Pokhriyal et al. *EPJ Data Science* (2023) 12:26
https://doi.org/10.1140/epjds/s13688-023-00405-6

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE Open Access

Quantifying participation biases on social media

Neeti Pokhriyal[1], Benjamin A. Valentino[2] and Soroush Vosoughi[1*]

*Correspondence:
soroush.vosoughi@dartmouth.edu
[1]Department of Computer Science, Dartmouth College, Hanover, NH, USA
Full list of author information is available at the end of the article

Abstract

Around seven-in-ten Americans use social media (SM) to connect and engage, making these platforms excellent sources of information to understand human behavior and other problems relevant to social sciences. While the presence of a behavior can be detected, it is unclear who or under what circumstances the behavior was generated. Despite the large sample sizes of SM datasets, they almost always come with significant biases, some of which have been studied before. Here, we hypothesize the presence of a largely unrecognized form of bias on SM platforms, called *participation bias*, that is distinct from selection bias. It is defined as the skew in the demographics of the participants who opt-in to discussions of the topic, compared to the demographics of the underlying SM platform. To infer the participant's demographics, we propose a novel generative probabilistic framework that links surveys and SM data at the granularity of demographic subgroups (and not individuals). Our method is distinct from existing approaches that elicit such information at the individual level using their profile name, images, and other metadata, thus infringing upon their privacy. We design a statistical simulation to simulate multiple SM platforms and a diverse range of topics to validate the model's estimates in different scenarios. We use Twitter data as a case study to demonstrate participation bias on the topic of gun violence delineated by political party affiliation and gender. Although Twitter's user population leans Democratic and has an equal number of men and women according to Pew, our model's estimates point to the presence of participation bias on the topic of gun control in the opposite direction, with slightly more Republicans than Democrats, and more men compared to women. Our study cautions that in the rush to use digital data for decision-making and understanding public opinions, we must account for the biases inherent in how SM data are produced, lest we may also arrive at biased inferences about the public.

Keywords: Social media data; Probabilistic modeling; Bias quantification

1 Introduction

Governments have increasingly sought to employ big data to improve policy making, as evidenced by initiatives such as the Foundations for Evidence-based Policy making Act in the United States [1], the European Union's data strategy [2] and the National Policy development framework in South Africa [3]. Although representative surveys are traditionally used as the primary tool to design policies that require information on public opinion and

Springer

Figure 1 A pictorial illustration of the participation bias, selection bias for a social media platform, and the underlying population distribution

behavior, they are time-consuming, expensive, and, in recent years, have been known to suffer increasingly from reduced participation and decreased accuracy.

As an alternative, social media (SM) data are frequently touted as a cheaper, more easily accessible, and timely source of data in diverse problem settings related to public opinion and mass behavior [4, 5]. Applications of SM data include epidemiology [6–8], migration [9–13], economics [14–17], politics and elections [18–20], and many more use cases.

Researchers have raised important concerns, however, about the bias resulting from the self-selection of individuals onto SM platforms [21–27], and have also explored it for select platforms [28, 29]. But this is not the only form of bias inherent to SM data. The platforms don't elicit responses from individuals across all topics in a uniform manner, i.e., individuals who participate on certain topics are not reflective of all the individuals who are on that platform. Motivated by this observation, we define the *participation bias* for a given topic as the skew in the demographics of the participants who opt-in to discussions on that topic, compared to the demographics of the underlying SM platform, as pictorially depicted in Fig. 1.

Researchers have studied the selection bias on SM platforms [21–29]. Here, we hypothesize the presence of participation bias on SM platforms, which is distinct from the selection bias discussed previously. Participation bias is associated with the outcome of a study on SM and is motivated by the observation that individuals who participate are not representative of all the individuals who are on that platform. In other words, whilst it is known that SM users are not typically representative of the underlying population (selection bias), we hypothesize and demonstrate, here, that users participating in (discussions of) different topics on SM are not representative of the underlying SM population. We define *participation bias* for a given topic as the skew in the demographics of the participants who opt-in to discussions on that topic, compared to the demographics of the underlying SM platform, as pictorially depicted in Fig. 1.

To get the demographic information for individuals, extant works have either used their profile images, names, and other metadata [30–32] or have imputed these variables. This raises critical privacy (scraping metadata from user's profiles to get demographic information) and ethical (is it *correct* to do so without explicit consent) questions [33, 34], as well as questions about accuracy, when pre-trained image models are used to infer gender, for example, they are less accurate for dark-skinned individuals [35, 36]. Recent attempts that match Twitter accounts with administrative data (voter files) present important privacy challenges [37, 38]. Methods that explicitly link individuals from a representative poll to their SM profiles upon consent to determine demographics suffer from a limited sample size of surveys and non-response biases [28, 37]. Such efforts are also costly. These meth-

ods also introduce unknown biases (depending on who shares their photos and personal information online).

1.1 Contributions

Contrasting existing approaches, we propose a novel *probabilistic generative model* to learn the demographic distribution of the participants for a given topic. Our model links information from representative surveys and SM data at the granularity of demographic subgroups and not individuals. To do so, we assume that the innate propensity of an individual to hold a specific view on a given topic remains unchanged if they are asked for their views via surveys or are *sensed* passively by examining their posts and behavior on SM. As an example, for the topic of pro-gun control if the individual is polled via polling agency or *sensed* on SM, it is natural to assume that they will have the same opinion).[1] Using this formalism, the model directly outputs the demographic subgroups of the participants and never infringes on individual-level privacy.

We play with the *strengths* of the two data sources. First, the availability of representative survey responses delineated by demographic subgroups, see Table 1, which is taken from NPR/PBS Newshour/Marist poll survey on the question of the importance of controlling gun violence and provides responses delineated by Democratic men, Democratic women, Republican men, and Republican women. Second, the availability of cheap and timely sensing of the opinion of SM users as demonstrated by [18, 39–42]. While the opinion of SM users on a topic can be calculated, the demographic information of the users who are generating this information is unknown. However, by utilizing the surveys, we know how a *typical* SM user belonging to a certain demographic subgroup would behave. This insight allows us to link three disparate quantities: the knowledge of how a demographic subgroup responds to a given topic, the opinions extracted from SM corresponding to that topic, and the demographic distribution of the participants generating these opinions on SM. This linkage forms the central tenet of our modeling framework.

Two significant *challenges* arise in this learning task. First, the task of linking surveys with SM data, and estimating opinions from SM data injects noise in our framework. The second challenge is the limited availability of representative surveys for a given topic at a time point. To handle these challenges, we propose a probabilistic generative model, which explains the observations (the *noisy* opinions on SM) in terms of the demographic distribution of the participants and their typical responses gathered from surveys. We model the "noisy" opinion on SM as a Beta random variable, with the mean as the estimate of opinion and variance as the noise associated with the mean. The learning task is *"What is the most likely demographic distribution that generated this opinion on SM?"*. The demographic distribution is modeled as a Dirichlet random variable, with predicted mean as the demographic subgroups and predicted variance as the uncertainty in these estimates. To mitigate the challenges of learning from limited data, our model incorporates existing knowledge of the demographic structure of the SM platform as informed prior.

To *validate* our model's estimates, we design a simulation that models multiple SM platforms (each with different underlying demographics), simulates discussion on diverse topics on these platforms, and injects varying levels of noise. For each of these settings, the

[1] For the topic of gun control polling suggests that attitudes on this subject vary strongly by political party affiliations and gender, which motivates our example.

simulation allows us to generate the distribution of the participating population apriori and this serves as the target distribution (against which our model's estimates are compared). Our validation results demonstrate that our model can recover the demographic distribution of the participants with high accuracy across different topics, different simulated platforms, and varying levels of noise.

We perform a case study on *Twitter data with NPR/PBS Newshour/Marist polls* on the topic of gun control in the United States. Twitter's population leans Democratic and has roughly equal numbers of men and women according to Pew [28]. We show that in discussions about gun control, our model estimates the demographics of the participating population to be skewed in the opposite direction with more Republicans and men participating on that topic. These results underscore the importance of quantifying the biases in how the SM data is produced. If we hope to use SM data to draw valid conclusions about the broader public or, even, to correctly contextualize studies done on SM, then we need to understand and account for these biases.

The rest of the paper is organized as follows: In the Methodology section, we begin by describing the intuition behind our model, followed by its problem formulation and model details. Then, the section, titled Simulation, details the design of the statistical simulator. This is followed by the section describing how Twitter data on gun control was gathered and processed to get an estimate of aggregated opinion. The results section describes the results of the validation results on simulated as well as Twitter data; followed by discussions and conclusions.

2 Methodology

In this section, we begin by providing an intuition of our model, followed by the problem formulation and its details. We, then, provide the inference procedure employed to get the demographic subgroups.

2.1 The proposed computational framework

Figure 2 describes our proposed computational framework for a topic (t). At its core lies the *probabilistic bias quantification (Biq)* model. There are two sets of inputs to our model. First is the response probabilities corresponding to each demographic subgroup for N survey questions, as shown in Table 1, these values for Democratic-men, Democratic-women, Republican-men, Republican-women for the topic of controlling gun-violence (column 4) are 0.86, 0.92, 0.20, 0.26 respectively. Public opinion surveys and polls provide such representative responses delineated by different demographic subgroups of interest and are used in this work. The second set of inputs is the estimates of aggregated opinion on SM calculated for each of the N survey questions. The output is the demographic distribution of the participants on SM for topic t. The participation bias is, then, the skew in the participant's demographics compared to the demographics of the underlying SM platform. It is important to note that the granularity of our analysis is the demographic subgroups of interest, and never individuals.

Our proposed Biq model belongs to the class of generative probabilistic models that describe a hypothetical random process by which the observed data are generated. The "noisy" aggregated opinions on SM are observed, while the underlying demographic structure that generated the opinion remains hidden. Biq reduces the process of how opinions are generated on SM to a set of simple probabilistic steps, as defined later in this section.

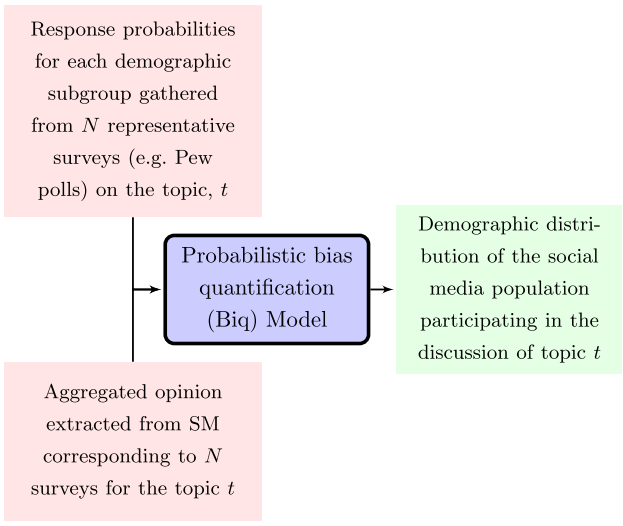


Figure 2 The proposed computational framework with the probabilistic bias quantification (Biq) model at its core. The two inputs are shown in pink. The first input is the response probabilities for each demographic subgroup gathered from *N* representative surveys on a given topic. The second input is the aggregated opinions extracted from social media data corresponding to those *N* survey questions. The output of the Biq model is the demographic distribution of the participants on SM

Table 1 Snapshot of survey data from the NPR/PBS Newshour/Marist poll

Party ID and Gender	National Adults		
	Do you think it is more important to:		
	Protect gun rights Row %	Control gun violence Row %	Unsure Row %
Democrat men	13%	86%	1%
Democrat women	5%	92%	3%
Republican men	74%	20%	5%
Republican women	63%	26%	12%

The inference task is to use the observed data to infer the hidden demographic structure of users. By modeling the parameters of interest as random variables we are able to handle the noise injected into our computational framework during translating survey questions into SM queries and extracting opinions from SM data.

We convey our *model's intuition* using an example provided in Table 1, which is a snapshot from NPR/PBS Newshour/Marist poll survey. Focusing on the question of the importance of controlling gun violence (column 3), we see that 86% of Democratic men respond with a yes. Our model's assumption states that Democratic men on SM will likely respond with a yes/pro opinion to the query of controlling gun violence with a 0.86 probability. For Democratic women, Republican men, and Republican women these (pro) response probabilities are 0.92, 0.20, and 0.26 respectively.

If we *sense* individuals on SM who have opted-in to discussions on the topic of gun control and get their aggregated pro-opinion, then this aggregated (pro)opinion can be seen as the weighted sum, where each addend is the (pro)response probability of a demographic subgroup (for Dem. men it is 0.86) weighted by the fraction of the participants on SM be-

longing to that demographic subgroup (the number of Dem. men, which is not known).[2] Each survey question is, thus, translated to an equation (See (1)). We demonstrate that with a set of survey questions and opinions as described below, we can estimate the demographic distribution of participants on SM.

2.2 Problem formulation

We assume D demographic subgroups indexed by j. For a given topic, we assume N surveys, indexed by i. Each survey gets translated into an equation with three sets of variables defined as follows:

1. For the ith survey, $\mathbf{x}^i \equiv [x_1^i, x_2^i, \ldots, x_D^i]$, denotes a D length vector consisting of response probabilities for the D demographic subgroups. Thus, x_j^i is the response probability of the jth demographic subgroup for the ith survey, shown in blue in (1).
2. The aggregated opinion from SM is taken as a random variable, whose mean is denoted by δ_i, and variance is denoted as ϕ_i, shown in red in (1). The N surveys are translated to N queries to SM and, thus, we get N equations.
3. The demographic distribution on SM, to be estimated, is denoted by the vector $\mathbf{w} \equiv [w_1, w_2, \ldots, w_D]$ and w_j is the fraction of the participating population in SM that belongs to the jth demographic subgroup as shown in green in (1).

The aggregated SM opinion for a survey i, δ^i, is calculated as follows:

Demographic distribution to be estimated

Survey responses for demographic subgroups

Opinion from social media

$$\begin{aligned} w_1 x_1^1 + w_2 x_2^1 + \ldots + w_D x_D^1 &= \delta^1 \\ w_1 x_1^2 + w_2 x_2^2 + \ldots + w_D x_D^2 &= \delta^2 \\ &\vdots \\ w_1 x_1^N + w_2 x_2^N + \ldots + w_D x_D^N &= \delta^N \end{aligned} \tag{1}$$

If an exact estimate of δ^is were available, N surveys would be needed to exactly estimate $\mathbf{w}$. But, since δ^i is noisy in SM, one could estimate $\mathbf{w}$ is by minimizing the residual sum of squares given as:

$$J(\mathbf{w}) = \sum_{i=1}^{N} \left(\delta^i - \mathbf{w}^\top \mathbf{x}^i\right)^2. \tag{2}$$

But this neither incorporates the uncertainty associated with δ^i nor guarantees that the values in the estimated vector, $\mathbf{w}$, lie between 0 and 1, and sum up to 1. To address these issues, we propose a generative probabilistic model, as outlined below.

2.3 The bias quantification (Biq) model

The *graphical representation* of the proposed generative probabilistic model is given in Fig. 3. We assume the following *generative process* for δ^i for the N surveys:

[2] Existing research has dealt with getting the aggregated opinion on SM corresponding to a topic of interest. Readers are directed to the survey [42].

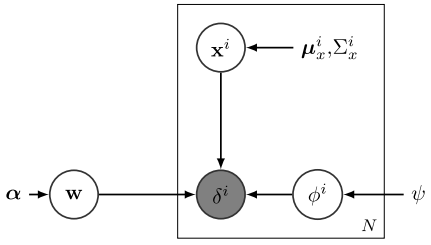


Figure 3 Graphical representation of the proposed probabilistic model. Each node is a random variable (r.v). Shaded node indicates that r.v is observed and the unshaded ones indicate hidden values. The rectangles denotes replication. The N plate denotes number of surveys

1. Randomly sample the demographic distribution weight vector, $\mathbf{w}$, from a *Dirichlet* distribution, i.e., $\mathbf{w} \sim \text{Dirichlet}(\boldsymbol{\alpha})$, where α is the hyper-parameter for the *Dirichlet* distribution.
2. For the ith survey:
   (a) Randomly sample the noise term associated with the ith survey, ϕ^i, from an *Inverse-Gamma* distribution, i.e., $\phi^i \sim \text{InvGamma}(\psi)$.
   (b) Randomly sample the survey response vector, $\mathbf{x}^i$ from Gaussian distribution with mean, $\boldsymbol{\mu}_x^i$, and covariance matrix, Σ_x^i, i.e., $\mathbf{x}^i \sim \text{Gaussian}(\boldsymbol{\mu}_x^i, \Sigma_x^i)$.
   (c) Calculate the parameters, α^i and β^i, for the *Beta* distribution that will be used to sample the opinion for the ith survey, as follows:

$$\mu^i = \frac{1}{1 + \exp\left(-\mathbf{w}^\top \mathbf{x}^i\right)}, \tag{3}$$

$$\alpha^i = \mu^i \phi^i, \tag{4}$$

$$\beta^i = \left(1 - \mu^i\right)\phi^i, \tag{5}$$

   (d) Randomly sample the opinion for the ith survey, δ^i, from a *Beta* distribution, with parameters α^i and β^i, i.e., $\delta^i \sim \text{Beta}(\alpha^i, \beta^i)$.

We assume a Dirichlet prior on the demographic distribution vector, $\mathbf{w}$, with a specified prior parameter, $\boldsymbol{\alpha}$. The survey response vector for the demographic subgroups, $\mathbf{x}^i$, is assumed to be a Gaussian-distributed random variable.[3] The noise associated with the SM aggregated opinion is denoted as ϕ^i, and is sampled from an *Inverse-Gamma* distribution, which ensures that it takes only positive values. The shape parameter, ψ, associated with the *Inverse-Gamma* prior controls the variance in the magnitude of the sampled values. Finally, the SM aggregated opinion, δ^i, is modeled as a *Beta* distributed random variable, whose parameters, α^i and β^i, are derived using the mean and variance values, μ^i and ϕ^i, respectively (See (4) and (5)). The mean, μ_i, is a function of the demographic distribution and the survey responses, as shown in (3). The *sigmoid* transformation in (3) ensures that μ^i is between 0 and 1.

Given the parameters, $\boldsymbol{\alpha}$, ψ, $[\boldsymbol{\mu}_x^i]_{i=1}^N$ and $[\Sigma^i]_{i=1}^N$, the joint distribution of the SM demographic distribution, $\mathbf{w}$, the SM aggregated opinion, $\boldsymbol{\delta} \equiv \{\delta^1, \ldots, \delta^N\}$, the variance terms, $\boldsymbol{\phi} \equiv \{\phi^1, \ldots, \phi^N\}$, and the set of survey responses, $[\mathbf{x}^i]_{i=1}^N$, is given by:

$$p\left(\mathbf{w}, \boldsymbol{\delta}, \boldsymbol{\phi}, \left[\mathbf{x}^i\right]_{i=1}^N | \boldsymbol{\alpha}, \psi, \left[\boldsymbol{\mu}_x^i\right]_{i=1}^N, \left[\Sigma_x^i\right]_{i=1}^N\right)$$

[3] The prior on $\mathbf{x}^i$ accounts for the errors associated with survey results. We assume that the prior covariance matrix is diagonal with equal variance, i.e., $\Sigma_x^i = \sigma_x^{2i} I$, where the σ_x^i corresponds to the error margin associated with the ith survey.

$$= p(\mathbf{w}|\boldsymbol{\alpha}) \prod_{i=1}^{N} p\big(\delta^i|\alpha^i, \beta^i\big) p\big(\phi^i|\psi\big) p\big(\mathbf{x}^i|\boldsymbol{\mu}_x^i, \Sigma_x^i\big), \tag{6}$$

where α^i and β^i are derived using (3), (4) and (5). Integrating over ϕ^is and $\mathbf{x}^i$s, we obtain the following marginal distribution:

$$\begin{aligned} &p\big(\mathbf{w}, \boldsymbol{\delta}|\boldsymbol{\alpha}, \psi, \big[\boldsymbol{\mu}_x^i\big]_{i=1}^N, \big[\Sigma_x^i\big]_{i=1}^N\big) \\ &= p(\mathbf{w}|\boldsymbol{\alpha}) \prod_{i=1}^{N} \int \int p\big(\delta^i|\alpha^i, \beta^i\big) p\big(\phi^i|\psi\big)\, d\phi^i\, d\mathbf{x}^i \end{aligned} \tag{7}$$

Further integrating over $\mathbf{w}$, we obtain the marginal distribution for the SM aggregated opinion, $\boldsymbol{\delta}$:

$$p\big(\boldsymbol{\delta}|\boldsymbol{\alpha}, \psi, \big[\boldsymbol{\mu}_x^i\big]_{i=1}^N, \big[\Sigma_x^i\big]_{i=1}^N\big) = \int p(\mathbf{w}|\boldsymbol{\alpha}) \prod_{i=1}^{N} \int \int p\big(\delta^i|\alpha^i, \beta^i\big) p\big(\phi^i|\psi\big)\, d\phi^i d\mathbf{x}^i\, d\mathbf{w} \tag{8}$$

Note that α^i and β^i depend on $\mathbf{w}$ and hence the inner term in (8) cannot be moved out of the integral.

To infer the posterior distribution for the demographic distribution vector, $\mathbf{w}$, given the observed SM aggregated opinion, $\boldsymbol{\delta}$, we apply Bayes rule to obtain:

$$p\big(\mathbf{w}|\boldsymbol{\delta}, \boldsymbol{\alpha}, \psi, \big[\boldsymbol{\mu}_x^i\big]_{i=1}^N, \big[\Sigma_x^i\big]_{i=1}^N\big) = \frac{p(\mathbf{w}, \boldsymbol{\delta}|\boldsymbol{\alpha}, \psi, [\boldsymbol{\mu}_x^i]_{i=1}^N, [\Sigma_x^i]_{i=1}^N)}{p(\boldsymbol{\delta}|\boldsymbol{\alpha}, \psi, [\boldsymbol{\mu}_x^i]_{i=1}^N, [\Sigma_x^i]_{i=1}^N)}, \tag{9}$$

where the numerator and denominator terms are calculated using (7) and (8), respectively.

The presence of the integral term within the product in (7) and (8) makes the above posterior intractable to compute. Hence, we employ a *Normalized Importance Sampling* to obtain the approximate posterior distribution for $\mathbf{w}$, which is also modeled as a *Dirichlet* distribution, as described below.

2.4 Biq model inference

We employ self-normalized importance sampling [43] to calculate the posterior. The idea behind importance sampling is to sample from an easy-to-sample distribution q and, then, to reweigh the samples, so as to get an approximation of the original posterior (p), which is difficult to sample from.

Our task is to find $\mathbb{E}_{x\sim p}[f(\boldsymbol{X})] = \int_D f(x)p(x)\, dx$. We have the following importance distribution q:

$$\begin{aligned} \mathbb{E}_{x\sim p} \int_D f(x)p(x)\, dx &= \int_D \frac{f(x)p(x)}{q(x)} q(x)\, dx \\ &= \mathbb{E}_{x\sim q} \frac{f(\boldsymbol{X})p(\boldsymbol{X})}{q(\boldsymbol{X})}, \end{aligned} \tag{10}$$

where $\mathbb{E}_{x\sim q}$ is expectation for $\boldsymbol{X} \sim q$ and $w(x) = p(x)/q(x)$ is the weight. So, samples are taken from q and weighted by $w(x)$. The expected value of this Monte Carlo approximation is the original integral.

However, we are interested in calculating the unnormalized version of our posterior probability as its denominator is intractable. Let the unnormalized version of p be $p_u(x) = cp(x)$, where c is unknown. We compute the ratio $w_u(x) = p_u(x)/q_u(x)$.

3 Design of simulation

We designed a simulation to simulate different SM platforms, a variety of topics on each SM platform, and a mechanism to inject different levels of noise in these scenarios. The simulation helps to generate the target distribution for each of the above-mentioned scenarios and thus provides a robust validation of our model's estimates.

3.1 Overview of simulation

Figure 4 shows the three phases of the simulation – simulating polls, simulating an SM platform, and extracting opinion from SM, as described below:

1. *Phase 1: Simulating polls*: The first phase is analogous to conducting opinion polls or surveys for a set of topics, where for each topic a set of survey questions are asked.[4] We simulate a set of topics and for each topic, we sample the response probabilities for each demographic subgroup from a statistical distribution.

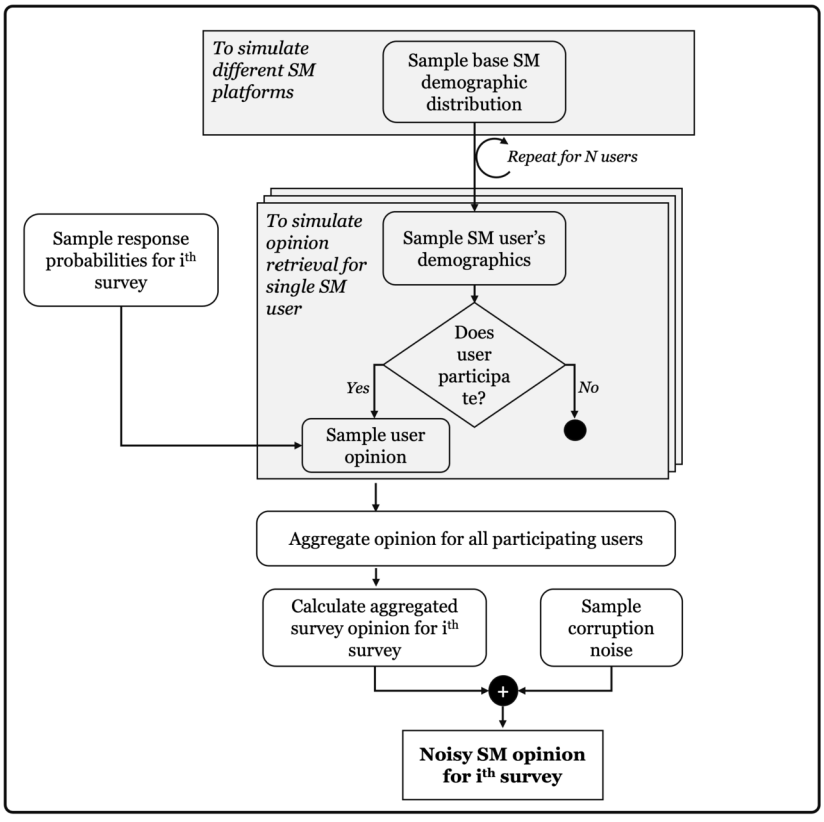


Figure 4 Flowchart illustrating the simulation of polls and opinion extraction from SM. Every sampling step uses an appropriate statistical prior

[4]For example, for gun control, the survey questions correspond to understanding the public's support for issues on gun curbing violence, the necessity of background checks, assault weapon bans, etc.

2. *Phase 2: Simulating SM platform*: The second phase is analogous to creating an SM population, where we generate users belonging to different demographic subgroups. We employ priors that let us simulate different SM populations with varying demographic distributions.
3. *Phase 3: Simulating opinion retrieval on SM*: The third phase is analogous to eliciting an aggregated opinion for a given survey question on SM. It involves determining if an SM user will participate or not. If yes, then the user's opinion about the topic is sampled based on the response probability corresponding to that demographic subgroup (obtained from step 1). By aggregating the opinions of all the participating individuals, we generate an aggregate opinion for that survey question. We also generate a variance associated with this quantity, signifying the uncertainty, using another statistical distribution. We also corrupt the aggregated opinion using a noisy term.

The three phases described above involve statistical priors and changing these priors allows us to simulate different scenarios. Some of the example scenarios are different SM platforms, where each platform has a different underlying demographic distribution. The simulation also allows us to get the target distribution, against which our estimates can be compared.

3.2 Details of the simulation

The simulation procedure described above is devised to generate three *quantities* for a given topic:

1. The response probabilities, $\mathbf{x}^i$, for a set of N surveys.
2. The demographic subgroup vector, $\mathbf{w}$, which is the unknown.
3. The SM aggregated opinion, δ^i (and the corresponding uncertainty ϕ^i), for each of the N surveys.

While the simulation process is agnostic to the demographic subgroup studied, here, we illustrate the process for gender demographic (*female* (*f*) and *male* (*m*)), and two political party affiliations (*Republican* (*r*) and *Democrats* (*d*)). Thus, the vectors $\mathbf{x}^i$ and $\mathbf{w}$ will be of length 4. We will also assume, without loss of generality, that the entries in these vectors correspond to 1-*Republican-female* (*rf*), 2-*Republican-male* (*rm*), 3-*Democratic-female* (*df*), and 4-*Democratic-male* (*dm*).

For the *first* quantity, we sample each entry in the vector, $\mathbf{x}^i$, from a Beta distribution with a different parameter for each demographic subgroup (See Additional file 1 Section Parameter Choices for exact details).

For the *second* quantity, we assume the availability of three hyper-parameters as follows:

1. The first hyper-parameter is the *base* demographic distribution for the entire population. For example, for the gender demographic, this could be $[\bar{w}_f, \bar{w}_m] \equiv [0.50, 0.50]$.
2. The second hyper-parameter, $\kappa_s \in [0, 1]$, denotes the SM platform bias towards a political affiliation. $\kappa_s = 0.5$ denotes a neutral platform, $\kappa_s \approx 1$ denotes a platform with almost all Democrats, and $\kappa_s \approx 0$ denotes a highly Republican platform.
3. The third hyper-parameter, $\kappa_p \in [-1, +1]$, denotes the preference of individuals belonging to the two political affiliations to participate in the discussions related to the topic. Positive values for κ_p indicate that more Democrats are participating, negative values for κ_p indicate that more Republicans are participating, while $\kappa_p = 0$ indicates that both party affiliations are participating equally on that topic.

These hyper-parameters are used to sample users, specified by their gender and political affiliation, and a participation indicator, *yes* or *no*. The users for which the participation indicator is *no* are dropped from further analysis. Let $\mathcal{U}$ denote the set of participating users.Their gender and political affiliations are used to determine the distribution vector, **w**. For example, for *Republican-females*, the corresponding entry in **w** will be:

$$w_{cf} = \frac{|u : (u \in \mathcal{U}) \wedge (u_g = f) \wedge (u_p = r)|}{|\mathcal{U}|}, \tag{11}$$

where u_g and u_p denote the gender and political affiliation of the user u, respectively.

For the *third* quantity, we sample a response to the ith survey question for each of the participating users in $\mathcal{U}$, by sampling a binary response (0 or 1) from a Bernoulli distribution whose parameter is chosen from the entry in vector $\mathbf{x}^i$ corresponding to the gender and political affiliation for that user. For example, if the jth user is Democrat and female, the response, r_j^i, will be a sample from a Bernoulli distribution with parameter x_{df}^i. Given the sampled responses for all of the users, the raw aggregated opinion, $\bar{\delta}^i$, is calculated as:

$$\bar{\delta}^i = \frac{\sum_{j=1}^{|\mathcal{U}|} r_j^i}{|\mathcal{U}|}. \tag{12}$$

We, then, add a noise term to $\bar{\delta}^i$ to obtain the final δ^i. To add noise, we sample δ^i from a Bernoulli distribution whose mean is $\bar{\delta}^i$ and the variance is given by a noise hyper-parameter, ϵ.

Finally, we sample the *uncertainty* associated with δ^i from an *Inverse-Gamma* distribution whose shape parameter is given by $\frac{1}{\psi}$, where ψ is a positive uncertainty hyper-parameter. The hyper-parameter choices for the simulation study are given in the Additional file 1 Section Parameter Choices.

4 Case study on Twitter data: estimating the aggregated opinion

To evaluate using real-world SM data, we use Twitter data on the topic of gun violence. The distribution of responses on gun violence across party identification and respondent gender are gathered from traditional surveys conducted by *NPR/PBS NewsHour/Marist polls* in February and September 2019 on a representative sample of the American public [44, 45]. We use data from 8 different survey questions that asked subjects to indicate their positive or negative responses on a variety of questions related to gun violence. For the list of survey questions and their responses see Additional file 1 Table 1.

About 3 Million raw Tweets were analyzed to get the aggregated opinion of the Twitter users corresponding to each survey question. More quantitative details about the Twitter data used in this study are given in Additional file 1 Table 2.

The aggregated opinion extracted from Twitter for each survey is assumed to be a random variable, generated from a *Beta* distribution. This formulation helps to encode the *uncertainty* as the variance of the random variable.

The aggregated opinion of Twitter's user population corresponding to each survey question is calculated as the ratio of the individuals who are sensed as having a "pro" opinion on that topic and the total individuals who expressed opinion (via an original tweet or a retweet) on that question. Thus, our goal is to assign individuals to either "pro" or "anti" camps based on their tweet content, which we accomplish in the following steps:

1. *Getting the tweets*: We translate each survey question into a set of keywords and hashtags (See Additional file 1 Table 3 for more details). Using Twitter's Developer API, we obtained all tweets and retweets (content and associated metadata) that contained one of the keywords and hashtags for the survey question and were posted during the time frame coinciding with the surveys. We acknowledge that this translation is only approximate, and can contribute to significant noise in our framework. We model this noise as uncertainty and study its impact on the accuracy of our model's estimates.
2. *Filtering the tweets*: The initial set of tweets (and retweets) were filtered to retain only those written in English and by users located in the US, where location is inferred from the location field in the metadata by performing geographic entity matching [46]. Tweets from automated bots and verified accounts (mostly belonging to news media sources, etc.) were removed.
3. *Labeling a small subset of tweets as "pro" or "anti"*: From the cleaned collection of tweets, we extract the most commonly occurring hashtags (via frequency count) and manually identify the hashtags corresponding to the "pro" and "anti" camps. We use the hashtags in each camp to label a small subset of tweets that contain those hashtags as "pro" or "anti", which is a low-cost, fast, and reliable method to assign stance to a large number of tweets [18, 40]. A careful human inspection of a random set of tweet content in each camp was done to ensure that the tweet content reflected the stance of that camp. This subset of tweets with pro/anti labels is our training set. For details of how the Tweets were processed, please see Additional file 1 Fig. 1.
4. *Labeling all the tweets using a classifier*: To assign a label to other tweets that did not contain the above hashtags, we trained a binary logistic regression classifier on the *Sentence-BERT* [47][5] embeddings of the labeled set of tweets (obtained in the previous step). The embeddings were reduced to 20 features using the top 20 principal components obtained from the embeddings. The labeled tweets were re-sampled to ensure that both classes were equally represented in the training data. The 10-fold cross-validation accuracy of the classifier (cv) was also noted as a measure of the inaccuracy of the trained classifier.
5. *Calculating the aggregated opinion*: The trained classifier was applied to each tweet, t, in the original collection, to obtain its probability to belong to the "pro" camp, denoted as $p_{t,\text{pro}}$. The probability was then adjusted using the cross-validation accuracy to propagate the uncertainty associated with the classifier downstream, as follows $\hat{p}_{t,\text{pro}} = cv * p_{t,\text{pro}} + (1 - cv) * p_{t,\text{anti}}$.

   We collected all tweets posted by a given user and computed the average of the per-tweet probability, which represents the probability of the user u belonging to the "pro" camp for the given survey question. We assume that the per-user probabilities are samples from a *Beta* distribution, and we estimate the parameters of this distribution from these samples. We then calculate the expected value (mean) and the variance for the fitted distribution, as the estimates of the aggregated opinion (δ^i) and its uncertainty (ϕ^i), respectively. The aggregated opinion of the Twitter user population corresponding to each survey question, along with their uncertainty, are given in Additional file 1 Table 1.

[5] We note that other classifiers and embeddings can be explored to learn parity of the Tweets. Here, we use a popular method for our learning task.

5 Results

In this section, we detail the results, first on the simulated data and then on Twitter data for the topic of gun violence.

5.1 Experimental setup

We conducted three sets of experiments on simulated data to answer the following questions related to the efficacy of our model:

1. How close are our model's estimates to the target distribution for *different topics on the same SM platform*? Each topic elicits a different demographic distribution of the participants.
2. How close are our model's estimates to the target distribution for *different topics across different SM platforms*? Each SM platform has a different underlying demographic distribution.
3. For a given SM, how robust are our model's estimates with *increasing noise* with our estimates on the SM platform?

The topics are varied across a continuous spectrum of political affiliations ranging from those that elicit mostly Republican participation to those that elicit mostly Democratic participation.

For each topic, we simulate a set of surveys analogous to survey questions in a traditional poll. As an example, on the topic of gun control topic, survey questions include whether the individual supports gun control in general, whether they support background checks, and whether they support banning assault weapons. For each topic, we simulate *30 survey questions* that are used to estimate the demographic distribution using the Biq model. For more details on experiment setup, see Additional file 1 Section Parameter choices.

Metrics We compare our model's estimates against the target distribution and measure the *Pearson correlation coefficient* and the *Mean Absolute Error* (MAE), which is the average absolute difference between the estimated and true values across all demographic subgroups. We also provide the r-squared values for each subplot.

5.2 Simulation results

Overall, our results demonstrate that our model can successfully uncover the demographic distribution of the participants in each of the above scenarios.

The topics range from those inducing mostly Republican participation (indicated in deep red) to neutral topics all the way to those inducing mostly Democratic participation (indicated in deep blue). Overall, we run the simulation for 21 such topics (See Additional file 1 Section Parameter Choices for details). For each topic, our model estimates the distribution of four demographic subgroups, namely Democratic-men, Democratic-women, Republican-men and Republican-women, shown by different shapes. Since there are 21 topics and 4 demographic subgroups, in the scatter plots below there are 21×4 points.

Detailed results for each of the above settings are given below:

5.2.1 Robustness to different topics on same SM platform

Figure 5 shows the scatter plot of our model's estimates and target distribution, and we show that our model can accurately uncover the participating demographics for different topics with a mean absolute error of 0.03 and Pearson's r coefficient of 0.93.

In our simulation, given that males and females are taken to be equally distributed and each gender is equally divided by party affiliation, the fraction of Republican-men,

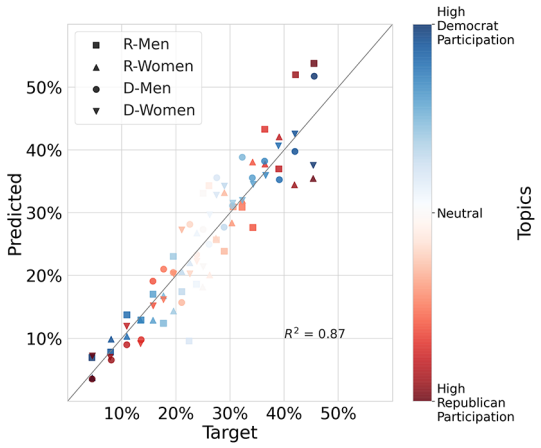


Figure 5 Scatter plot depicting how our model's estimates compare against the oracle for a set of topics on the same SM platform. Each topic induces a different set of the demographic subgroup to be preferentially active

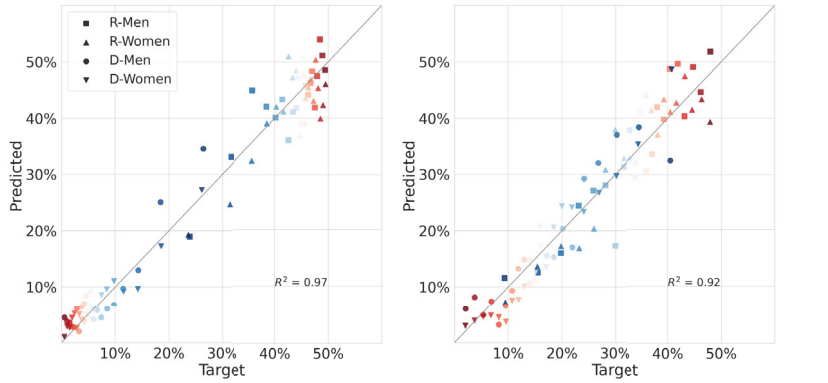
Republican-women, Democratic-men and Democratic-women is 25% each. For these experiments, the underlying demographic distribution on the platform is taken to be neutral (unbiased), and moderate noise is assumed. There are three distinct regions in the scatter plot corresponding to each topic: (1) For neutral topics, our model retrieves this distribution, which is shown as light-colored shapes and lies at the center of the plot. (2) For topics eliciting mostly Democratic participation, our model recovers these high percentages of Democrats shown by deep blue circles and inverted triangles and they fall on the top right corner of the plot. Our model also recovers the low percentages of Republicans in this setting, which are demonstrated by deep blue squares and triangles and fall on the lower left corner of the plot. (3) Similar inference can be drawn for the topic that elicits mostly Republican participation, which is shown by deep red shapes on either end of the diagonal.

We, also, notice that in the very extreme case of highly skewed participation by either democratic subgroup, our model slightly overestimates men and underestimates women falling in the majority demographic subgroup, shown by the vertical deviation in the 4 points on the top right of the plot. However, for both of these cases, our estimates are highly accurate for the minority demographic subgroup, shown by points superimposed on each other on the bottom left.

5.2.2 Robustness to different SM platforms

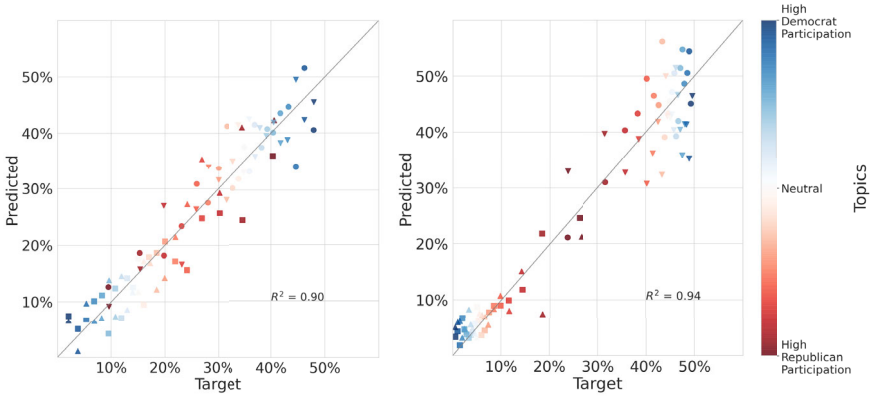
Each subplot in Fig. 6 is the result of estimating participant demographics for the full spectrum of topics for four different SM platforms – ranging from a platform where the underlying demographic distribution consists of mostly Republicans (leftmost) to that consisting of mostly Democrats (rightmost). We demonstrate that our model can uncover the demographic distribution on different SM platforms. Our model's errors never increase more than 0.03 and the Pearson's correlations are always greater than 0.95 with very low p-values, and we have stable r-squared values too.

Focusing on Fig. 6(a), where the platform is populated mostly by conservatives and the topic elicits mostly Republican participation, we notice that all deep red points are towards the extremes of the diagonal. Most squares and triangles are pushed toward the top right



(a) Platform demographics: Almost all Republicans

(b) Platform demographics: Mostly Republicans



(c) Platform demographics: Mostly Democrats

(d) Platform demographics: Almost all Democrats

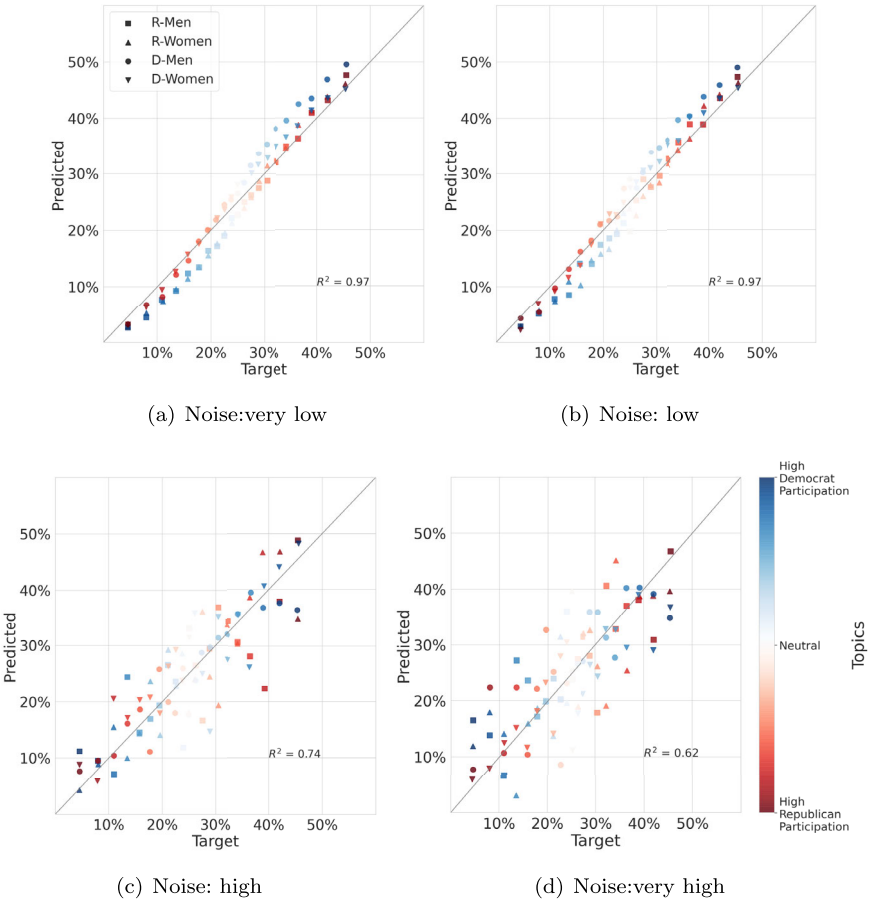
Figure 6 Robustness to different SM platforms: Accuracy of our model's estimates in uncovering participation bias different SM platforms

signifying high numbers of Republicans and all circles and inverted triangles are pushed toward the lower left signifying very less Democrats. Comparing Fig. 6(d) with Fig. 6(b), we notice similar inferences for a platform that is mostly Democratic. We, also, notice that on an extremely Democratic platform, if the topic elicits neutral participation, then we see much more Democrat participation than Republicans. This is evident by the dull red and blue points lying toward the edges of the diagonals. Similar inference for Fig. 6(a) too.

Moving to Fig. 6(b), we see how our model performs on a platform with a less extreme skew towards conservatives, with red points *spreading* towards the center of the plot and all points lying close to the diagonal. Similar inferences can be drawn for Fig. 6(c).

5.2.3 Robustness to noise on the aggregated opinion gathered from SM

In Fig. 7, we notice that the errors of our model's estimates increase with increasing noise. Highly accurate results, depicted by the close proximity of points to the diagonal on the left-most plot, are obtained for lower values of noises, with an error of 0.03. For moderate noise values of about 10%, our estimates deviate by 0.02, as seen by the spread of points along the diagonal. However, for very high values of noise (>20%), rightmost plot, our



(a) Noise:very low (b) Noise: low

(c) Noise: high (d) Noise:very high

Figure 7 Varying noise: Accuracy of our estimates with increasing noise to SM aggregated opinion

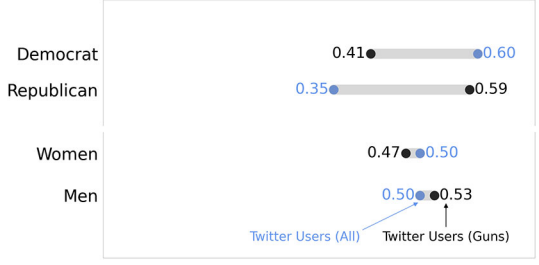


Figure 8 Participation bias on the topic of the gun on Twitter estimated by our model. The numbers in blue show the underlying Twitter demographic distribution from Pew [28]

estimates incur an error of 0.04 with Pearson's r coefficient of 0.83 and r-squared values of 0.62.

5.3 Gun violence on real data from Twitter as a case study

Figure 8 shows the *participation bias for the gun violence topic*. Although the underlying demographic distribution of Twitter is 60% Democratic, 35% Republican [28] and an equal number of males and females, our model estimates the participant's distribution to be

composed of 41% Democrats, 59% Republicans, and 53% males and only 47% females. Our model's estimates, along with the uncertainties, for each of the demographic subgroups are shown in Additional file 1 Fig. 2.

We also attempt to see how *comparative methods* that infer demographics using Twitter users' photos, names, and other metadata might work on our data (even though there is no direct comparison since they infringe on an individual's privacy). To the best of our knowledge, there are no existing methods that can directly infer if a Twitter user is a Democratic-man/Democratic-woman/Republican-man/Republican-woman.

Extant methods can infer the gender information for Twitter users with high accuracy [48, 49]. We applied one such method [48] to our dataset for Twitter users, which inferred 57% males and 43% females. These percentages again point to the over-representation of males in these topical discussions, and concur with our findings. Inferring the political affiliation of a user on Twitter is a highly contested subject, with studies claiming that the task is not easy and cautioning the generalizability of classifiers used in existing works [50, 51]. It should be noted that existing works [50, 52] have validated their methods for Twitter users who are either legislators or state their political preferences publicly or whose political contributions are known. It is important to have methods that can infer if a Twitter user is a Democrat vs Republican reliably with high accuracy if they are used as ground truth for our validation purposes, thus we did not extract political affiliations for Twitter users in our dataset.

We, also, *qualitatively validate* our findings using data from the 2019 Pew Research Center's poll, which found that men were 2.9 times more likely than women to say that they often or sometimes visited websites about guns, and Republicans were 3.4 times more likely than Democrats to do so [53], meaning that men and especially Republican men might be more participating in online discussions. Recent works highlighting that political right enjoys higher amplification on Twitter than political left [54], and the partisan asymmetries of left and right use of digital media [55] might also offer clues in support of our findings. However, more concerted research efforts along with human validation are needed to conclusively verify these findings.

6 Conclusions

Our work begins with the assumption that there are significant asymmetries in demographic participation in SM for different topics, and, thus, puts forth the notion of the existence of participation bias, induced by individuals who chose to be on the platform. We contribute to the existing scholarship by proposing a novel computational framework to estimate the biased demographic distribution by linking surveys with SM data at the granularity of demographic subgroups, without relying on individual-level data.

Our model has several attractive properties that include robustness to noise (that is attributed to the linking of survey and "noisy" SM data) and the ability to learn even with limited survey data. Our formulation has the potential to estimate a much finer degree of demographic subgroups, e.g., Republican male 65+, provided that we have availability of survey responses at that granularity. Since recent works [56] point that a significant proportion of Twitter users can be non-political, our formulation can be used to incorporate the non-political dimension with the political ones. However it is dependent on the availability of representative survey responses along those dimensions. Our approach can potentially be applied to SM platforms that do not collect or make public user images and other metadata.

6.1 Limitations

A significant challenge posed by our computational framework is the difficulty of translating survey questions into efficient queries on the SM platform. Here, we focused on questions that elicit binary answers (the wording of each question is provided in Additional file 1 Table 1), and these questions are translated into a set of keywords and hashtags to query Twitter. We acknowledge that this translation may only approximately capture the intent of the question, and, thus, contributes to noise in our computational framework. We study the impact of increased noise on our model's estimates using simulated data and demonstrate their robustness. However, more concerted research efforts are needed on how to translate a broader range of survey questions into queries on SM.

We also discuss some of the caveats associated with our Twitter results. First, the term *participation* of a user account on Twitter means accounts that are tweeting and retweeting on the topic of the gun. A more encompassing definition of *participation* that includes following, and liking other accounts or news media or prominent political figures can be explored. Second, the participation bias uncovered for guns topic is limited by the survey questions we investigated on Twitter during the Feb-Sep 2019 timeline. While our survey questions covered major aspects of the gun control topic, namely overall support for gun control, red-flag laws, requiring background checks, etc. for the given period, a broader set of questions might produce a different participation bias. Third, our initial set of tweets is collected via keyword search, which can potentially induce some selection bias in our analysis. Additionally, the filtering steps that include removing non-English Tweets can also induce potential selection bias. Fourth, we assign a tweet to be pro/anti for a stance based solely on its content and employ an embedding-based method for classification. Again, more information from accounts, likes, followers, etc., and improved classifiers can be employed to see if further accuracy gains are observed.

Our modeling framework's ability to quantify participation bias can pave the way to understanding the mechanisms of political polarization [57], how it interacts with the algorithmic mechanisms of content visibility on Twitter [54], and with offline participation [55]. Additionally, once the participation bias is known, methods from survey science can be employed to re-weight the estimates to get population-level generalizations [58, 59].

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-023-00405-6.

Additional file 1. (PDF 516 kB)

Acknowledgements
We thank one anonymous reviewer who provided valuable feedback on the manuscript.

Funding
This work is supported in part by NSF Award 2242072 and the cybersecurity cluster programmatic fund from the Institute for Security, Technology and Society (ISTS). Society (ISTS), Dartmouth College, NH.

Abbreviations
SM, Social Media platform; NPR, National Public Radio; PBS, Public Broadcasting Service; Biq, Bias Quantification Model; BERT, Bidirectional Encoder Representations from Transformers; MAE, Mean Absolute Error.

Availability of data and materials
All data and code to generate the results described in the paper are given at:
https://github.com/neetip/participation-bias-sm.

Declarations

Competing interests
The authors declare that they have no competing interests.

Author contributions
NP, BAV, and SV designed the research and wrote the paper. NP performed research and analyzed data. All authors read and approved the final manuscript.

Author details
[1]Department of Computer Science, Dartmouth College, Hanover, NH, USA. [2]Department of Government, Dartmouth College, Hanover, NH, USA.

Received: 19 December 2022 Accepted: 13 July 2023 Published online: 28 July 2023

References

1. (2019) Foundations for evidence-based policymaking act of 2018. https://www.cio.gov/policies-and-priorities/evidence-based-policymaking/
2. (2021) A European strategy for data. https://digital-strategy.ec.europa.eu/en/policies/strategy-data
3. (2020) National policy development framework. https://www.gov.za/sites/default/files/gcis_document/202101/national-policy-development-framework-2020.pdf
4. Conrad F, Gagnon-Bartsch J, Ferg R, Schober M, Pasek J, Hou E (2019) Social media as an alternative to surveys of opinions about the economy. Soc Sci Comput Rev 39(4):489–508
5. Sen I, Flöck F, Weller K, Weiß B, Wagner C (2021) Applying a total error framework for digital traces to social media research. In: Handbook of computational social science, vol 2. Routledge, London, pp 127–139
6. Aiello AE, Renson A, Zivich PN (2020) Social media and Internet-based disease surveillance for public health. Annu Rev Public Health 41:101–118
7. Yousefinaghani S, Dara R, Poljak Z, Bernardo TM, Sharif S (2019) The assessment of Twitter's potential for outbreak detection: avian influenza case study. Sci Rep 9(1):1–17
8. Masri S, Jia J, Li C, Zhou G, Lee M-C, Yan G, Wu J (2019) Use of Twitter data to improve Zika virus surveillance in the United States during the 2016 epidemic. BMC Public Health 19(1):1–14
9. Zagheni E, Garimella VRK, Weber I, State B (2014) Inferring international and internal migration patterns from Twitter data. In: Proceedings of the 23rd ACM International Conference on World Wide Web, pp 439–444
10. Fiorio L, Abel G, Cai J, Zagheni E, Weber I, Vinué G (2017) Using Twitter data to estimate the relationship between short-term mobility and long-term migration. In: Proceedings of the 9th ACM web science conference, pp 103–110
11. Kim J, Sîrbu A, Giannotti F, Gabrielli L (2020) Digital footprints of international migration on Twitter. In: International symposium on intelligent data analysis. Springer, Berlin, pp 274–286
12. Barchiesi D, Moat HS, Alis C, Bishop S, Preis T (2015) Quantifying international travel flows using Flickr. PLoS ONE 10(7):0128470
13. Zagheni E, Weber I, Gummadi K (2017) Leveraging Facebook's advertising platform to monitor stocks of migrants. Popul Dev Rev 43(4):721–734
14. Pokhriyal N, Dara A, Valentino B, Vosoughi S (2020) Social media data reveals signal for public consumer perceptions. Proceedings of the ACM International Conference on AI in Finance
15. Pasek J, Yan HY, Conrad FG, Newport F, Marken S (2018) The stability of economic correlations over time: identifying conditions under which survey tracking polls and Twitter sentiment yield similar conclusions. Public Opin Q 82(3):470–492
16. Antenucci D, Cafarella M, Levenstein M, Ré C, Shapiro MD (2014) Using social media to measure labor market flows. National Bureau of Economic Research, Inc. NBER working papers
17. O'Connor B, Balasubramanyan R, Routledge BR, Smith NA (2010) From tweets to polls: linking text sentiment to public opinion time series. In: International conference on web and social, Media
18. Bovet A, Morone F, Makse HA (2018) Validation of Twitter opinion trends with national polling aggregates: Hillary Clinton vs Donald Trump. Sci Rep 8(1):1–16
19. Beauchamp N (2017) Predicting and interpolating state-level polls using Twitter textual data. Am J Polit Sci 61(2):490–503
20. Barberá P, Rivero G (2015) Understanding the political representativeness of Twitter users. Soc Sci Comput Rev 33(6):712–729
21. Tufekci Z (2014) Big questions for social media big data: representativeness, validity and other methodological pitfalls. In: Proceedings of 8th international AAAI conference on weblogs and social, Media
22. Ruths D, Pfeffer J (2014) Social media for large studies of behavior. Science 346(6213):1063–1064
23. Baeza-Yates R (2020) Biases on social media data: (keynote extended abstract). In: Companion proceedings of the web conference. WWW '20. Assoc. Comput. Mach., New York
24. Gayo-Avello D (2011) Don't turn social media into another literary digest poll. Commun ACM 54(10):121–128
25. Baeza-Yates R (2018) Bias on the web. Commun ACM 61(6):54–61
26. Hargittai E (2020) Potential biases in big data: omitted voices on social media. Soc Sci Comput Rev 38(1):10–24
27. Kim JW, Guess A, Nyhan B, Reifler J (2021) The distorting prism of social media: how self-selection and exposure to incivility fuel online comment toxicity. J Commun 71(6):922–946. https://doi.org/10.1093/joc/jqab034
28. (2019) Sizing up Twitter users. https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/
29. Ribeiro FN, Benevenuto F, Zagheni E (2020) How biased is the population of Facebook users? Comparing the demographics of Facebook users with census data to generate correction factors. In: 12th ACM conference on web science, pp 325–334
30. Nguyen D, Gravel R, Trieschnigg D, Meder T (2013) "how old do you think I am?" a study of language and age in Twitter. In: ICWSM

31. Pennacchiotti M, Popescu A-M (2011) A machine learning approach to twitter user classification. ICWSM 11
32. Vijayaraghavan P, Vosoughi S, Roy D (2017) Twitter demographic classification using deep multi-modal multi-task learning. In: Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: short papers). Assoc. Comput. Linguistics, Vancouver, pp 478–483. https://doi.org/10.18653/v1/P17-2076
33. Hamidi F, Scheuerman MK, Branham SM (2018) Gender recognition or gender reductionism? The social implications of embedded gender recognition systems. In: Proceedings of the 2018 ACM CHI conference on human factors in computing systems, pp 1–13
34. Raji ID, Gebru T, Mitchell M, Buolamwini J, Lee J, Denton E (2020) Saving face: investigating the ethical concerns of facial recognition auditing. In: Proceedings of the AAAI/ACM conference on AI, ethics, and society, pp 145–151
35. Buolamwini J, Gebru T (2018) Gender shades: intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency, pp 77–91. PMLR
36. Fosch-Villaronga E, Poulsen A, Søraa RA, Custers B (2021) Gendering algorithms in social media. ACM SIGKDD Explor Newsl 23(1):24–31. https://doi.org/10.1145/3468507.3468512
37. Hughes AG, McCabe SD, Hobbs WR, Remy E, Shah S, Lazer DMJ (2021) Using administrative records and survey data to construct samples of Tweeters and Tweets. Public Opin Q 85(S1):323–346. https://doi.org/10.1093/poq/nfab020
38. Grinberg N, Joseph K, Friedland L, Swire-Thompson B, Lazer D (2019) Fake news on Twitter during the 2016 U.S. presidential election. Science 363(6425):374–378. https://doi.org/10.1126/science.aau2706
39. Tillery AB (2019) What kind of movement is black lives matter? The view from Twitter. J Race Ethn Polit 4(2):297–323. https://doi.org/10.1017/rep.2019.17
40. Darwish K, Stefanov P, Aupetit M, Nakov P (2020) Unsupervised user stance detection on Twitter. In: Proceedings of the international AAAI conference on web and social media, vol 14, pp 141–152
41. Lyu H, Wang J, Wu W, Duong V, Zhang X, Dye TD, Luo J (2021) Social media study of public opinions on potential COVID-19 vaccines: informing dissent, disparities, and dissemination. Intell Med
42. Küçük D, Can F (2020) Stance detection: a survey. ACM Computing Surveys 53(1). https://doi.org/10.1145/3369026
43. Tokdar ST, Kass RE (2010) Importance sampling: a review. Wiley Interdiscip Rev: Comput Stat 2(1):54–60
44. (2019) NPR/PBS NewsHour/Marist Poll: february 2019 gun violence, 2019 [Dataset]. Roper #31116083, Version 2. Marist College Institute for Public Opinion [producer]. Cornell University, Ithaca, NY: Roper Center for Public Opinion Research [distributor]
45. (2019) NPR/PBS NewsHour/Marist Poll: september 2019 gun violence, 2019 [Dataset]. Roper #31116763, Version 1. Marist College Institute for Public Opinion [producer]. Cornell University, Ithaca, NY: Roper Center for Public Opinion Research [distributor]
46. Zheng X, Han J, Sun A (2018) A survey of location prediction on Twitter. IEEE Trans Knowl Data Eng 30(9):1652–1671. https://doi.org/10.1109/TKDE.2018.2807840
47. Reimers N, Gurevych I (2019) Sentence-bert: sentence embeddings using Siamese bert-networks. In: Proceedings of the 2019 conference on empirical methods in natural language processing. Assoc. Comput. Linguistics, Vancouver
48. Preoţiuc-Pietro D, Ungar L (2018) User-level race and ethnicity predictors from Twitter text. In: Proceedings of the 27th international conference on computational linguistics, pp 1534–1545
49. Wang Z, Hale S, Adelani DI, 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. Assoc. Comput. Mach., New York, pp 2056–2067. https://doi.org/10.1145/3308558.3313684
50. Preoţiuc-Pietro D, Liu Y, Hopkins D, Ungar L (2017) Beyond binary labels: political ideology prediction of Twitter users. In: Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers), pp 729–740
51. Cohen R, Ruths D (2013) Classifying political orientation on Twitter: it's not easy! In: ICWSM. AAAI Press, Menlo Park
52. Barberà P (2015) Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. Polit Anal 23(1):76–91. https://doi.org/10.1093/pan/mpu011
53. (2017) America's complex relationship with guns. https://www.pewresearch.org/social-trends/2017/06/22/americas-complex-relationship-with-guns/
54. Huszár F, Ktena SI, O'Brien C, Belli L, Schlaikjer A, Hardt M (2022) Algorithmic amplification of politics on Twitter. Proc Natl Acad Sci 119(1):2025334119. https://doi.org/10.1073/pnas.2025334119
55. Freelon D, Marwick A, Kreiss D (2020) False equivalencies: online activism from left to right. Science 369(6508):1197–1201. https://doi.org/10.1126/science.abb2428
56. Mukerjee S, Jaidka K, Lelkes Y (2022) The political landscape of the us twitterverse. Polit Commun 39(5):565–588
57. Bail CA, Argyle LP, Brown TW, Bumpus JP, Chen H, Hunzaker MBF, Lee J, Mann M, Merhout F, Volfovsky A (2018) Exposure to opposing views on social media can increase political polarization. Proc Natl Acad Sci 115(37):9216–9221. https://doi.org/10.1073/pnas.1804840115
58. Park DK, Gelman A, Bafumi J (2004) Bayesian multilevel estimation with poststratification: state-level estimates from national polls. Polit Anal 12(4):375–385
59. Little RJ (1993) Post-stratification: a modeler's perspective. J Am Stat Assoc 88(423):1001–1012

Publisher's Note