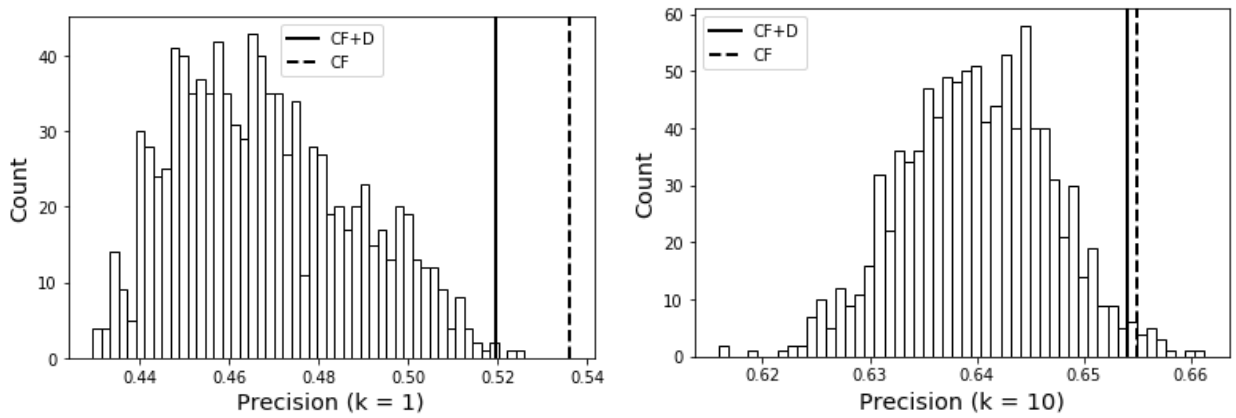
SUPPLEMENTARY FIGURE 6. Number of users with k domains in the test set for neighborhoods (the set of the $n = 10$ most similar users to a given user) computed using the correlation coefficient of Kendall (solid line) and Pearson (dashed line). In general, Pearson leads to shorter lists of recommendations.



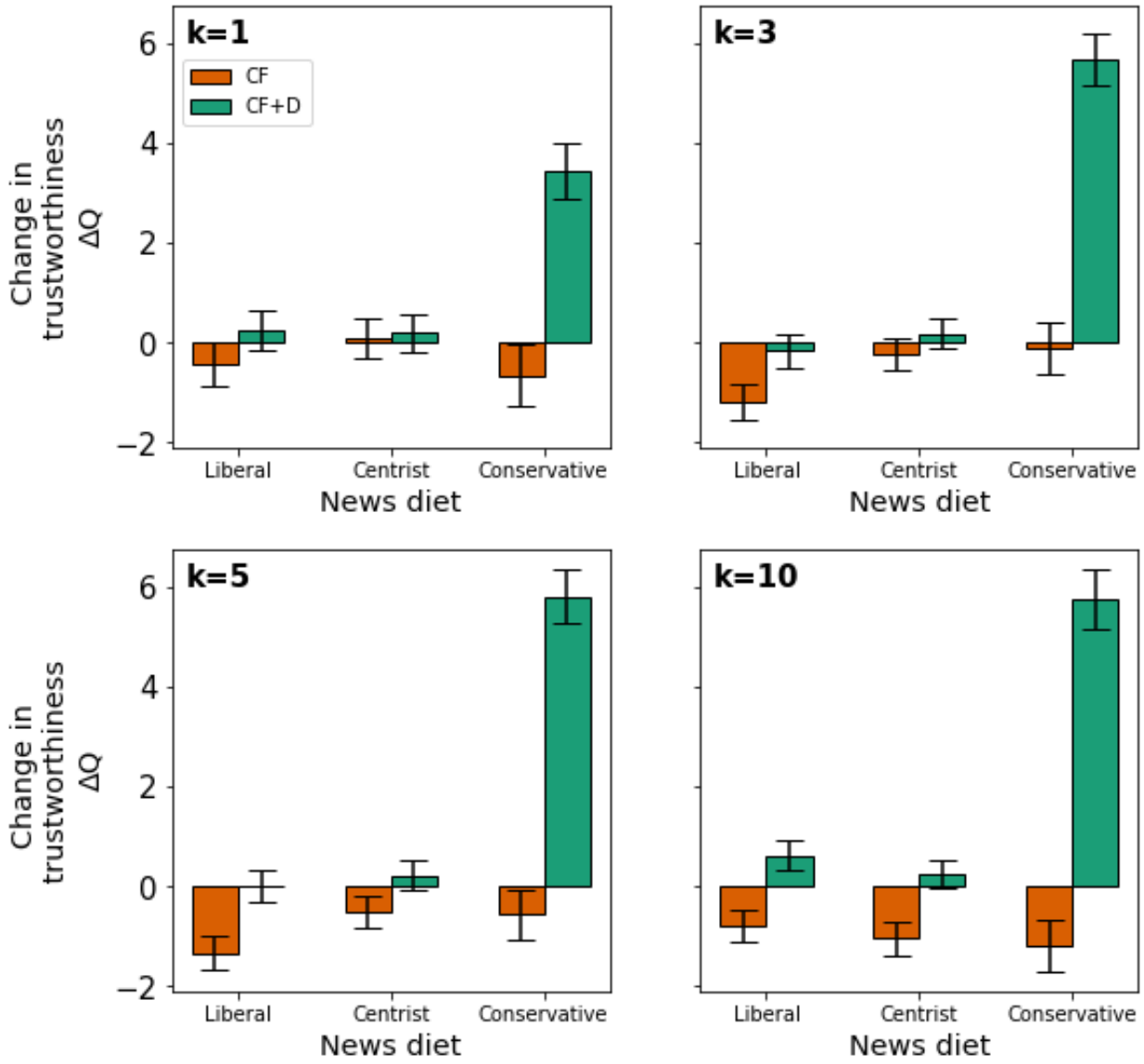
SUPPLEMENTARY FIGURE 7. Trustworthiness of recommended domains by length of ranked list k when the data for training and test sets for the first wave of users are split based on time Left: Trustworthiness based on scores from NewsGuard [3]. Right: proportion of domains labeled as 'trustworthy,' also by NewsGuard. Actual visits v are normalized using TF-IDF (see Methods C). 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 standard error of the mean. In this figure, both CF and CF+D compute the similarity between users using the Kendall τ correlation coefficient (see Methods C).



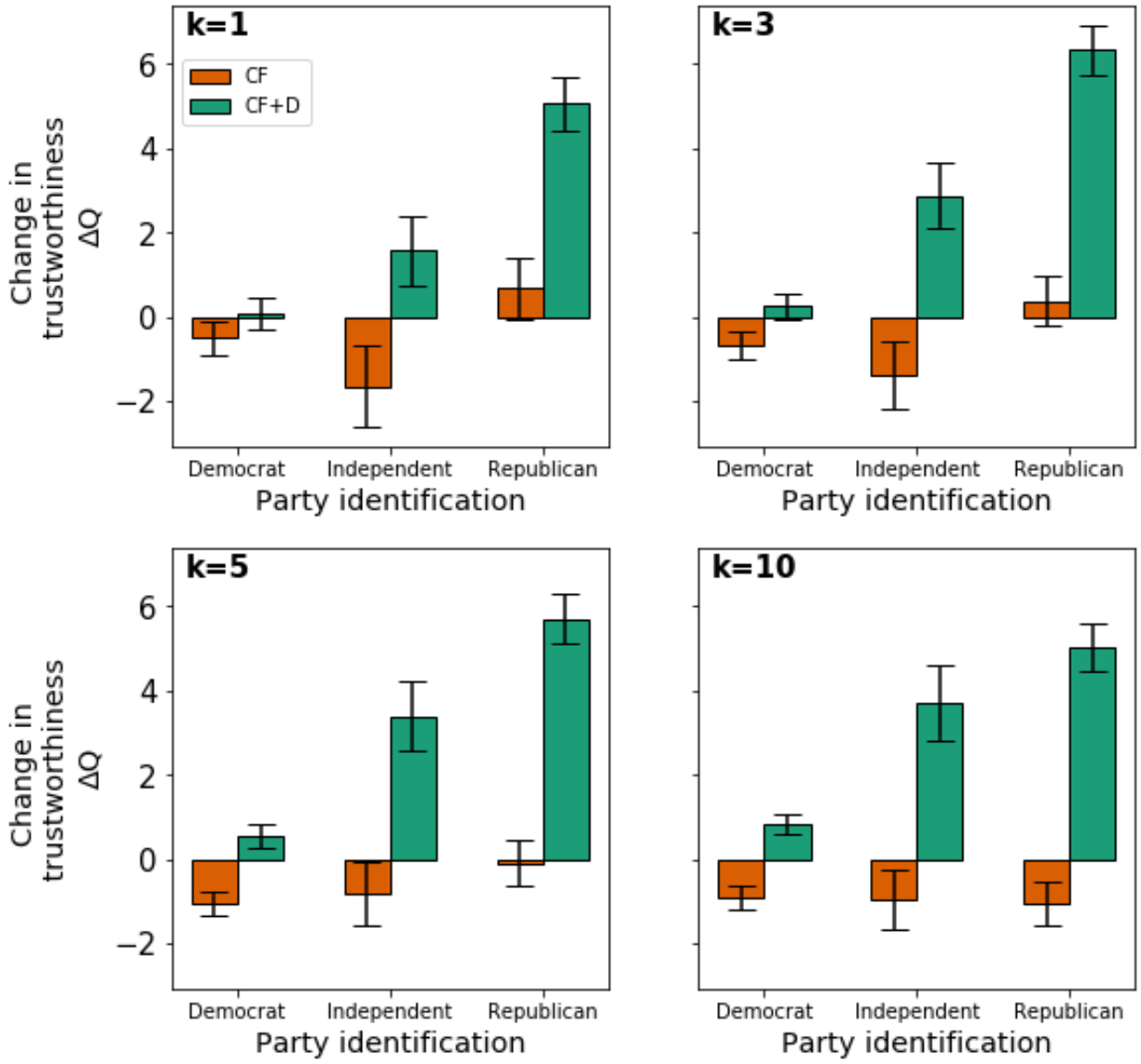
SUPPLEMENTARY FIGURE 8. Accuracy of domain recommendations by length of ranked list k when the data for training and test sets for the first wave of users are split based on time. Left: Precision (proportion of correctly ranked sites) by length of ranked list k (higher is better). Right: RMSE (root mean squared error) 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 standard error of the mean. In the last bin ($k = 73$) precision is 100% for all users. In this figure, both CF and CF+D compute the similarity between users using the Kendall τ correlation coefficient (see Methods C).



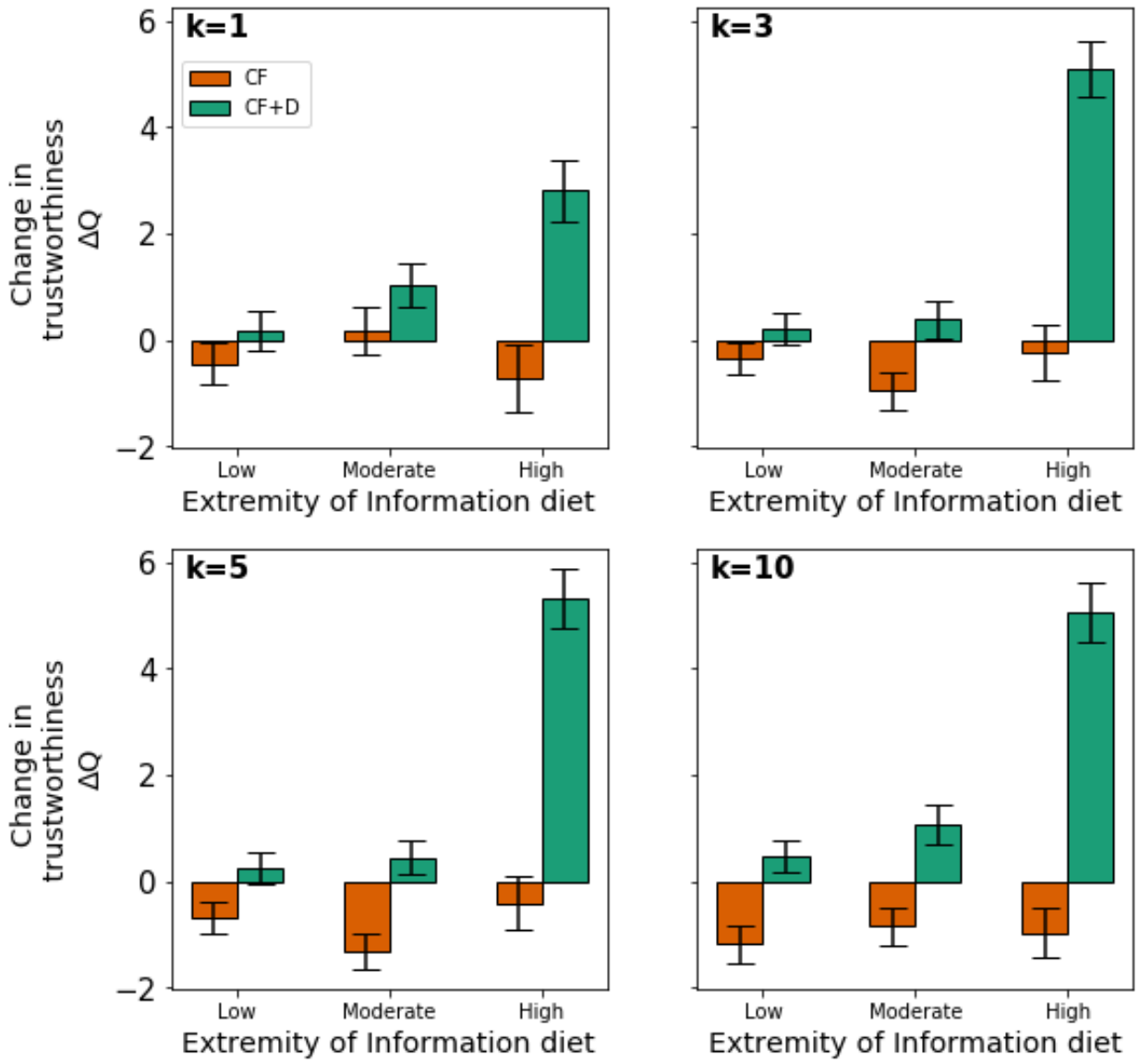
SUPPLEMENTARY FIGURE 9. Distribution of precision obtained after re-ranking the domains, by means of re-shuffling the diversity signal values $g(\delta_d)$ from the CF+D ratings calculation (see Eq. 5 and Eq. 6). The re-shuffling was repeated $1,000$ times. The two distributions correspond to different values of k. The (one-sided) p-values are 0.002 ($k = 1$) and 0.021 ($k = 10$). The two vertical lines correspond to the observed precision values of CF+D (solid) and CF (dashed).



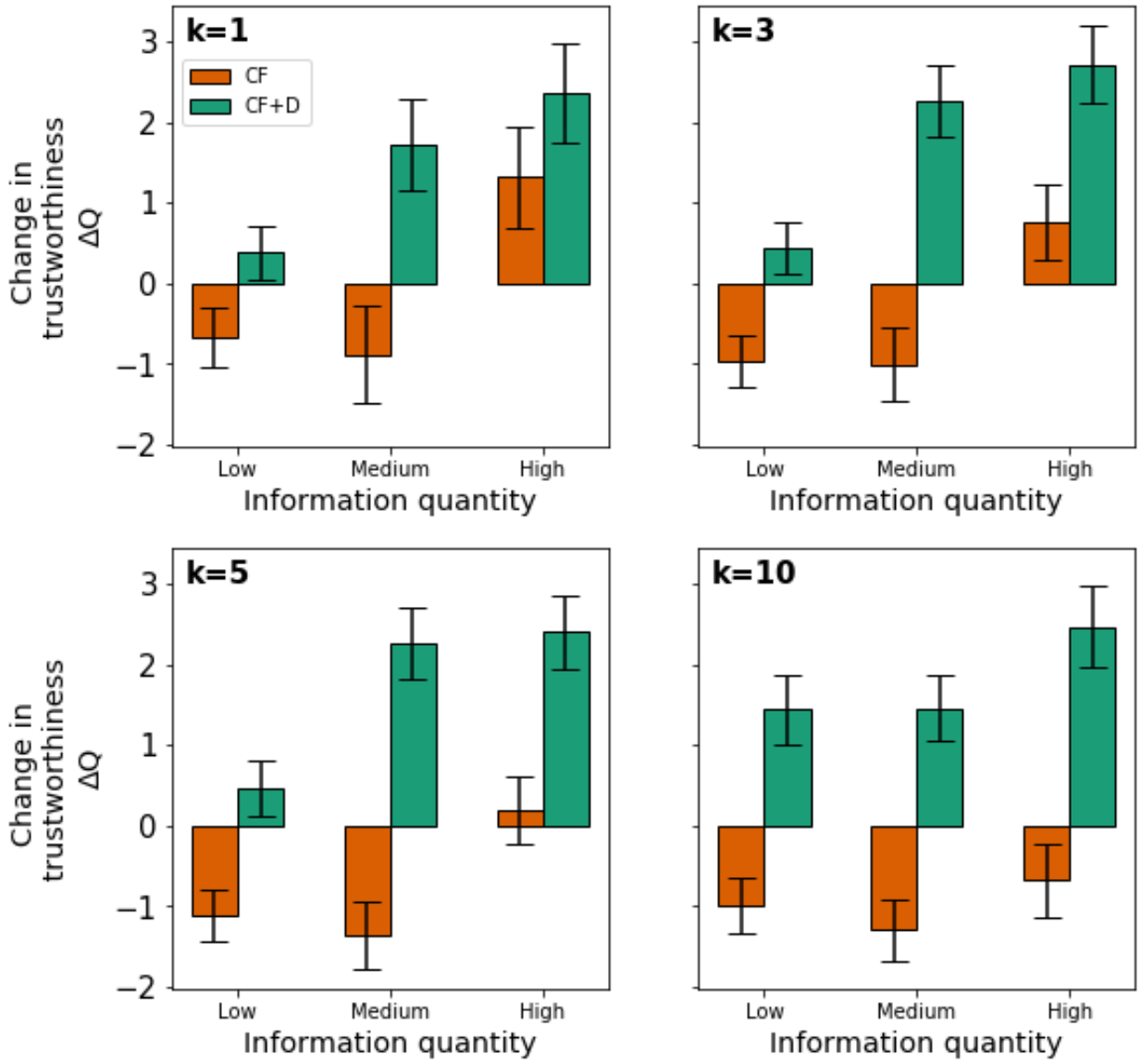
SUPPLEMENTARY FIGURE 10. Effect of CF and CF+D versus baseline by ideological slant of visited domains (terciles using scores from Bakshy et al. [1]) and by length of ranked list k. In this and the following plots, bars represent the standard error of the mean. Change in trustworthiness ΔQ based on scores from NewsGuard [3].



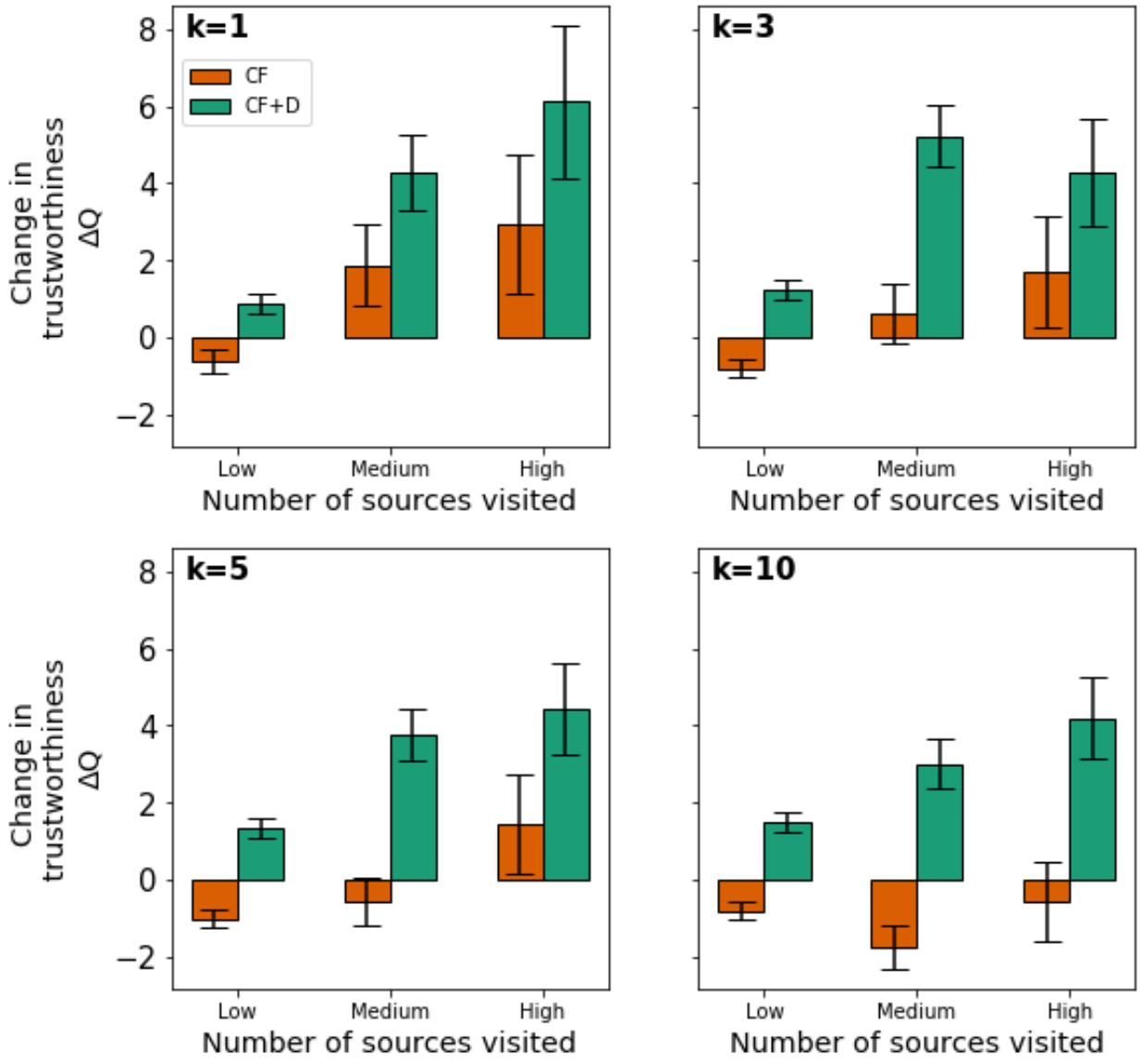
SUPPLEMENTARY FIGURE 11. Effect of CF and CF+D versus baseline by self-reported party ID from YouGov Pulse responses as measured on a 7-point scale (1–3: Democrats including people who lean Democrat but do not identify as Democrats, 4: Independents, 5–7: Republicans including people who lean Republican but do not identify as Republicans) and by length of ranked list k. Change in trustworthiness ΔQ based on scores from NewsGuard [3].



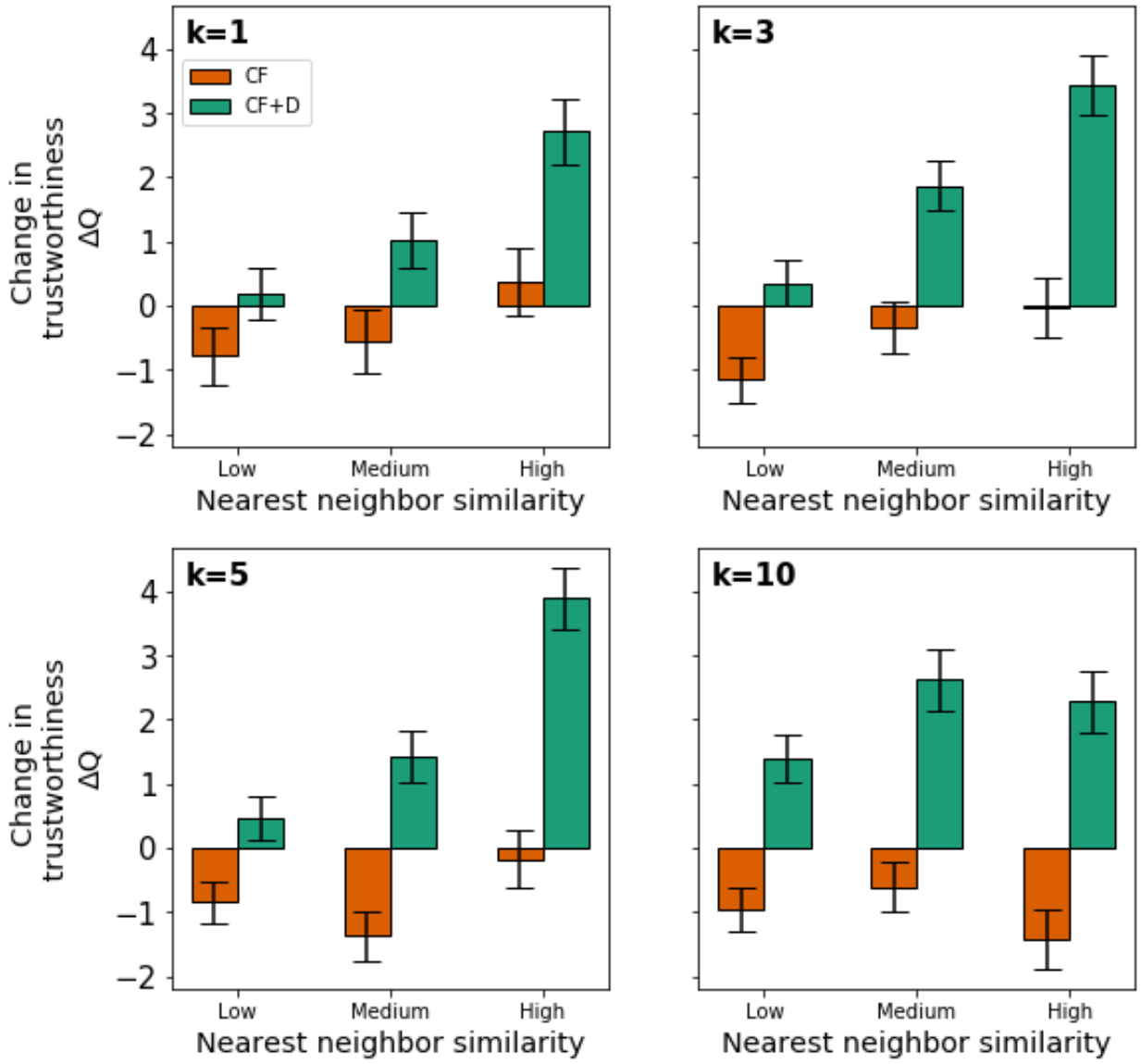
SUPPLEMENTARY FIGURE 12. Effect of CF and CF+D versus baseline by absolute slant of visited domains (terciles using scores from Bakshy et al.) and by length of ranked list k. Change in trustworthiness ΔQ based on scores from NewsGuard [3].



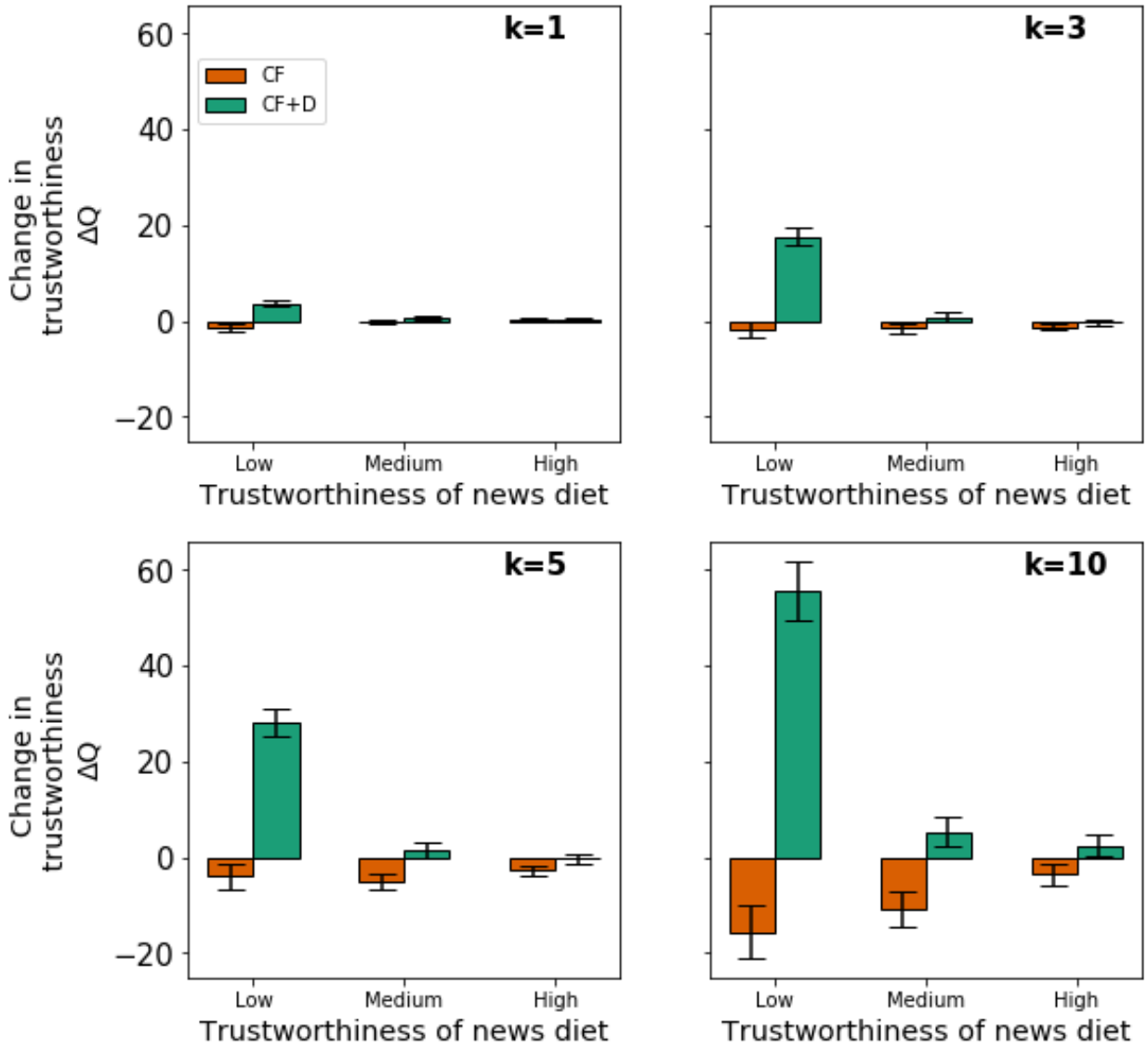
SUPPLEMENTARY FIGURE 13. Effect of CF and CF+D versus baseline by total online activity (TF-IDF-transformed pageviews; terciles) and by length of ranked list k. Change in trustworthiness ΔQ based on scores from NewsGuard [3].



SUPPLEMENTARY FIGURE 14. Effect of CF and CF+D versus baseline by distinct number of domains visited (terciles) and by length of ranked list k. Change in trustworthiness ΔQ based on scores from NewsGuard [3].



SUPPLEMENTARY FIGURE 15. Effect of CF and CF+D versus baseline by average user–user similarity with nearest $n = 10$ neighbors in training set (terciles) and by length of ranked list k. Change in trustworthiness ΔQ based on scores from NewsGuard [3].



SUPPLEMENTARY FIGURE 16. Effect of CF and CF+D versus baseline by baseline trustworthiness of domains visited by users (terciles) and by length of ranked list k. Change in trustworthiness ΔQ based on scores from NewsGuard [3].

SUPPLEMENTARY TABLE 5. Analysis of bipolar audiances

Metric	n	r	95% C.I.	two-sided p-value
Extremists (FE)	1,680	0.064	[0.02, 0.11]	0.0083
Bimodality (BC)	1,680	−0.097	[−0.14, −0.05]	$< 10^{-4}$
Harmonic Mean (FE, BC)	1,680	−0.053	[−0.10, −0.00]	0.0304

S6. BIPOLAR AUDIENCES

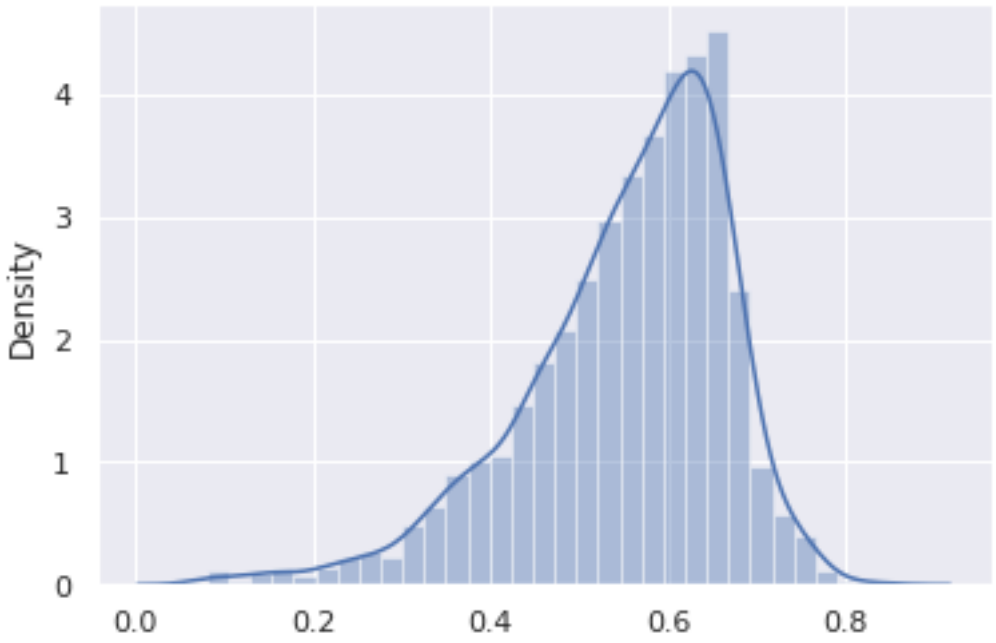
To capture the idea of an ideologically diverse audience, our analysis relies on the second moment of the distribution of audience partisan slant of the visitor of a site. However, two distributions may have the same mean and variance but different numbers of modes. This is a reasonable concern, since in the main text we show that, while keeping average partisanship constant, having more variance in partisanship will increase quality. To capture the idea of a bipolar audience we considered the following two metrics:

- The bimodality coefficient (BC), which is based on the skewness γ and kurtosis k of the distribution and is defined as $\frac{\gamma^2+1}{k}$;
- The fraction of extremists (FE), defined as $\frac{n_1+n_2+n_6+n_7}{\sum_{k=1}^{7} n_k}$, where n_k is the number of respondents who reported a slant equal to k (where $k = 1, \ldots, 7$).

The BC is a metric used in behavioral research to test for the presence of dual processes [2] that seems appropriate in this case. Values above 5/9 indicate the presence of two or more modes. The distribution of BC values in our data is displayed in Fig. 17, showing a prevalence of likely bimodal audiences in the YouGov data. The FE is our own attempt at capturing the idea that most of the probability mass should be located at the extremes of the distribution.

We repeated the analysis in the main text, but this time we computed three different partial correlation coefficients between quality and diversity: the first given the BC of the distribution of slants, the second given its FE, and the third given the harmonic mean of BC and REM. This last metric is high only if both FE and BC are large, and thus should discriminate between truly bipolar audiences and audiences that are merely skewed to one extreme but not both (high FE and low BC), or that are bimodal but not extreme (high BC and low FE). The results are summarized in Table 5.

We find a small negative correlation for the harmonic mean and for the BC alone, and a small positive correlation for FE alone. Taken together, these results suggest that audience bipolarity is



SUPPLEMENTARY FIGURE 17. Distribution of bimodality coefficients of the distributions of audience partisanship slants of the websites in the YouGov data. The solid line represents a kernel density estimate.

SUPPLEMENTARY TABLE 6. YouGov Pulse respondent data summary

Dates	Sample	Male	College	White	Age	Dem.	GOP	Domains	Pageviews
10/7–11/14/16	3,251	47.6%	29.2%	68.2%	58	37.2%	26.1%	158,706	26,715,631
10/25–11/21/17	2,100	47.6%	27.2%	69.1%	45	34.6%	25.0%	104,513	14,247,987
6/11–7/31/18	1,718	48.1%	30.2%	64.9%	54	38.4%	27.4%	108,953	15,212,281
7/12–8/2/18	2,000	48.8%	29.3%	65.2%	57	38.4%	26.0%	74,469	9,395,659
10/5–11/5/18	3,332	48.3%	29.0%	64.6%	55	39.5%	26.5%	98,850	19,288,382
11/12/18–1/16/19	4,907	48.7%	28.8%	64.1%	50	36.4%	26.8%	117,510	21,093,638
1/24–3/11/19	2,000	48.4%	29.6%	65.2%	55	34.9%	28.0%	113,700	27,482,462

Note: The participants for each data collection period were different. Some participants took part in multiple waves but overlap was small.

a much weaker signal of quality than partisan diversity as measured by the variance.

S7. YOUGOV PULSE RESPONDENT DATA SUMMARY

Supplementary Table 6 gives a demographic summary of respondents in the YouGov Pulse panel, broken down by collection period.

SUPPLEMENTARY TABLE 7. Table for Figure 7

k	Degrees of freedom	p-value	Effect size statistic (t)
1	853	$< 10^{-4}$	3.56
2	1004	$< 10^{-4}$	5.78
3	1059	$< 10^{-4}$	4.17
4	997	$< 10^{-4}$	2.32
5	920	0.503	0.89
6	824	0.054	1.45
7	752	0.238	1.23
8	678	0.683	−0.34
9	607	0.270	−0.56
10	525	0.022	−2.56
11	475	$< 10^{-4}$	−4.67
12	416	$< 10^{-4}$	−4.89
13	371	$< 10^{-4}$	−6.21
14	345	$< 10^{-4}$	−8.32
15	308	$< 10^{-4}$	−10.56
16	266	0.0244	−3.67
17	242	0.002	−5.89
18	210	0.005	−6.78
19	197	0.138	−1.45
20	183	0.261	−1.12
21	162	0.206	−1.36
22	144	0.031	−3.78
23	128	0.007	−5.34
24	117	0.001	−7.54
25	102	$< 10^{-4}$	−8.68
26	89	$< 10^{-4}$	−6.45
27	76	0.009	−4.79
28	67	0.166	−2.45

REFERENCES

[1] Bakshy, E., Messing, S., and Adamic, L. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239):1130–1132.

[2] Freeman, J. B. and Dale, R. (2013). Assessing bimodality to detect the presence of a dual cognitive process. *Behavior Research Methods*, 45(1):83–97.

[3] NewsGuard, Inc. (2020). Rating process and criteria. Retrieved from *Internet Archive*: https://web.archive.org/web/20200630151704/https://www.newsguardtech.com/ratings/

rating-process-criteria/.