CESifo WORKING PAPERS

11138
2024

May 2024

Mapping the Risk of Spreading Fake-News via Wisdom-of-the-Crowd & MrP

François t'Serstevens, Roberto Cerina, Giulia Piccillo

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 11138*

Mapping the Risk of Spreading Fake-News via Wisdom-of-the-Crowd & MrP

Abstract

The alleged liberal bias and high costs of fact-checker ratings (Nieminen & Rapeli, 2019) have prevented their ability to limit the spread of fake news. Wisdom-of-the-crowd-based approaches, recognized as a credible alternative to traditional fact-checking methods, have gained prominence for their independence from alleged biases, real-time availability and cost-effectiveness (Allen, Arechar, Pennycook, & Rand, 2021). Instead of fact-checking, this paper utilises a large-scale crowd-sourcing experiment on tweets related to US politics and COVID-19. The objectives of this paper are twofold. First, it develops a method to compute consensus-based post-accuracy indices that are representative of the broader crowds despite their initial non-representative reviewing sample. The computed metrics indicate that Democrat and Republican reviewers have a non-overlapping definition of fake news. Though less accurate than state-of-the-art models, the presented methods provide a deeply explicable, impartial estimate usable for automated content moderation. Second, using the aforementioned accuracy indices, this paper identifies the characteristics of fake news sharers and generates state-wide indices of fake news sharing in the United States. This paper's findings suggest the tweet author's political alignment did affect the likelihood of spreading fake news, with Republicans sharing more fake news than Democrats even if an equal number of Republicans and Democrats were in the reviewing sample. The model parameters are subsequently extended to make state-wide fake news-sharing estimates. The resulting state indices and their transparent methodology, provide policymakers with a tool to estimate where fake news policies are most needed.

JEL-Codes: C100, C900, D900.

Keywords: fake news, representative inference, content moderation, wisdom of the crowds.

*François t'Serstevens**
Department of Data Analytics and Digitilisation
Maastricht University / The Netherlands
f.tserstevens@maastrichtuniversity.nl

Roberto Cerina
Department of Media Studies
University of Amsterdam / The Netherlands
r.cerina@uva.nl

Giulia Piccillo
Department of Economics
Maastricht University / The Netherlands
g.piccillo@maastrichtuniversity.nl

*corresponding author

May 17, 2024

1 Introduction

The spread of fake news has emerged as a formidable challenge to the integrity of public discourse. To this day, most quantitative studies on fake news have relied on fact-checkers to objectively determine the veracity of claims. However, concerns about potential biases within fact-checking practices prompt a critical reevaluation of the method's validity (Nieminen & Rapeli, 2019; Poynter, 2019; Rich, Milden, & Wagner, 2020; Walker & Gottfried, 2019). Instead of fact-checking, this paper employs the wisdom of crowds to establish a ground truth free of political bias, it then uses these ratings to infer the characteristics of fake-news sharers. This study utilises a large-scale experiment on tweets posted in 2022 on American politics and the COVID-19 pandemic.

This paper's objectives are twofold: First, it aims to contribute to the methodological foundations of the wisdom of the crowd post-accuracy metrics. Four aggregation methods and two estimation methods are presented. The aggregation methods were designed to weigh the accuracy ratings provided by survey participants in a manner that reflects the broader sentiments of the crowd, either as a whole or within specific segments (e.g., an accuracy score representative of the average Republican or the entire population) Estimation methods, either naive or model-based refer to the input used for these aggregations. Naive estimation took the accuracy scores as-is from the sample. Model-based estimation first estimated the predisposition of each socio-demographic stratum towards a tweet and its topic, then used the fitted values of the model as the input accuracy. Model-based estimates aimed to provide a representative estimate of the wisdom of the crowd. Given the non-random sampling, naive estimates may not accurately represent the crowd's sentiment. Naive and model metrics were highly correlated across all aggregation methods. The deeply explainable and representative accuracy metrics presented in this paper contribute to the

groundwork for automated content moderation on social media.

Second, this paper aims to infer the characteristics of fake news sharers across the United States. To do so, it leverages a multilevel regression with post-stratification combined with the previously calculated aggregated accuracy scores and the data-rich social media environment. The model is subsequently extended to generate state-wide indices of fake news spread in the United States. The state indices provide policymakers with a tool to evaluate where fake news sharing is most prominent. They aim to bridge the gap between policies which are often taken at a regional level and fake news which is often pictured as an online phenomenon.

Consistent with fact-checker-based research, Republicans were found to share more fake news than their Democrat counterparts (Guess, Nagler, & Tucker, 2019). This trend persisted even when an equal number of Republicans and Democrats were sampled. Across all accuracy metrics, the estimated state-wide fake news sharing indices maintain consistent rankings, with states like Indiana and DC consistently showing a higher and lower propensity for sharing fake news respectively.

The paper is structured as follows, section 2 briefly summarizes prior research on fake news and representative inference. Section 3 describes the survey setup, detailing the collection of data from crowd workers and the integration of external datasets. Section 4 details the calculation and use cases of seven different accuracy metrics. Finally, section 5 builds upon the calculated tweet-accuracy metrics to estimate the spread of fake news across the United States.

2 Fake-News Detection and Sharing on Social Media

This section provides a concise review of the academic literature regarding fake news. The authors begin by synthesizing prevalent methodologies utilized to assess the accuracy of news claims. Subsequently, they summarize influential factors contributing to the dissemination of fake news across social platforms. Lastly, the section outlines distinctions between online and offline data samples, addressing their potential impacts on the interpretation of fake news phenomena.

2.1 Fake News Definition and Identification

In academic literature, fake news is often defined as "fabricated information that mimics news media content in form but not in organizational process or intent" (Lazer et al., 2018). This study's wisdom of the crowds-based definition departs from (Lazer et al., 2018)'s definition, rather it defined fake news as news that falls outside of the consensus. Regardless of the granularity of the definition, an assessment must be made to assess the veracity of the claim or outlet. This assessment can be performed through three broad techniques.

Expert Review. Journalistic fact-checking has been widely adopted by social media with Facebook, Instagram and X formerly Twitter all implementing a form of fact-checker labels and verifications over the last years (Instagram, 2019; Meta, 2021; X, 2023b). Besides social media platforms, academics have also resorted to fact-checkers to identify fake news reliably (Allcott & Gentzkow, 2017; Mena, 2020; Pennycook & Rand, 2019a; Vosoughi, Roy, & Aral, 2018). Republicans and Democrats are known to perceive fact-checkers differently with Republicans contesting the validity of fact-checkers ratings (Nieminen & Rapeli, 2019). The alleged bias of fact-checkers partly reduces

the credibility of the findings for Republicans who were consistently found as culprits of the spread of fake news (Guess et al., 2019).

Crowd-sourced Review. Scholars and practitioners have recently explored the wisdom of the crowds to identify fake news (Pennycook & Rand, 2019a). This crowd-based definition of the truth comes with two main advantages over fact-checking, (1) a universal definition of the truth and, (2) superior scalability. Provided an initially representative sample, crowd-sourced assessments are significantly correlated to fact-checkers' ratings, even when a limited number of reviews per claim are available (Allen et al., 2021). The difficulty of crowd-sourced ratings lies in the representativeness of the initial reviewer sample, i.e. a Democrat-leaning pool of laypeople would favour a Liberal narrative over a Republican one.

Computational Methods. Computational models are enabled by the data-rich environment of social media. Computational methods mainly leverage available information on (1) content features such as the source and headline and, (2) social context data such as the network, the user profile and others and, (3) linguistic features (Shu, Sliva, Wang, Tang, & Liu, 2017). Computational models have shown promise to assess fake news automatically with a handful of models coming close to human assessment (Agarwal, Sultana, Malhotra, & Sarkar, 2019; Hussain, Hasan, Rahman, Protim, & Al Hasan, 2020). Recent advances made in AI are likely to further enhance the accuracy scores of these models, possibly perpetrating the biases of the initial crowd. It is worth noting that all machine learning models require an exogenous definition of the truth, either crowdsourced or from expert review.

Notably, the methods outlined above are not mutually exclusive and have been combined by researchers and social media platforms. For instance, X's recently implemented community notes leverage both a machine learning algorithm and user input (X, 2023a). Likewise, user flags have been used to select review-worthy claims in an effective manner (Tschiatschek, Singla, Gomez Rodriguez, Merchant, & Krause, 2018).

2.2 Online Sharing Dynamics

Across social media, fake news consumption is highly heterogeneous, with a handful of individuals primarily responsible for posting and sharing fake news (Grinberg, Joseph, Friedland, Swire-Thompson, & Lazer, 2019; Guess et al., 2019). The echo chamber structures and bots have often been accused of promoting the spread of fake news (DiFonzo, 2011). While their presence is indisputable, substantial evidence establishing their definitive role in its dissemination remains elusive (Cinelli, De Francisci Morales, Galeazzi, Quattrociocchi, & Starnini, 2021; Guess, Nyhan, Lyons, & Reifler, 2018; Vosoughi et al., 2018). Rather, behavioural elements, such as analytical thinking and perceived accuracy or "Fear of Missing Out" have been identified as consistent factors in the sharing process (Pennycook & Rand, 2019b; Talwar, Dhir, Kaur, Zafar, & Alrasheedy, 2019; t'Serstevens, Piccillo, & Grigoriev, 2022).

Although behavioural elements are central to fake news sharing, they are typically not observed on social media and require a deeper analysis (Osmundsen, Bor, Vahlstrup, Bechmann, & Petersen, 2021). The academic literature therefore investigates the significance of demographic factors. In the United States 2016 presidential elections, Republicans in older age groups were consistently found to share more fake news than their Democrat counterparts (Guess et al., 2019; Grinberg et

al., 2019). A noteworthy caveat of these findings is the fact-checkers that lie at the centre of the veracity ratings of the claims. Unlike behavioural elements, age, gender, and political alignment can often be found in the information-rich social media environment and are thus included in this analysis.

3 Data Collection

This section covers the setup of the experiment in this study. It details how the survey respondents and the tweets they were shown were selected. The reviews were later used to determine the accuracy of tweets. The survey design, the gathered data (excluding private information) and the R, python and stan code used for the analyses are available online through GitHub (https://github.com/ftserstevens/mapping). The experiment was approved by an ethics committee review reference number ERCIC_348_15_04_2022.

3.1 Tweet Selection

A total of 247,725 tweets, posted from May 2022 to December 2022, were downloaded from the Twitter API using a list of COVID-19 related keywords[1]. U.S. Geotagged tweets were exclusively selected as reliable localisation data was necessary to enable post-stratification at the state level. Therefore, Twitter users who opted out of the geolocalisation feature were excluded from the analysis apriori. From this initial pool of tweets, a final set of 5513 was selected for review. Although this represents a considerable downsizing, reducing the tweet pool enabled the crowd-sourced approach.

[1] *Corona, COVID, COVID-19, Coronavirus, Facemasks, Vaccine*

The reasoning and mechanics of this selection process are as follows:

1. **Non-Relevant Tweets**. A majority of downloaded tweets did not contain any verifiable information that either laypeople or fact-checkers could review. Consider the following tweet: "Vaxxed and ready to tackle life with extra protection! Grateful for science and a brighter future.". It does not contain any verifiable information nor is it relevant to the societal debate on fake news. Most tweets found on X were akin to this example in that they did not contain any form of news or propaganda As such, they were irrelevant within the scope of this study. Twitter's built-in context annotations[2] were used to select tweets that related to both COVID-19 and political events. Though the filtering did not fully exclude non-informative tweets, it drastically reduced their prevalence in the sample.

2. **Effective Usage of Resources**. An equilibrium between the number of tweets in the review pool and the average amount of reviews per tweet needed to be determined to ensure efficient usage of the finite amount of reviews. On the one hand, the greater the number of tweets in our pool the greater the diversity of Twitter users in the sample. On the other hand, the fewer tweets in the final pool the more crowd-sourced reviews per tweet. Preceding studies have suggested that there were diminishing returns after 5 reviews per tweet (Allen et al., 2021).

3. **Optimal Number of Reviews per Tweet**. Given the importance of the first five reviews on a given tweet, the optimal number of reviews per tweet was estimated to maximise the number of tweets with more than 5 reviews. Given a finite number of available reviews ($\tilde{4}0.000$), an average of 6 to 7 reviews per tweet would maximize the number of effective tweets, i.e. tweets

[2]Twitter provides the option to download context annotations through its API. These context annotations are used by Twitter to create its trending topics. They denote whether a tweet is about one or multiple personalities, events, objects, etc. Section 3.2 details their use in this paper.

with more than 5 reviews. This translates into an initial sample of approximately 6500 to 5500 tweets. Consequently, tweets posted from May 2022 onwards were exclusively chosen. This time-bound filter and the aforementioned relevance filter resulted in a final pool of 5513 tweets. An average of 7.73 reviews per tweet were gathered from the survey respondents for all 5513 tweets. The number of reviews per tweet here is higher due to a surplus of survey participants.

3.2 Estimating X Users' Socio-Demographics

In addition to the data provided through the X API, author information was supplemented with estimates of author gender, age and political alignment. Author age and sex were determined through the m3 inference pipeline (Z. Wang et al., 2019). The m3 inference algorithm uses the author's profile picture and description to generate a probabilistic estimate of age category and gender. A majority of the sample (57%) was classified as 40+ years old by the algorithm. The remaining percentages are spread across the 18-30 and 30-40 categories equally with only a handful of minors in the sample (1%). Figure 1a depicts the proportions of voters within the sample and how they differ from the population.

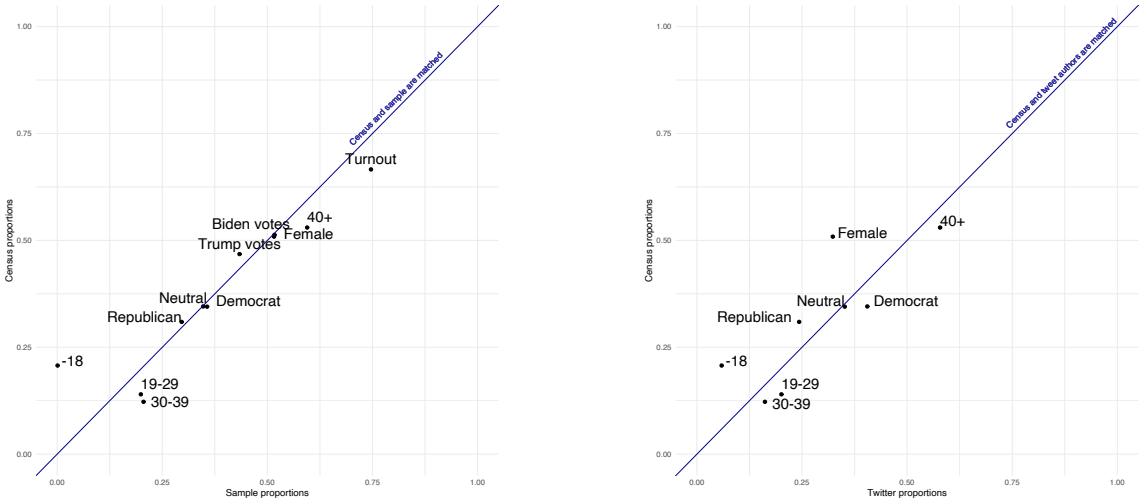
Political affiliation was estimated through the elite misinformation-exposure estimation tool (Mosleh & Rand, 2022). The tool identifies elite accounts within the following networks of X users. Elite accounts are typically well-known X profiles whose partisanship was scored by PolitiFact. The elite scores are aggregated per user to estimate the political affiliation of the user. Users above below or in between $-.25$ and .25 were classified as either Democrats, Republicans or Neutrals, users that did not follow any elite accounts, were also classified as Neutrals. Across all authors, 36% were

Democrat-leaning, 17% had neutral scores and 21% were Republican-leaning, the remaining 26% did not follow any elite accounts. The socio-demographic characteristics of the tweet authors and their differences with the census data are summarized in figure 1b.

Information on tweet content and topic was derived through X's native context annotations. X uses context annotations to classify tweets in one or multiple topics, for instance, the annotations are used to identify the trending topics on X. Only the 30 most mentioned context annotations were recorded. A given tweet could include several annotations at once, e.g. a tweet might address both "Governmental institutions" and "Joe Biden" simultaneously. Highly concurrent annotations (over 60% co-occurrence both ways) or present in all tweets were excluded.

3.3 Crowd-Workers Socio-Demographics

Between January and March 2023, we recruited 5154 U.S.-based participants from Lucid. From the initial pool of participants, 2196 participants passed the attention checks, and 166 participants did not complete the survey fully. A total of 2030 unique participants were selected in the final dataset. The survey demographics and their difference with the census data are summarized in figure 1. The survey was designed with quotas ensuring equal representation of both genders, representative age groups, and a maximum limit of 100 participants from each state. Every ZIP code in the US was represented at most once in the sample, participants from an already represented ZIP code were rejected. This feature was implemented to ensure a diverse sample population, it enforces a more geographically diverse sample suited to the subsequent implementation of MrP.



(a) Lucid Raters (b) X Users

Figure 1: The y-axis represents the census proportions and the x-axis represents the Sample and Tweet proportions. Any point on the red line represents a matching proportion. Points that are away from the red line are over- or under-represented.

3.4 Survey Protocol

The survey participants assessed the accuracy of 20 tweets. The tweets were selected randomly from the aforementioned curated pool and were shown one at a time. For each tweet, laypeople indicated their perceived accuracy using a 4-point accuracy scale akin to previous research: *Not at all accurate, Not very accurate, Somewhat accurate and Very accurate* (Pennycook & Rand, 2019a). As the analyses require a numerical quantification of the perceived accuracy, the aforementioned accuracy scale was converted to a numerical scale from 1 to 4, 1 being *Not accurate at all* and 4 being *Very accurate*. The first ten tweets shown to participants only included the tweet text corpus and the day on which it was posted. The subsequent 10 tweets included all commonly available profile information, i.e., Twitter profile picture, Twitter bio, localisation (if disclosed by the user), followers and following numbers, and tweet-specific social cues (retweets, likes, comments). The

decision to include tweets with and without author information is motivated by the need to assess the impact of additional contextual cues on perceived accuracy judgments. That is the author's profile picture affects the reviewer's judgement. This feature was not included as a parameter of the subsequent model.

Three attention checks were placed in the assessment phase. Participants who failed more than one were excluded from the survey. Attention checks prompted laypeople to select pre-defined accuracy and confidence ratings, e.g., "Very accurate" and "Very confident". The attention checks had the same format as the tweet accuracy rating tasks. For the first ten tweets, for which only the tweet text and date of posting were displayed, the tweet text was replaced by the attention check query. For the last ten tweets, where more information was displayed, in one-half of the cases, the author bio was changed by the attention query, in all other cases the tweet text was changed instead.

4 Aggregating Crowd Assessments

To estimate the wisdom of the crowd accuracy this paper introduces a total of eight accuracy metrics.Two aggregation methods are presented, Naive and Model-Based. Naive accuracy scores do not use participant information such as their political party, model-based estimates do. The necessity of the model-based estimates stems from the original possible non-representativeness of the tweet-reviewing sample. The model-based metrics aim to estimate the tweet rating as if a representative crowd reviewed it. For both Naive and model-based estimation four aggregation methods are presented: *Sample* based, the average estimate of the sample, (2) *Balanced*, where Democrats and Republicans are equally represented, (3) *Population* based, where census data is used to enforce representation of the entire population and, (3) *Partisan* based, which represents

the average opinion of a given political party. These accuracy metrics would later be used as input to estimate fake news spread in the United States. Table 1 summarizes the main accuracy metrics. This section details their aggregation process and assesses the similarity of these metrics.

Aggregation \ Estimation	**Naive**	**Model**
Sample	Observed assessment average	Fitted values average
Balanced	One Democrat for one Republican	Mean of Fitted Democrat & Republican personas
Population	Raked average (IPF)	Fitted census population
Partisan	Only Republicans or Democrat ratings	Fitted Republican or Democrat personas

Table 1: Operations are done at the tweet level, e.g. the average assessment for a given tweet.

4.1 Naive Aggregation

The first naive estimation method calculates the average accuracy score per tweet. In an ideal scenario with a sufficiently large group of reviewers for each tweet and a representative sample, the characteristics of the reviewers would accurately reflect those of the broader population for all tweets. In the experiment at hand, an average of 7.73 reviews per tweet were made reviews, it is unlikely that the reviewer pool of every tweet is indeed representative of the population. This measure is hereafter referred to as $Sample\ Accuracy^{Naive}$, its formal definition can be found in Equation 1. The appeal of the latter lies in the simplicity and straightforwardness of its calculation.

$$Sample\ Accuracy_t^{Naive} = \frac{1}{n}\sum_{p=1}^{n} A_{t,p} \tag{1}$$

Where A is the vector of assessments made by participant p on a tweet t.

The second form of naive estimation was the "Balanced" method. This method effectively ensures that an equal number of Democrat and Republican voters are considered for all tweets despite the random nature of the experiment. It is formally defined in equation 2. A bootstrapping procedure was implemented for participants in the over-represented political group. This involved resampling the reviews from the over-represented political party 10 times. For each resampling iteration, an equal number of reviews from both Democrats and Republicans were chosen. Specifically, the reviews from the over-represented political group were resampled for each iteration. This enabled the use of all the reviews within the over-represented category. This method aligns with the established practice in this research field, as indicated by previous work (Allen et al., 2021; Pennycook & Rand, 2019a). Though $BalancedAccuracy^{Naive}$ enforces a political balance, it is worth noting that both the independent- and non-voters were excluded from the aggregation altogether.

$$\begin{aligned} n_t &= \min(||A_t^{Dem}||, ||A_t^{Rep}||) \\ Balanced\ Accuracy_t^{Naive} &= \frac{1}{2n_t} \sum_{p=1}^{n_t} A_{t,p}^{Dem} + A_{t,p}^{Rep} \end{aligned} \tag{2}$$

Where A^{Dem} and A^{Rep} are the assessment vectors of Democrat and Republican voters in the 2020 presidential elections respectively.

The third form of naive estimation was the raked average of the assessments. The raking procedure, also known as iterative proportional fitting (IPF), adjusted the weights of survey participants to represent the population assumed from census data. While the Lucid survey quotas enforced

representation of age categories and gender, some socio-demographic variables, such as the vote in the 2020 presidential elections, were excluded from this procedure. Raking was employed to re-weight the sample and ensure its conformity with known population-level demographic marginal distributions. The raking procedure was done using the `anesrake` function from the American National Election Study (ANES) (DeBell & Krosnick, 2009) akin to the methods of other scholars (Deming & Stephan, 1940; Kolenikov, 2014).

The last form of aggregation exclusively utilizes the inputs of Biden or Trump voters in the 2020 presidential election. The partisan accuracy metrics serve as a benchmark against which other accuracy metrics are compared. The partisan-based metrics explain how the wisdom of the crowds assessments relates to political ideologies. For instance, there might be a tendency for the consensus to consistently lean towards Democratic ideologies due to their higher representation in the sample, potentially leading to a "tyranny of the majority". Equation 3 formally defines $PartisanAccuracy$.

$$\forall \mathrm{S} \in \{\text{Vote Biden, Vote Trump}\}$$
$$Partisan\ Accuracy_{t,S}^{Naive} = \frac{1}{n} \sum_{p \in \{\mathrm{S}\}}^{n} A_{t,p} \tag{3}$$

Where S denotes the subsets of Biden and Trump voters and $A_{t,p}$ are the assessments made by participants in a given subset S on a given tweet t.

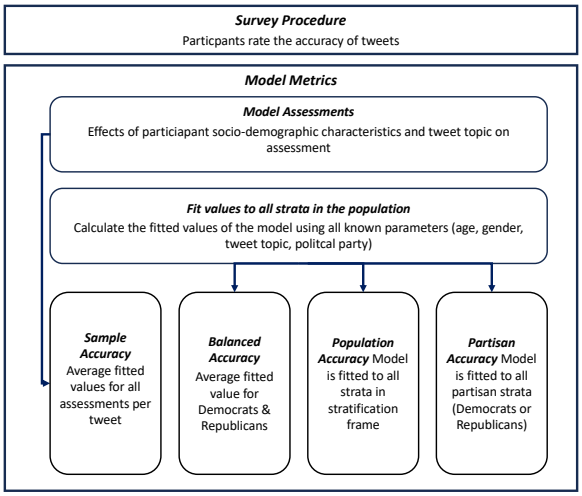


Figure 2: Summary of the estimation process of model-based metrics.

4.2 Model-based Aggregation

Unlike the naive approaches, model-based approaches included information from the laypeople in their estimations. A Bayesian multilevel ordinal-logistic regression was employed to estimate the predispositions of socio-demographic groups towards given tweet topics. It is denoted in equations 4 to 7. Figure 2 provides a brief overview of the aggregation process of model metrics.

$$A_i \sim \text{Ordinal}(\pi_1, \pi_2, \pi_3, \pi_4) \tag{4}$$

$$
\begin{aligned}
logit(\pi_i) = & \alpha^{tweet}_{a[i]} + \alpha^{context}_{b[i]} + \\
& \beta^{gender}_{c[i]} + \beta^{age}_{d[i]} + \beta^{state}_{e[i]} + \beta^{party}_{f[i]} + \\
& + \gamma^{context:party}_{b[i],f[i]}
\end{aligned} \tag{5}
$$

$$
\forall\, x \in \{\alpha, \beta, \gamma\} \qquad\qquad x^u \sim \mathcal{N}(0, \sigma^u) \tag{6}
$$

$$
\forall\, u \in \{\text{tweet, context, gender, } \ldots \mid \exists\, x^u\} \qquad\qquad \sigma^u \sim \mathcal{N}^+(0, 1) \tag{7}
$$

The regression model incorporates information from both the tweets and the participants. Participant characteristics such as *gender*, *age*, *state*, and *party*, were included as covariates. These captured the participants' inherent likelihood to rate any post as accurate or not. e.g. Young participants were less likely to believe in a post than older participants. *Context* refers to the context annotation of the tweet as explained in section 3.2. The *context* parameter is included to capture any inherent inclination of the sample towards a given topic. It interacted with the *party* covariate such that the interactions capture the predispositions of a political group on a given annotation, e.g. the average accuracy rating of Republicans on tweets mentioning the White House. Although the $context - party$ interaction could be extended to all other covariates (e.g. age-context interaction), this would drastically increase the required statistical and computational power. The $context - party$ interaction was chosen due to the highly political nature of fake news and the selected tweets. The remaining *tweet* parameter then represents the tweet's accuracy ceteris paribus.

As with the naive accuracy metrics, *Sample*, *Balanced* and *Population* accuracy scores were calculated for all tweets. Using the estimated model parameters, their computation is as follows:

1. **Sample Accuracy**. Model-based *Sample Accuracy*^{Model} scores were calculated by taking the average of the fitted values. That is, instead of using the assessment of the participants as input, their predicted assessments were used.

2. **Balanced Accuracy**. The model equivalent of *Balanced Accuracy*^{Model} was generated by computing the fitted values for all Republican and Democrat strata for all tweets. These fitted values were then averaged, with the size of each stratum within each political class weighted according to their representation in the census. This process yielded an estimated average Republican rating and an average Democrat rating, which were then averaged again to generate the model-based *Balanced Accuracy*^{Model}. As with its naive homologue, equal weights were given to Republicans and Democrats and non-voters and independents were fully excluded.

3. **Population Accuracy**. To estimate *Population Accuracy*^{Model}, the fitted values for every tweet for all strata present in the census were calculated. These fitted values were then averaged and weighted by the importance of the population. This process provides an overall estimation of the accuracy across the entire population, a synthetic version of the wisdom of the crowds.

4. **Partisan Accuracy.** As for *Balanced Accuracy*^{Model}, *Partisan Accuracy*^{Model} accuracy score used the fitted values of Republicans and Democrats in the stratification frame. Unlike *Balanced Accuracy*^{Model}, which combined the fitted values of Democrats and Republicans, *Partisan Accuracy*^{Model} only averaged either Republicans or Democrats fitted values for its

prediction.

The value of model metrics, particularly $Balanced\,Accuracy^{Model}$ and $Population\,Accuracy^{\,Model}$, lies in their ability to capture the perspectives of all strata within the stratification frame. Unlike naive metrics, which depend solely on a limited number of reviews collected through surveys per tweet, model metrics (except $Sample\ Accuracy^{model}$) utilize the predicted values across the entire population. This broader approach ensures a more comprehensive representation of diverse viewpoints, enhancing the reliability and robustness of the evaluation process.

4.3 Model Evaluation

The model was implemented and processed using the Stan programming language (Gelman, Lee, & Guo, 2015). It utilized a binary tree depth of 10 and was run with a total of 7 chains of 500 post-warm-up iterations. All chains and parameters of the model converged ($\hat{R} \leq 1.05$, $N_{eff}/N \geq 10\%$, $N_{eff} \geq 400$, $se_{mean}/\sigma < 10\%$). The computed model, and visualizations of the posterior distributions of the parameters, can be found in the supplementary material, the necessary code to run the model is also provided but requires several days of computation on a household computer[3].

The accuracy assessments were mostly unaffected by demographic variables, as evidenced by credible intervals consistently centred around 0. Despite no discernible trends across various categories, a handful of specific levels exhibited significant deviations from the baseline, this happened especially for less represented levels, e.g. doctoral levels of education.

Notably, the main effect of the reviewer's political affiliation showed no impact on the assess-

[3]Intel i5 with 8GB of RAM

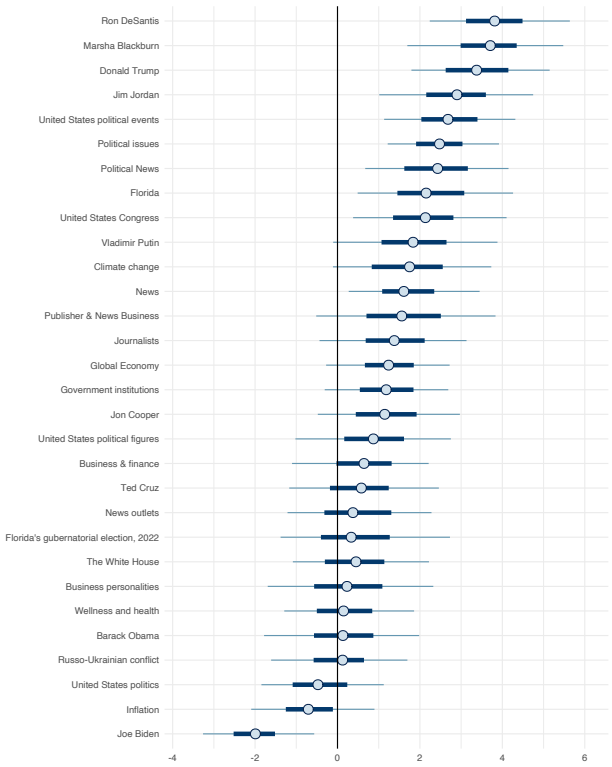


Figure 3: Difference between Democrat and Republican predispositions towards the selected context annotations. Negative and positive values respectively indicate that Republicans and Democrats are more likely to rate the topic accurately. Over a third of the tweets did not have an attributed topic.

ment. That is, Republicans, Democrats, and Neutrals provided similar ratings across the political spectrum. i.e. None of the political alignments being particularly sceptical or gullible. However, interaction effects between political affiliation and context were statistically significant, indicating partisans differed in their understanding of 'truth' for some context annotation. This suggests that both the overall crowd and its political subgroups displayed predispositions towards specific topics Figure 3 visually illustrates the difference in the logged odds of Democrats and Republicans for the chosen annotations.

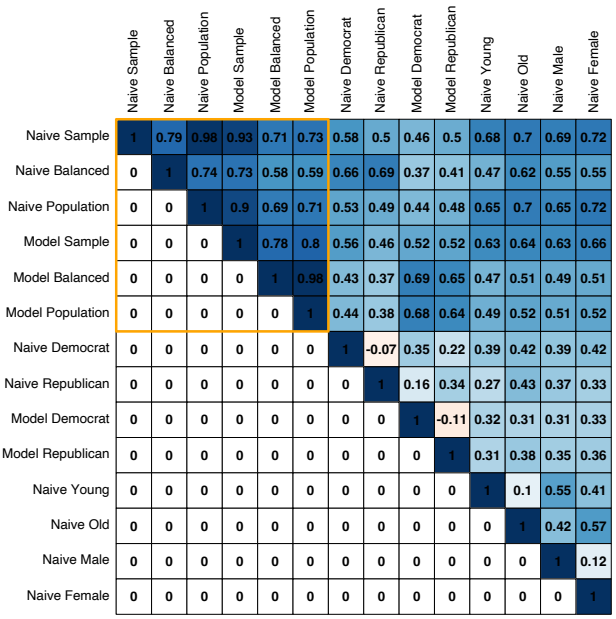


Figure 4: The upper half of the matrix indicates the correlation coefficients; the lower half the p-values of the respective correlations. The yellow rectangle highlights the wisdom of the crowd metrics, non-highlighted metrics are unrepresentative of the crowd by design.

4.4 Accuracy Metric Comparison

The correlations across all accuracy estimates are summarized in Figure 4. Estimating the correlations across all metrics serves multiple purposes. First, it validates the model-based estimates. The high correlation between the model and Naive *Sample Accuracy* indicates that the model can replicate the assessments made in the survey to a significant extent. Second, the correlations offer an indication of the consistency and agreement between different evaluation approaches. If two metrics are highly correlated, it suggests that they tend to yield similar results. All metrics exhibited significant correlations except for the *Partisan Accuracy* metrics. All naive and model-based correlations were significant, indicating consistency across both methods.

Establishing a best-in-class accuracy metric poses significant challenges. While journalistic fact-checking can help identify which metric aligns most closely with expert ratings, it may not definitively establish the optimal expression of the wisdom of the crowds. Credibly determining the best metric would require significantly larger sample sizes per tweet to ensure convergence of crowd wisdom across all instances. However, the high correlations among all metrics suggest that various measures ultimately converge to a stable estimation of crowd wisdom. The ensemble of provided metrics would likely capture the essence of a best-in-class metric.

It is worth noting that all *Partisan Accuracy* metrics, both naive and model-based, were the only ones to be negatively correlated.

The partisan metrics express the underlying polarization around fake news sharing online. Though Democrats and Republicans both agree that fake news is a problem within the political landscape, they have very different definitions of what fake news is. Table 2 displays 4 example tweets along with their accuracy scores, it exemplifies the differences and similarities in Democrat and Republican ratings along with the inner workings of the $SampleAccuracy^{Naive}$ and $ModelAccuracy^{Model}$ scores.

In contrast to the partisan metrics, analyses of other dichotomous classifications (i.e., young and older reviewers, or female and male reviewers) reveal statistically significant correlations within their respective categories. This underscores the distinctive nature of partisan perspectives towards fake news, while age and gender may influence perceptions there is a common understanding or agreement across these categories regarding the nature of fake news. When it comes to political affiliation, the situation is markedly different. Despite shared recognition of fake news as a problem within the political landscape, Democrats and Republicans have a significantly different interpretation of what fake news is. This lack of consensus across political groups highlights the

Tweet Text	Accuracy Metrics	
Vaccines have been saving us for centuries so why stop being vaccinated now?! we took vaccines for chickenpox measles polio! Without these vaccines millions would have died! What's any different than this from Covid!? And those that don't take it make others vulnerable!	**Naive Sample**	3.21
	Model Population	3.72
	Only Democrats	4.00
	Only Republicans	2.57
That's for 2022, general gas prices Pre COVID where around $2 to $2.50 that's an increase of 61% when looking at 3.71. Get out inflation is killing us and they keep adding more dollars with no productivity back up in the economy.	**Naive Sample**	3.25
	Model Population	3.60
	Only Democrats	4.00
	Only Republicans	4.00
In case anyone has forgotten, we have rampant crime, 9% inflation, blown-up retirement savings, an open border with millions pouring across it illegally, soaring interest rates, COVID-19 mandates and government assaults on moral decency and biology. No joke, man.	**Naive Sample**	2.82
	Model Population	3.46
	Only Democrats	1.67
	Only Republicans	4.00
The state of Florida needs experience and steady hands to lead the recovery after the devastating hurricane and the Covid 19 pandemic. So many lives have been lost because of an inept governor in Ron Desantis and the destruction is unspeakable. We need Charlie Crist and Val Demm.	**Naive Sample**	2.30
	Model Population	2.10
	Only Democrats	3.33
	Only Republicans	1.33

Table 2: Mentions (e.g. @WhiteHouse) were removed from the tweet for readability.

deep ideological polarization that characterizes discussions around fake news. It lies at the heart of addressing misinformation polarization online.

5 Mapping 'Fake News' Sharing Risk

The second objective of this paper was to leverage the wisdom of the crowd ratings, as calculated in section 4, to infer who shares fake news and where is shared in the United States. To do so the paper leverages a Multilevel Regression with Post-stratification (MrP). The adoption of the MrP approach in this study is motivated by the need to address disparities between online and real-life

samples (W. Wang, Rothschild, Goel, & Gelman, 2015; Mellon & Prosser, 2017; Mislove, Lehmann, Ahn, Onnela, & Rosenquist, 2011). It enables the representative estimation of state-level indices of fake news sharing.

Populations on social media are known not to be representative of real-life populations. Specifically online samples are found to be younger, more educated and liberal-leaning (Mellon & Prosser, 2017). Though social platforms provide rich data, any inferences must consider the nature of their sample if they are to be generalized to the broader population.

For small area-level estimation of public opinion, Multilevel regressions with Post-stratification (MrP) have been commonly implemented to mitigate the unrepresentativeness of the data. Notably, Wang et al. (2015) implemented MrP by using a heavily biased Xbox-gaming sample to make country-wide representative polling inferences. Since its introduction the MrP framework has been widely adopted even outside of political forecasts, e.g. to predict anti-migrant sentiment and extremism recruitment (Butz & Kehrberg, 2016; Cerina & Duch, 2021). In the context of social media, MrP offers a means to account for the differences between online and real-life environments.

In summary, all six different accuracy metrics yield a similar conclusion despite their inherent philosophical differences. That is, (1) a handful of states consistently share more or less fake news than others, and (2) Republicans are more likely to participate in this spread. The methodology of this paper is free of any form of judgment from both the researchers and the journalist.

5.1 Model Parameters

Tweets with average accuracy ratings at or below the 10^{th} percentile i.e. the least accurate tweets, were designated as fake news, whilst the remaining 90% of tweets were categorized as generic news. The independent variables of the MrP featured both author- and state-level information. Author gender, age and political affiliations were gathered through the use of external tools as outlined in section 3.2. The states were inferred through the geo-tags provided by Twitter. The percentage of white individuals and the educational attainment in the state were taken from census data (Ruggles et al., 2023).

For all models except for the naive balanced approach, all 5513 tweets were used for an average of 7.74 reviews per tweet. The naive balanced approach used only 4710 tweets with an average of 3.55 reviews. This difference was caused by the sampling strategy, i.e. one Democrat needed to be sampled for every Republican and vice versa. The model and the post-stratification procedure are detailed in equations 8 to 14.

$$Model\ Accuracy_i \sim Bernouilli(\theta_i) \tag{8}$$

$$\begin{aligned} logit(\theta_i) =& \alpha_0 + \beta_{a[i]}^{gender} + \beta_{b[i]}^{age} + \beta_{c[i]}^{party} + \beta_{d[i]}^{state} \\ &+ \gamma^{white\%} X_1 + \gamma^{education} X_2 + + \gamma^{density} X_3 \end{aligned} \tag{9}$$

$$\alpha \sim \mathcal{N}(0, 1) \tag{10}$$

$$\forall\, u \in \{\text{gender, age, party}\} \qquad \beta^u \sim \mathcal{N}(0, \sigma^u) \tag{11}$$

$$\sigma^u \sim \mathcal{N}^+(0, 1) \tag{12}$$

$$\forall\, v \in \{\text{white\%, education, density}\} \quad \gamma^v \sim \mathcal{N}(0, 1) \tag{13}$$

$$\hat{\theta}_s^{PS} = \frac{\sum_{j \in J_s} N_j \theta_j}{\sum_{j \in J_s} N_j} \tag{14}$$

Where j is a cell (or a stratum) in the stratification frame, s is a state, and $\hat{\theta}_j$ is the estimated probability of sharing fake news for cell j. The multi-index notation of post stratified estimates, $\hat{\theta}_s^{PS}$, signifies the probability of state s to share a tweet classified as fake news, i.e. rated within the first ten accuracy percentiles. The model was implemented and processed using the Stan programming language (Gelman et al., 2015). A total of *10* chains with 1250 post-warm-up iterations were run with a binary tree depth of 15. All chains were initialized with random parameters with generic weakly informative priors for all parameters and hyperparameters.

The variables were stratified to census data and political surveys gathered through IPUMS USA and the American National Election Studies (ANES) (Ruggles et al., 2023; American National Election Studies, 2021). As the joint distributions of voter affiliation and demographic information were unavailable, the MrP with adjusted synthetic joint distributions (MrsP) procedure was used to build the stratification frame (Leemann & Wasserfallen, 2017).

5.2 Fake News Sharing Estimates

All accuracy metrics except *PartisanAccuracy* yielded similar coefficient estimations, a summary is provided in the Supplementary materials, and their fully computed form is available online on GitHub. All models fully converged with chains mixing and low autocorrelation ($\hat{R} \leq 1.05$, $N_{eff}/N \geq 10\%$, $N_{eff} \geq 400$, $se_{mean}/\sigma \leq 10\%$). Given the six accuracy metrics, naive *SampleAccuracy* naive *BalancedAccuracy*, and model-based *PopulationAccuracy* were used as primary benchmarks. Specifically, *BalancedAccuracy* was chosen because it replicates existing designs in the field (Allen et al., 2021; Pennycook & Rand, 2019a), *SampleAccuracy* due to its simplicity and transparency and, model-based *PopulationAccuracy* because of its novelty and efficient use of data. All models are available online and yielded similar conclusions.

The model intercepts had notably low logged odds (Naive Sample = -2.01, Naive Balanced = -1.97, Model Population = -2.02). This was explained by the model's dependent variable, which captured the least accurate 10% of tweets in the sample. The average estimate of the likelihood should therefore be at or close to 10%.

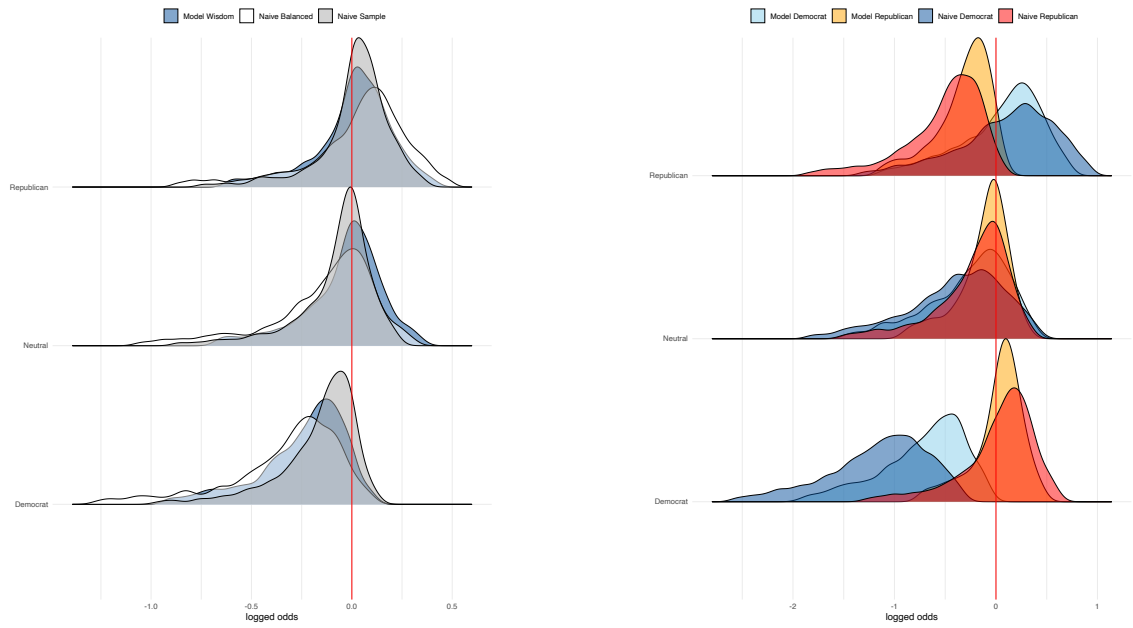
State-level estimates of population density, white percentage, and educational attainment were centred around 0 in all models. Similarly, author user gender and age estimates exhibited insignificant impact on fake news sharing, with credible intervals also centred around 0 across all models. However, the political affiliation of Twitter users was strongly correlated with fake news sharing propensity. Republicans consistently demonstrated a higher likelihood of sharing fake news compared to Democrats and Independents across all models. This trend persisted even when Democrats and Republicans were given equal weights, as observed in the Balanced metrics, indicating a greater responsibility for fake news sharing among Republicans. Posterior distributions for the selected

models are depicted in Figure 5a.

This political divide also manifested when using *PartisanAccuracy* metrics. When using Democrat partisan metrics, both model and naive-based, Republicans are the primary sharers of fake news, the opposite is true when Republican reviewers are selected exclusively. Figure 5b displays the posterior distributions of the author party parameter using the *PartisanAccuracy* metrics.

Notably, using the model-based definitions of truths typically has a negligible effect over the naive definitions of truths, except for the impact on the sharing coefficient of the democrats. The latter remains negative indicating Democrats are less likely to share fake news under their definition — but it is substantially closer to the average (i.e. the Model penalizes Democrats more than its naive equivalent). This is partly explainable by the model's 'regression to the mean' effect which impacts the overall variance of the dependent variable in this case and points to systematic differences between the demographic typing of Democrats in the sample of raters and Democrats in the population as a whole. This emphasises the importance of representativeness, here addressed through MrP, in the context of wisdom-of-the-crowd-driven content moderation.

Given the polarizing nature of these findings, it is essential to contextualize and consider their implications carefully. While the findings suggest that Republicans are more likely to share fake news on average, they do not imply that Democrats are immune to fake news sharing or that all Republicans actively engage in it. Previous research has outlined that a small number of individuals are primarily responsible for posting and sharing fake news (Grinberg et al., 2019; Guess et al., 2019). Such individuals can likely be found in Democrat and Republican communities alike, the analysis suggests that they are predominantly, not exclusively, Republican.



(a) Balanced and Net

(b) Democrats and Republicans

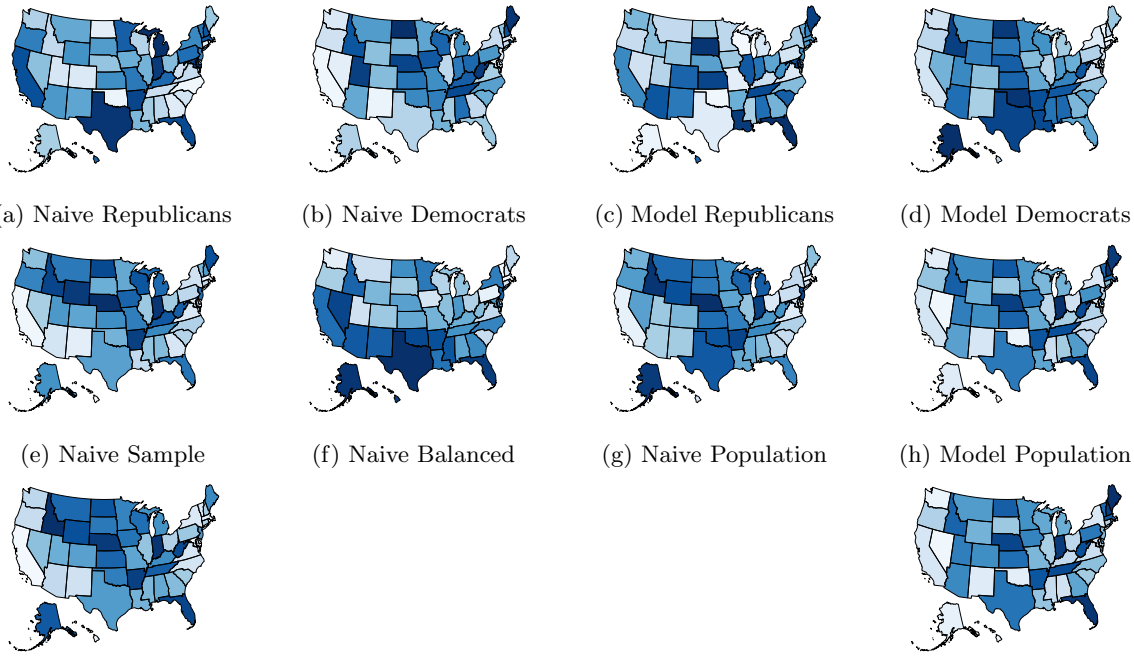
Figure 5: Posterior distribution of the party parameters for fake news sharing estimation. Negative values indicate a reduced likelihood of posting fake news.

5.3 State-Level Estimates

Using the estimated parameters, Bayesian post-stratified estimates were computed for all states. Figure 6 displays the median estimates of all accuracy metrics across the US states, the supplementary materials include the complete posterior distributions of all states and the output of all models. The high correlations between the presented models yielded similar state estimates across all models but the naive $BalancedAccuracy$ model. This difference in state estimates would be explained by the difference in sampling procedures. Whereas all other models used the full set of reviews for an average of 7.73 per tweet, the naive balanced procedure had an average of 3.55 reviews per tweet.

Figure 7 summarizes the average position of all states across all models excluding the $PartisanAccuracy$ models and the $BalancedAccuracy^{Naive}$ models. The former accuracy scores were excluded because of their inherent biases, the latter because of the low reviews per tweet. The models indicate that certain states were consistently ranked as high and low fake news spreaders. Middling states tended to have more uncertainty in their rankings across models. Future research could investigate the reasons behind the phenomena. Unsurprisingly the state estimates replicate the bias of Republicans towards fake news with most consistent fake news spreaders being predominantly Republican in the 2020 elections. However, this was not always the case, for instance, Florida was consistently ranked as a high fake-news-spreading state despite being a swing state in the latest presidential election.

The negative correlation between Democrat- and Republican-only ratings was visually exemplified in figures 6d, and 6a. Using both $PartisanAccuracy$ metrics, the state-level estimates are seemingly inverted. For instance, Texas transitions from being perceived as a low fake-news spreading state when considering Republican ratings, to being among the top fake-news spreaders based on



(a) Naive Republicans (b) Naive Democrats (c) Model Republicans (d) Model Democrats

(e) Naive Sample (f) Naive Balanced (g) Naive Population (h) Model Population

(i) Model Sample (j) Model Balanced

Figure 6: State-wide indices of fake news sharing. A darker colouring indicates that a state is likelier to share fake news.

Democrat-only ratings. The phenomenon is also observable in highly Democratic states such as New York. This inversion highlights the significant influence of partisan perspectives on the perception of fake news and the importance of a balanced sample when using the wisdom of the crowds in online samples.

6 Discussion

In the ever-evolving landscape of information dissemination, this paper sought to make two main contributions to the ongoing discourse on fake news. (1) It aimed to refine the methodological

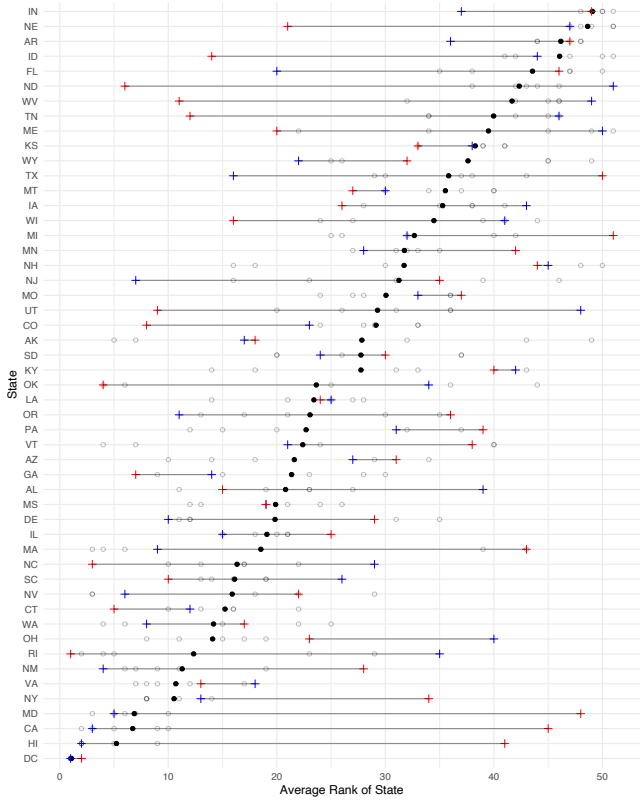


Figure 7: The black dot represents the average estimates across models. Light grey dots indicate the estimates of specific models. Red and blue crosses represent the partisan estimates for that state.

foundations of the wisdom of the crowd by proposing various ways to aggregate the ratings of laypeople. (2) It sought to identify the characteristics of fake news sharers and generate state-wide indices of fake news sharing in the United States.

First, this study finds that a relatively simple statistical model replicated the wisdom of the crowds in a synthetic manner, with model-based estimates being closely correlated with their naive equivalent. The subsequent analysis of the accuracy metrics revealed that the definition of fake news heavily depends on the reviewing crowd in a wisdom of the cross setting. Specifically, there was no overlap between accuracy metrics provided by exclusively Democrat and Republican reviewers. On the other hand, some common agreement could be found across genders and age categories.

Second, this identified that among all used socio-demographic variables the political affiliation of the tweet authors was the most notable predictor of fake news sharing. All wisdom of the estimates consistently identified Republicans as fake news sharers, even if an equal amount of Republican and Democrat ratings were used.

Lastly, the state-wide fake news sharing indices provided a comprehensive and granular understanding of fake news dissemination in the United States. While slight variations in state estimates were observed based on the used accuracy metrics, some states were consistently ranked as the high or low-spreading states with more uncertainty for middling states.

6.1 Wisdom of the Crowds and Fact-Checking for Online Content Moderation

The foundation of any quantitative study on the propagation of (misleading) information hinges on establishing accuracy or veracity metrics. To achieve this, researchers and social media platforms have often utilized fact-checkers to determine the veracity of claims objectively (Pennycook & Rand, 2021; Lazer et al., 2018; Vosoughi et al., 2018; Meta, 2021). However, concerns have arisen regarding potential biases within fact-checking practices, prompting a reevaluation of the method's validity (Nieminen & Rapeli, 2019).

This paper's underlying hypothesis has been that crowd-sourced accuracy metrics could be used to moderate online discussions. Of course, using the crowd to define the truth raises a moral question - can a crowd of laypeople effectively rate the veracity of a tweet correctly? Collective misconceptions have long existed (Miller & Prentice, 1994) and are likely to affect the wisdom of the crowds as well, but fact-checkers are not immune to mistakes.

Beyond an indication of veracity, a crowdsourced metric also indicates the polarization surrounding a post or statement. In the realm of online discussions and echo chambers (Cinelli et al., 2021), wisdom of the crowds ratings could provide valuable insights into the divergence or convergence of public opinion on any given post. Fake news debates often go hand in hand with political polarization discussions, the latent goal of any metric is to create common ground for discussion, with the hope of safeguarding a democratic process. Where fact-checker ratings would be rejected by default because of their partly subjective process, the wisdom of the crowd metrics are exempt from this fallacy.

6.2 Methodological Concerns for the Wisdom of the Crowds

While the wisdom of the crowds-based metrics could serve as a useful addition or alternative to the fact-checker ratings, it is important to consider the possible pitfalls of this method when considering their implementation. Beyond the moral concerns of using the crowds as arbiters of truth, any critique of wisdom of the crowds ratings would hinge on the reliability of such metrics to accurately reflect the broader population. This study's findings suggested that Republicans and Democrats had non-overlapping definitions of fake news, to a lesser extent this was also the case for males and females and young and old reviewers. It is unlikely that any crowdsourced metric would be universally accepted if any political or demographic category is over-represented, i.e. if more Democrats are used in the ratings, said ratings would likely face allegations of bias as fact-checkers have.

Though this may seem like a simplistic concern at first, the issue gains traction when considering the recommendation algorithms of social media. These algorithms imply that the audience viewing content is not randomly sampled and thus not representative of the broader crowd. If the viewing crowd is used as the basis for the metrics, a weighting of the ratings likely needs to be implemented such that the metric reflects the broader population instead. The apparent solution to this issue would involve setting up clinical trials where each post is reviewed fairly, though the sheer volume of posts on social media makes this solution hardly scalable.

To address the non-representativeness of the reviewing sample, this paper and other scholars have proposed various methods (Tschiatschek et al., 2018). Their findings suggest that algorithmic processes can effectively adjust for the initial non-representativeness. While it's likely that a machine learning model that uses more features and a more complex algorithm would outperform the met-

rics detailed in this paper. Yet, incorporating any algorithmic revision will likely raise suspicions among the intended end users. If that is indeed the case then implementation of the crowdsourced ratings must also consider the transparency of their model to avoid allegations of bias. In contrast to computationally complex models, this paper's methodology offers a relatively simple and transparent solution. The required data are openly accessible online, notably, any differences from the reviewing crowd average can be explained through the model parameters.

While the methodology outlined in this paper offers a transparent and straightforward approach, it is not devoid of limitations. Tweets about the COVID-19 pandemic and its handling by United States politics in 2022 were specifically considered. Expanding the model to encompass a broader spectrum of topics is feasible with relative ease, more reviews need to be gathered on a more varied sample of social media posts. Computational power should not be an issue given the relative simplicity of the model. More significantly, reviewers' opinions on various topics may have evolved over time. What was deemed accurate previously might now be perceived as false, and vice versa. If a similar model were to be employed in practice, the review-gathering process would need to occur continuously over time rather than at a singular point in time.

6.3 Fake news Spread Online and Offline

The spread of fake news is often linked with the rise of social media, yet its implications extend beyond the digital realm, constituting real societal challenges. Yet, policymakers are often constrained to make policies within a country, state or county and have little decision power to regulate online environments. For instance, prior research identified that media literacy and more broadly education were related to a higher tendency to identify fake news (Arin, Mazrekaj, & Thum, 2023;

Adjin-Tettey, 2022). Specifically in the United States, educational policy is primarily decided at the state level because of the 10^{th} amendment (of Education, 2024). More so, the high heterogeneity in fake news belief (Grinberg et al., 2019; Guess et al., 2019), would likely render any national media literacy policy inefficient.

A geographical prediction of fake news spread allows policymakers to judge whether policies are needed in their area, and to what extent they should be implemented. The challenge of such predictions stems from the population differences between social media and the real world (Mellon & Prosser, 2017). Beyond the fact that some areas are more densely populated than others, models need to account for the difference in population across those areas. A Washington DC resident is on average younger, has a higher income, and living in a city, is likelier to use social media than the average Wyomingite (Perrin, 2015). Geographical inferences on fake news spread must account for these differences lest their results be biased.

Acknowledgement

This paper was supported by a grant provided by the University Maastricht Behavioral Insights Center seeding grant provided by the School Of Business and Economics, Tongersestraat 53, 6211LM Maastricht, Netherlands.

References

Adjin-Tettey, T. (2022). *Combating fake news, disinformation, and misinformation: Experimental evidence for media literacy education. cogent arts & humanities, 9 (1), 2037229.*

Agarwal, V., Sultana, H. P., Malhotra, S., & Sarkar, A. (2019). Analysis of classifiers for fake news detection. *Procedia Computer Science*, *165*, 377–383.

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives*, *31*(2), 211–236.

Allen, J., Arechar, A. A., Pennycook, G., & Rand, D. G. (2021). Scaling up fact-checking using the wisdom of crowds. *Science advances*, *7*(36), eabf4393.

American National Election Studies. (2021). *ANES 2020 Time Series Study Full Release.* July 19, 2021 version. Retrieved from `www.electionstudies.org`

Arin, K. P., Mazrekaj, D., & Thum, M. (2023). Ability of detecting and willingness to share fake news. *Scientific Reports*, *13*(1), 7298.

Butz, A. M., & Kehrberg, J. E. (2016). Estimating anti-immigrant sentiment for the american states using multi-level modeling and post-stratification, 2004–2008. *Research & Politics*, *3*(2), 2053168016645830.

Cerina, R., & Duch, R. (2021). Polling india via regression and post-stratification of non-probability online samples. *Plos one*, *16*(11), e0260092.

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, *118*(9), e2023301118.

DeBell, M., & Krosnick, J. A. (2009). Computing weights for american national election study

survey data. *nes012427. Ann Arbor, MI, Palo Alto, CA: ANES Technical Report Series*.

Deming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, *11*(4), 427–444.

DiFonzo, N. (2011). *The echo-chamber effect.* https://www.nytimes.com/roomfordebate/2011/04/21/barack-obama- (Accessed: 2023-12-01)

Gelman, A., Lee, D., & Guo, J. (2015). Stan: A probabilistic programming language for bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*, *40*(5), 530–543.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on twitter during the 2016 us presidential election. *Science*, *363*(6425), 374–378.

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science advances*, *5*(1), eaau4586.

Guess, A., Nyhan, B., Lyons, B., & Reifler, J. (2018). Avoiding the echo chamber about echo chambers. *Knight Foundation*, *2*(1), 1–25.

Hussain, M. G., Hasan, M. R., Rahman, M., Protim, J., & Al Hasan, S. (2020). Detection of bangla fake news using mnb and svm classifier. In *2020 international conference on computing, electronics & communications engineering (iccece)* (pp. 81–85).

Instagram. (2019). *Combatting misinformation on instagram.* https://about.instagram.com/blog/announcements/combatting-misinformation-on-instagram. (Accessed: 2023-12-02)

Kolenikov, S. (2014). Calibrating survey data using iterative proportional fitting (raking). *The Stata Journal*, *14*(1), 22–59.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... others

(2018). The science of fake news. *Science*, *359*(6380), 1094–1096.

Leemann, L., & Wasserfallen, F. (2017). Extending the use and prediction precision of subnational public opinion estimation. *American journal of political science*, *61*(4), 1003–1022.

Mellon, J., & Prosser, C. (2017). Twitter and facebook are not representative of the general population: Political attitudes and demographics of british social media users. *Research & Politics*, *4*(3), 2053168017720008.

Mena, P. (2020). Cleaning up social media: The effect of warning labels on likelihood of sharing false news on facebook. *Policy & internet*, *12*(2), 165–183.

Meta. (2021). *How meta's third-party fact-checking program work.* `https://www.facebook.com/formedia/blog/third-party-fact-checking-how-it-works`. (Accessed: 2023-12-03)

Miller, D. T., & Prentice, D. A. (1994). Collective errors and errors about the collective. *Personality and Social Psychology Bulletin*, *20*(5), 541–550.

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., & Rosenquist, J. (2011). Understanding the demographics of twitter users. In *Proceedings of the international aaai conference on web and social media* (Vol. 5, pp. 554–557).

Mosleh, M., & Rand, D. G. (2022). Measuring exposure to misinformation from political elites on twitter. *nature communications*, *13*(1), 7144.

Nieminen, S., & Rapeli, L. (2019). Fighting misperceptions and doubting journalists' objectivity: A review of fact-checking literature. *Political studies review*, *17*(3), 296–309.

of Education, U. D. (2024). *Laws guidance.*

Osmundsen, M., Bor, A., Vahlstrup, P. B., Bechmann, A., & Petersen, M. B. (2021). Partisan polarization is the primary psychological motivation behind political fake news sharing on

twitter. *American Political Science Review*, *115*(3), 999–1015.

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

Pennycook, G., & Rand, D. G. (2019b). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50.

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in cognitive sciences*, *25*(5), 388–402.

Perrin, A. (2015). Social media usage. *Pew research center*, *125*, 52–68.

Poynter. (2019). *Most republicans don't trust fact-checkers, and most americans don't trust the media.* Retrieved from `https://www.poynter.org/ifcn/2019/most-republicans-dont-trust-fact-checkers-and-most-americans-`

Rich, T. S., Milden, I., & Wagner, M. T. (2020). Research note: Does the public support fact-checking social media? it depends who and how you ask. *The Harvard Kennedy School Misinformation Review*.

Ruggles, S., Flood, S., Sobek, M., Brockman, D., Cooper, G., Richards, S., & Schouweiler, M. (2023). *Ipums usa: Version 13.0 [dataset].* Minneapolis, MN: IPUMS. doi: 10.18128/D010.V13.0

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, *19*(1), 22–36.

Talwar, S., Dhir, A., Kaur, P., Zafar, N., & Alrasheedy, M. (2019). Why do people share fake news? associations between the dark side of social media use and fake news sharing behavior.

Journal of Retailing and Consumer Services, *51*, 72–82.

Tschiatschek, S., Singla, A., Gomez Rodriguez, M., Merchant, A., & Krause, A. (2018). Fake news detection in social networks via crowd signals. In *Companion proceedings of the the web conference 2018* (pp. 517–524).

t'Serstevens, F., Piccillo, G., & Grigoriev, A. (2022). Fake news zealots: Effect of perception of news on online sharing behavior. *Frontiers in Psychology*, *13*, 859534.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *science*, *359*(6380), 1146–1151.

Walker, M., & Gottfried, J. (2019). Republicans far more likely than democrats to say fact-checkers tend to favor one side.

Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, *31*(3), 980–991.

Wang, Z., Hale, S., Adelani, D. I., Grabowicz, P., Hartman, T., Flöck, F., & Jurgens, D. (2019). Demographic inference and representative population estimates from multilingual social media data. In *The world wide web conference* (pp. 2056–2067).

X. (2023a). *About community notes on x — x help.* Author. Retrieved from `https://help.twitter.com/en/using-x/community-notes`

X. (2023b). *How we address misinformation on x.* `https://help.twitter.com/en/resources/addressing-misleading` (Accessed: 2023-12-03)