Faces in the Crowd: Twitter as Alternative to Protest Surveys

Christopher Barrie and Arun Frey *

October 30, 2021

Abstract

Who goes to protests? To answer this question, existing research has relied either on retrospective surveys of populations or in-protest surveys of participants. Both techniques are prohibitively costly and face logistical and methodological constraints. In this article, we investigate the possibility of surveying protests using Twitter. We propose two techniques for sampling protestors on the ground from digital traces and estimate the demographic and ideological composition of ten protestor crowds using multidimensional scaling and machine-learning techniques. We test the accuracy of our estimates by comparing to two in-protest surveys from the 2017 Women's March in Washington, D.C. Results show that our Twitter sampling techniques are superior to hashtag sampling alone. They also approximate the ideology and gender distributions derived from on-the-ground surveys, albeit with some bias, but fail to retrieve accurate age group estimates. We conclude that online samples are yet unable to provide reliable representative samples of offline protest.

*Both authors contributed equally. We thank Pablo Barberá, Michael Biggs, Neil Ketchley, Joshua Tucker, and Megan Metzger for comments on versions of this paper, Dana Fisher and Michael T. Heaney for making available survey replication data, as well as audiences at the 2020 American Political Science Association Online Conference and Nuffield Online Sociology Seminar. Christopher Barrie is Lecturer in Computational Sociology at the University of Edinburgh; Arun Frey is a DPhil Candidate in Sociology at the University of Oxford. Correspondence to: christopher.barrie@ed.ac.uk.

1 Introduction

Writing at the close of the revolutionary Nineteenth Century ([1896] 1947, 15), Gustave Le Bon saw a French society undergoing transition. And among "the most striking characteristics of our epoch of transition," he wrote, was the entry of the crowd into politics. But how to understand crowds? For his part, Le Bon claimed to find some element of "mental unity" among crowd members ([1896] 1947, 23). Unwilling to cast the crowd as a singular entity, George Rudé ([1964] 2005), would later set out to identify the "faces in the crowd," and to give names and significance to individual crowd members ([1964] 2005, 205). Giving life to individual crowd members, however, was a serious undertaking. This was because "participants... rarely leave records of their own," meaning the historian had to play archaeologist of revolt, piecing together whatever documentary evidence remained (Rudé, [1964] 2005, 12). Often, even where they were available, such records would not survive the impassioned context of their creation: the French National Archives, founded in 1790 to prevent the revolutionary destruction of public records, were later set aflame in the last weeks of the 1871 Paris Commune.

Subsequent generations of scholars have relied upon general population surveys to make inferences about participants in protesting crowds. Three problems often accompany this approach, which we summarize as: 1) question generality, 2) small positive n, and 3) preference falsification. General population surveys sometimes ask questions about protest participation but questions are often too vague to learn about the correlates of participants in particular protests. Surveys fielded after major protest events that do target particular protests commonly capture only a tiny fraction of actual participants. They are also limited by two types of response bias: to take one example, when a mass mobilization event is successful, respondents are likely to claim participation as the socially desirable response, even if they did not participate (Opp and Gern, 1993); conversely, when such mobilization fails, or participants are mobilizing counter to the initial protests (i.e., are counter-revolutionaries in the context of revolution), respondents may choose not to disclose participation for fear of repression or retribution (Beissinger, 2013; Kalinin and de Vogel, 2016).

The other option is to survey protesters in the field with in-protest surveys. This technique faces three further problems, summarized as: 1) sample selectivity, 2) non-response bias, and 3) logistical constraints. In-protest surveys select on the dependent variable, making it difficult to arrive at larger population-level inferences, and are undermined by considerable problems of non-response (Walgrave et al., 2016). What is more, the protest cascades that precipitate uprisings are not flagged in advance and often come as a surprise to participants and onlookers alike (Kuran, 1995). As a consequence, researchers rarely have the time to organize survey questionnaires, gain clearance from institutional review boards, and hire interviewers before streets once again empty. Finally, both general population and in-protest surveys pose financial costs that are prohibitive for most researchers without sources of external funding.

Given these constraints, and the increasing visibility of protest and dissent online, scholars have innovated by using social media as sources of information. Most often, social media researchers will sample data by the identifying hashtag associated with a protest or campaign (Barberá et al., 2015b; Steinert-Threlkeld, 2017). We do not know, however, if users who share information about a protest online have the same ideological outlook or basic attributes as offline protestors.

In what follows, we set out two techniques for the identification of protestors on the ground sampled from their online traces. We implement this technique on a sample of individuals tweeting about the Women's March—a series of protest events held across multiple cities in the USA in the first month of 2017 to advance women's rights and protest the presidency of Donald J. Trump (Berry and Chenoweth, 2018).[1] We first identify protesters on Twitter by locating individual Twitter users on march routes across ten US cities on the day of the protest. Using multidimensional scaling and machine-learning methods, we then estimate the ideological preferences of Twitter protestor-users, as well as their basic

[1] The full "Mission & Vision" statement of the march, outlining its uniting principles and goals, can be found here: https://web.archive.org/web/20170125034852/https://www.womensmarch.com/mission/.

demographic characteristics. For the largest of these marches—in Washington, D.C.—we benchmark our demographic and ideology estimates against those from two in-protest surveys, as well as against estimates from a random sample of `#WomensMarch` hashtag users on the day of the protest.[2] Finally, given the difficulties of obtaining a sufficiently large sample of geolocated users, we test the accuracy of a second technique for obtaining a protestor sample by manually coding photos shared by Twitter users in Washington D.C. on the day of the protests.

Our contribution is twofold. First, by elaborating techniques reliably to identify protestors on the ground from social media, we significantly improve on existing approaches that monitor only movement-specific hashtags. We show that by using this method we are able, with greater accuracy, to capture the ideological and demographic attributes of protestor crowds. Second, we evaluate these improved identification techniques by comparing them to benchmark data from protestors surveyed at protest marches. Here, we show that despite the improvements of our proposed technique, protestors who share information online still differ in systematic ways from the average protestor on the ground. Future research should build on our proposals for identifying protestors from their online traces, which represent an obvious advance on sampling by hashtag alone. In turn, the viability of surveying protestors from digital traces alone will depend on future levels of connectivity and further advances in the automated inference of online user attributes. Taken together, our results at once provide avenues for further research and reason to be cautious when inferring movement information on the basis of digital traces alone.

2 Surveying protest

Social movements and collective action constitute core fields within sociological research. And to pursue research in the field, scholars have made extensive use of both in-protest and

[2] For participation estimates, see the website of the "Crowd Counting Consortium" here: https://sites.google.com/view/crowdcountingconsortium/home.

retrospective surveys to understand the correlates of participation.

A first approach to gauging the correlates of participation involves using population surveys to capture both protestors and non-protestors in the sampling frame. Typically, such surveys are intended to be nationally representative. An early example is the work of Barnes and Kaase (1979) who used population surveys to study attitudes toward protest across five Western democracies. Questions on protest participation have more recently been included in major cross-national surveys like the World Values Survey (WVS). Unfortunately, these questions are generally unspecific and therefore cannot accurately identify which type of protest the individual took part in or when it took place (Biggs, 2015).

When a particular protest event *is* targeted within the survey design, researchers are often faced with the problem of a small positive n. By way of example, Wave II of the Arab Barometer surveys included questions on participation in the 2010-11 Arab Spring protests in Egypt and Tunisia—two large-scale mass-mobilization episodes. Despite the size of the Egyptian Revolution, only 8% of respondents (n=97) reported participating (Beissinger et al., 2015). Other examples—e.g., Beissinger (2013); Opp and Gern (1993)—do have a relatively large positive n.[3] But the "true" number of participants will often be smaller than the survey estimates: when mass mobilization events such as these are successful, asking retrospective questions about participation is subject to potential bias due to the "hero effect," whereby individuals claim participation despite the reality of their non-involvement (Opp and Gern, 1993). Beissinger (2013), for example, reports participation of 18.6% in the 2004 Orange Revolution in Ukraine; which would amount to 7.4 million people.[4] This estimate would make the event one of the largest mass mobilizations in world history. Further, this bias runs in both directions. In the same study of Ukrainian protestors, (Beissinger, 2013, 580) notes that "the number of counter-revolutionaries was likely twice as large as the [survey] indicated," since those protesting against the mood of the crowd are less likely to disclose

[3] The case of Beissinger (2013), however, relied on a regular survey being fielded at the time of protest outbreak—the kind of chance coincidence on which researchers cannot rely.

[4] 7.4m=40m*.186. Based on population estimate (of 15+ population) of 40 million in Ukraine in 2005, available at https://population.un.org/wpp/Download/Standard/Population/.

their true preferences.

The other survey tool available to researchers is the in-protest survey. To date, the most ambitious project to use these methods has been "Caught in the Act of Protest: Contextualizing Contestation" (CCC) (Klandermans et al., 2010), an effort by researchers across Europe to understand the sociological underpinnings of protest through in-protest surveys at some ninety-two protest events across seven European countries (Klandermans et al., 2014; Walgrave and Wouters, 2014). For the deployment of these instruments, researchers have also elaborated sophisticated random walk sampling frameworks to ensure the representativeness of the protestor sample (Fisher et al., 2005; Klandermans et al., 2010).

There are nonetheless several problems inherent to in-protest surveys. Most obviously, this method samples on the dependent variable, excluding non-protestors by design.[5] What is more, conducting in-protest surveys poses another set of challenges. The collection of protest data can be (literally) noisy: in nearly half of the protest surveys they carried out, interviewers in the CCC Project reported having difficulty hearing their interviewees; in one fifth of cases, interviewers reported difficulty given the chaotic nature of the demonstration, leading to increased non-response (Walgrave et al., 2016). Delayed refusal caused by individuals not returning postal questionnaires was even more pronounced, leading these authors to conclude that "noncooperation is a serious problem in protest surveying" (Walgrave et al., 2016, 93). Perhaps the biggest threat to this design, however, is the unpredictable nature of protest. Large-scale protest has a habit of breaking out all of a sudden (Kuran, 1995). This unpredictability necessarily confounds efforts to field survey teams at unexpected protest—for all protests covered in the CCC Project, protest organizers and police were contacted at least two weeks in advance of any action (Klandermans et al., 2010).

Against this backdrop, and the increasing visibility of protest on social media platforms, researchers have more recently started using digital trace data for the study of protest.

[5] Rosenfeld (2017) provides, to date, the only way out of this impasse, suggesting a version of the contaminated case-control design as a solution. This method relies on the availability of a random sample of "controls" (the population of interest) with which to twin "cases" (the selective protestor sample). Only in rare cases do there exist baseline survey-based estimates with comparable covariate profiles, however.

The most common platform for this research, given both its accessibility and popularity for campaigning, is the micro-blogging service Twitter. Researchers in this area have used Twitter data to study the dynamics of protest movement mobilization (Barberá et al., 2015b; Steinert-Threlkeld, 2017), recruitment (González-Bailón et al., 2011), polarization (Borge-Holthoefer et al., 2015), and change (Conover et al., 2013). Two problems attend this research. First, using samples derived from online platforms can provide insights into *online* mobilization dynamics but the generalizability of these insights to the *offline* world remains conjectural. As Steinert-Threlkeld (2017, 400) writes in his analysis of mobilization dynamics during the 2011 Arab Spring: "[the] article *assumes* that behavior on online networks parallels that of offline interpersonal ones" [emphasis added]. Similarly, given that both González-Bailón et al. (2011) and Barberá et al. (2015b) rely on online samples alone, they are naturally able to suggest only that their findings might inform theoretical models of (offline) collective action. Second, different sampling techniques may yield different results. Most often, to arrive at their sample, practitioners will filter on a set of hashtags related to the given protest campaign.[6] But as some of the same practitioners have noted, different filtering techniques can generate very different samples when studying online protest communication (González-Bailón et al., 2014). Rafail (2018) demonstrates in the case of the Occupy Wall Street (OWS) campaign, for example, that sampling on hashtag alone misrepresents the online network structure of the OWS movement, and underrepresents online mobilization activity.[7] In summary: existing research has taken samples from online sources to make important insights about the dynamics of collective action. However, the question of whether samples sourced online correspond to the characteristics of offline samples has yet to be examined.

[6]This is the case for all of Borge-Holthoefer et al. (2015); González-Bailón et al. (2011); Steinert-Threlkeld (2017) and Conover et al. (2013) cited above.

[7]Of course, in the below, our starting point is also a "hashtag sample" but we go on to outline two different approaches for filtering these data to recover a sample of (offline) protestors on the ground.

3 Data and Method

To fill this gap, we conduct two principal tests. The first compares our proposed techniques for capturing the digital traces of actual offline protestors to samples of users filtered by hashtag use alone; the second compares our Twitter-based sample of protestors to estimates from two in-protest surveys. In this, we are able to determine: 1) whether our proposed technique represents an improvement on more crude estimates from hashtag samples alone; and 2) the accuracy of our Twitter-based estimates when compared to the data from in-protest surveys.

To to build our dataset of protestors, we use two datasets of more than 8.6m tweets related to the 2017 Women's March. The first is taken from Littman and Park (2017), which records tweets across several hashtags related to the Women's March[8]; the second is taken from Ruest (2017) and records tweets containing the hashtag `#Womensmarch`. The Littman data was collected over the period December 19, 2016 to January 23, 2017 and the Ruest data from January 21, 2017 to January 28, 2017. The first sample we draw from these data is a random sample 5000 users who used one of the identifying hashtags on January 21, 2017.[9] We call this our "Random" Twitter sample and use this as a benchmark against which to compare estimates derived from our proposed techniques for capturing actual protestors on the ground.

3.1 Obtaining a sample of protestors

Identifying protest participants from the online behaviour of Twitter users alone is challenging: Protests often spark online commentary from participants, supporters, news reporters, and opponents alike. Those using the hashtag of a given protest may therefore be any of: 1) actual participants on the ground; 2) online supporters only; 3) online opponents only; 4)

[8]The full list of seeding words and hashtags are: "`#WomensMarch`, `#WMW`, `#WMWArchiveProject`, `#WhyIMarch`, `#ActivistADay`, `#WomensMarchWednesday`, `#WMWArt`, `#MarchMusicMondays`, and `#WMWYouth`." The overwhelming majority of tweets were captured through the `#WomensMarch` and `#WhyIMarch` hashtags.

[9]Included in this sample were all users for whom we could recover ideology and demographic estimates.

online commentators only.

To identify users who were posting on Twitter from within the march, we begin by filtering the tweet dataset to tweets sent on the day of the event (January 21, 2017).[10] We then filtered these data again to only those tweets that include location information in order to obtain digital traces of actual participants on the ground. Since only a small fraction of all Twitter users enable the geolocation of their tweets, this step considerably reduces our sample size from 3.8m to 17,120 tweets. To further restrict this data to actual protestors, our technique locates individual users to within a buffer of the protest march route on the day of the protest. To do this, we first sourced maps of the protest routes for ten of the largest protests during the Women's March online. A full list of the maps and their (archived) sources are in Appendix Table .1. Using the open-source geographic information systems software QGIS, these maps were georeferenced by locating landmarks and assigning relevant coordinates against reference coordinates from Open Street Map vector layers.[11] Using this technique, we were able to obtain samples of protestors across all ten US cities. Inclusion in these samples relied on the user tweeting about the Women's March from within a 1km buffer of the march route on the day of the protest. Of all 17,120 tweets for which location data was available, we identified 2,569 unique users whose tweet(s) located them at one of the protest marches. Appendix Figure .7 provides a visualization of the end result of this process. We refer to this sample as our "Geolocated" Twitter sample.

[10]Our analysis began two years after the Women's March. Although the original Tweet ID datasets by Ruest (2017) and Littman and Park (2017) contained ~14.4 and ~7.2m tweets respectively, only around half could be recovered for each source likely due to either account deletion, tweet deletion, or user removal by Twitter. The latter is the least concerning for our purposes as removed accounts will be mostly bots. While we cannot be sure of the magnitude of bias introduced by the omission of users, we see no obvious reason for account or tweet deletion to introduce bias along demographic or ideological dimensions. It is possible that our Geolocated sample would have included ideological opponents to the movement in the vicinity of the protest who subsequently deleted tweets, either because they did not want to be associated with a minority movement or otherwise. Our Photo-coded sample screens for opponents and so would have removed these accounts, had they remained in the sample. Where such bias would have affected findings is in the Random (hashtag) sample, for which inclusion is based on hashtag use alone. Here, subsequent tweet deletions by more conservative users may have skewed the ideology distribution leftwards. While we cannot determine the size of this possible bias, it does provide further support for our argument the hashtag sampling alone is unlikely to recover a close approximation of offline protestor ideology and demographics.

[11]This requires one of the (open-access) QGIS OSM Plugins.

Here, is worth noting that by using geolocation as our sole inclusion criterion, we do not exclude potential commentators who are reporting from within the protest (i.e., journalists as opposed to protestors on the ground). In the Appendix we discuss the size of any potential bias caused by their inclusion. We first calculate the percentage of users in our geolocated samples who are "verified"—an indication that a user may be a journalist or news organization in protest contexts—and then manually label a random subsample of our Washington D.C. geolocated tweets as "commentators" or "opponents." The percentage of users who are verified ranges from 0-7% across our ten cities. The percentage of tweets by commentators (rather than protestors) is ∼4% in our random Washington D.C. subsample; the percentage of tweets by opponents is .2%. Whether or not such individuals, who are "caught up" in a protest, satisfy inclusion criteria will depend on the research question at hand. In any case, exclusion of these accounts, on the basis of their verification status or (in the case of the Washington D.C. protest) manual codings, does not substantively alter our findings. As we detail below, we also evaluate a second, photo-coding, procedure for identifying protestors on the ground (where we screen for and exclude opponents and commentators) and are able to compare the findings from this approach to our results from the geolocation procedure for the Washington D.C. Women's March.

3.2 Obtaining ideology estimates of protesters

For both our Random and Geolocated samples we then estimate for each user their position on an ideology scale using a novel method originally developed by Barberá (2015), which computes ideology estimates of Twitter users by examining which political actors they follow (in Twitter parlance, their "friends"). This technique is broadly analogous to other multidimensional scaling techniques used to estimate the ideological leanings of individual legislators from roll call data (Poole and Rosenthal, 1985). However, in the place of voting, Barberá demonstrates that practitioners can leverage information on the friends of individual users to estimate their ideological position on a latent underlying dimension.

$$\begin{pmatrix} - & \texttt{@ABC}_1 & \texttt{@BarackObama}_2 & \texttt{@CynthiaLummis}_3 & \ldots & \texttt{@zeitgeist2o12}_{1186} \\ \texttt{@WomensMarcher}_1 & 0 & 1 & 0 & & 1 \end{pmatrix}$$

$$\begin{pmatrix} - & \texttt{@ABC}_1 & \texttt{@BarackObama}_2 & \texttt{@CynthiaLummis}_3 & \ldots & \texttt{@zeitgeist2o12}_{1186} \\ \texttt{@WomensMarcher}_2 & 1 & 0 & 0 & & 1 \end{pmatrix}$$

Figure 1: Example adjacency matrices

At its core, this estimation relies on the assumption that a user, given a set of otherwise similar political Twitter accounts with varying ideological beliefs, will prefer to follow those accounts that closely match her own ideological position. This is because the decision to follow a political account is costly: following a Twitter user entails the opportunity cost of not being exposed to alternative sources of information, and may induce cognitive dissonance if that information is at odds with one's own ideological outlook (Barberá, 2015).

Several multidimensional scaling techniques, including ideal point estimation and correspondence analysis, are suitable for estimating the ideology scores of individual users (Barberá et al., 2015a). In this article, we use a correspondence analysis procedure, since it gives effectively the same results as the Bayesian ideal point technique outline in Barberá (2015) while being computationally more efficient (Barberá et al., 2015a).

To estimate the ideology scores of our Random and Geolocated users, we begin by downloading the friends of each user using the Twitter REST API with the `rtweet` R package by Kearney (2019). We then follow the procedure set out by Barberá et al. (2015a), using the `R` package "`tweetscores`". This package includes a pre-specified list of US "elites" from politics and news media spanning a liberal-conservative dimension.[12] We then estimate individual user ideology scores by first arraying a sparse adjacency matrix of individual protestor user (rows) and elite friends (columns) as in Figure 1.

It is then possible to project each individual user matrix u back onto the latent ideological space already estimated by first taking the vector of the standardized residuals $u' = \frac{u}{\sum_i ui}$ for

[12] n=1,186, see Barberá et al. (2015a) for full list of accounts

each supplementary user then calculating the location of the new user on the latent ideological space $g = u'^T c$, where c represents the vector of column coordinates for individual political elites.[13] The estimation of users ideology score relies on her following network. Thus, if a user follows no elite accounts, their ideology score cannot be computed. For the Geolocated sample, this is the case for 111 observations, or 4.3% of the sample.[14]

The estimation procedure also accounts for "user- and elite-random effects" by including parameters for the political interest of user i (number of elites they follow) and the popularity elite j (number of followers of elite). The former acts as a proxy for the political interest of the user (i.e., a user may follow many accounts because they are simply interested in politics) and the latter accounts for the fact that a user may follow popular Twitter accounts (e.g. Barack Obama) simply due to their high profile and general relevance rather than as a function of ideological proximity (see supplementary material Barberá et al. 2015a and Barberá 2015) .[15]

3.3 Obtaining demographic estimates of protestors

We next supplement our ideology estimates by inferring basic demographic information from the Twitter profiles of individual users (Wang et al., 2019). Wang et al. (2019) propose a deep learning system that assigns each Twitter profile a probability of being male or female and belonging to a specific age group (≤18, 19–29, 30–39, 40+). To infer users' sex and age group, Wang et al. (2019) relies on four sources of information from Twitter: the username, screen name, biography, and profile image of each user. Each of these sources of information

[13] The "`tweetscores`" package is able efficiently to add users (or rows) to a correspondence analysis procedure without re-estimating the entire correspondence analysis. It does so by projecting the row coordinates of the new user onto the already-estimated latent ideological space by taking the row coordinates of the new user and looking up the corresponding column coordinates from a pre-estimated set of representative values. When the row coordinate does not have an exact match in this pre-estimated list of corresponding column coordinates, the function takes the closest corresponding column coordinate value and adds a value from a random normal distribution with mean 0 and standard deviation .05. This is why the estimated ideology score of each user will randomly vary by a small amount on each estimation.

[14] We describe the reasons for different types of missingness in more detail in Section 5 of the Appendix.

[15] We provide descriptive statistics on the number of elite accounts followed by users across the samples in Section 5 of the Appendix.

is evaluated using a separately trained text- or image-based neural model, before being combined for classification into a shared pipeline. Combined text and image information for each user is then classified with using the "`m3inference`" library in Python.[16] This estimation technique is preferable as it does not rely on large quantities of text produced by any individual user in order to generate demographic estimates, thus lowering computational costs. Despite its sparse input, the M3 model significantly outperforms state-of-the art techniques for inferring age and gender from image and text data.[17] By not relying on text output, it is also scaleable to multiple languages other than English.[18] We use this information to estimate the demographic composition of our sample. We were unable to recover demographic information for 148 users, or 5.8% of the sample. After removing the missing values for both ideology and demographic estimates, the Geolocated sample includes 2,319 unique users.

3.4 Alternative sampling procedure

Only a very small subset of users provide precise geolocation coordinates. This is one reason that research to date has opted to use alternative location information to estimate protestor crowd size (Sobolev et al., 2020). Recognising this constraint, we elaborated a second sampling procedure to capture protestors on the ground from their online traces. This second approach makes use of information contained in photographs shared by Twitter users. Given that the march in Washington D.C. saw the highest participation and we have in-protest survey evidence against which to compare our estimates, we only carry out this technique for Washington D.C. tweeters. To obtain a sample of protestors we first filtered our tweet dataset to users who posted original photographs and whose location ("Place") mentioned

[16] See `https://pypi.org/project/m3inference/` for details.

[17] Including "Face+++" (Jung et al., 2017), "Microsoft Face API" (Azure, 2018), "genderperformr" (Wang and Jurgens, 2018), "demographer" (Knowles et al., 2016), and Jaech and Ostendorf (2015).

[18] Wang et al. (2019) demonstrate that gender predictions are accurate across thirty-two separate languages, while there is a drop off in accuracy for age in some less widely used languages.

the city of Washington D.C, leaving 2,750 tweets.[19] We code a user as having participated in the protest if: a) the photo was taken from within the protest crowd during the Women's March in Washington D.C.; and b) if the image and accompanying text indicated protest attendance. We exclude tweets indicating news reporting rather than actual participation, as well as photos that could be stock images. We include in the Appendix of this article the full criteria that we used during the coding process. Each author independently coded half of the photographs dataset ($\sim$1300 tweets containing photographs) and jointly coded a subset of 200 photograph tweets. A comparison of our respective codings generated an inter-coder reliability Cohen's Kappa score of 0.8, indicating substantial agreement. Of the 2,750 tweets that included original imagery, 1,125 were coded as having been taken by protest participants. With this photo sample, we then repeated the same steps outlined above to generate ideology scores and estimates of crowd demographics. We refer to this sample as our "Photo-coded" sample. In total, we were unable to recover ideology estimates 201 users and demographic estimates for 49 users, resulting in a final sample size of 922.[20]

We summarize the entire workflow used to arrive at these estimations in Appendix Figure .8. The process detailed above results in three samples of Twitter users for whom we are able to recover ideological and demographic estimates. The first, Random sample of `#WomensMarch` hashtag users includes any user who posted with a relevant hashtag on the day of the protest; the second Geolocated sample includes any user identified on one of the protest routes across ten US cities; the third Photo-coded sample includes only users identified to the protest route in Washington, D.C.

Ethics

Before embarking on this research, we took account of a large number of ethical considerations. We summarize below what we determined on the basis of these considerations,

[19] Twitter aggregates location to a Twitter "Place." Twitter Places can refer to a specific place (like a stadium or monument) or an aggregate geographical location such as a city. For more information on Twitter Places, see https://developer.twitter.com/en/docs/tutorials/filtering-tweets-by-location.

[20] We describe the reasons for different types of missingness in more detail in Section 5 of the Appendix.

and detail in full the ethical framework according to which we approached this research in the Appendix. First, we gained authorization for this design from our institution's Central University Research Ethics Committee (Institutional Review Board equivalent).[21] We did not obtain informed consent from "participants" in this research as this was not deemed necessary. Consent is assumed as data is publicly available. Nonetheless, with a view to preserving contextual integrity Nissenbaum (2004) and user anonymity given the potential sensitivity of these data, we determined to: 1) elaborate an anonymization procedure prior to, and during, data ingestion to reduce any exposure to identifying information; 2) store all potentially identifying information locally on encrypted folders; 3) not to release tweet IDs of geolocated and photo-coded users in public replication folders.

4 Results and Validation

We first present the results from our geolocated users. Twitter-based estimates of crowd ideology distributions across ten US cities are depicted in Figure 2. We observe distributions centred to the left of ideological centre (depicted by a dashed grey line at 0). The distributions are very similar between cities, indicating a substantial degree of between-protest ideological homogeneity.

Our estimates of crowd demographics are displayed in Figures 3 and 4. Across most of our ten US cities, crowds are overwhelmingly female and tend to come, in the majority, from younger age groups. The exceptions are the cities of Portland and San Francisco, where the 30+ groups are in preponderance and there is almost gender parity. Both of these samples suffer from a very small n, however, and should therefore be treated with appropriate caution.

To scrutinise the validity of our results, we require a benchmark against which to compare them. These data are available for the Washington, D.C. Women's March where two in-

[21] The Reference Number of this Ethics Decision is: SOC_R2_001_C1A_20_16. The point of contact for this decision, and the individual to whom data requests may be sent is: Agnieszka Swiejkowska, DREC Secretary, Department of Sociology, University of Oxford, 42-43 Park End Street, Oxford, OX1 1JD. Email: research@sociology.ox.ac.uk; Tel.: + 44 1865 286177.

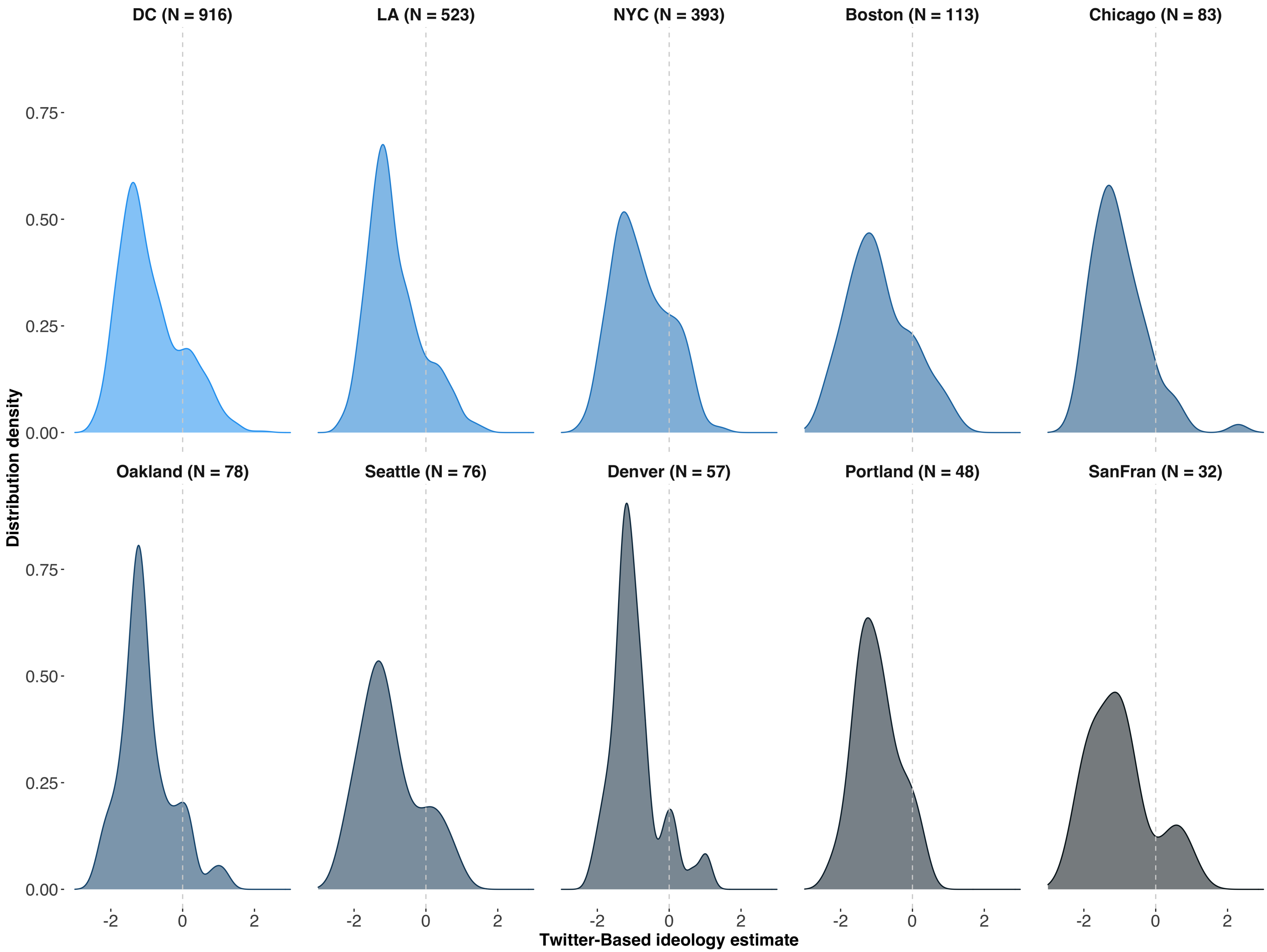


Figure 2: Distributions of aggregate crowd ideologies across ten cities in the 2017 US Women's Marches from geolocated users.

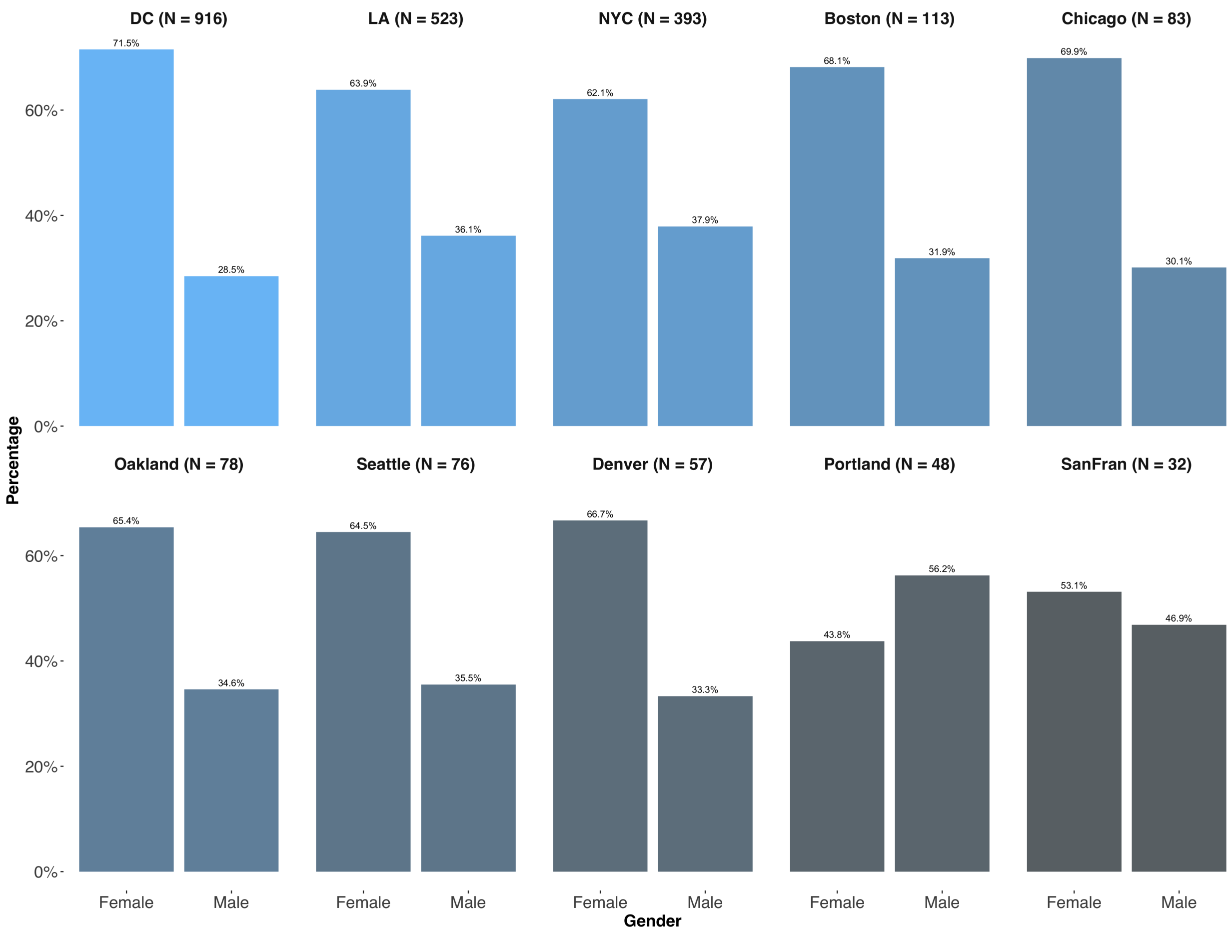


Figure 3: Gender distributions in the Twitter-based samples across ten cities in the 2017 US Women's Marches from geolocated users.

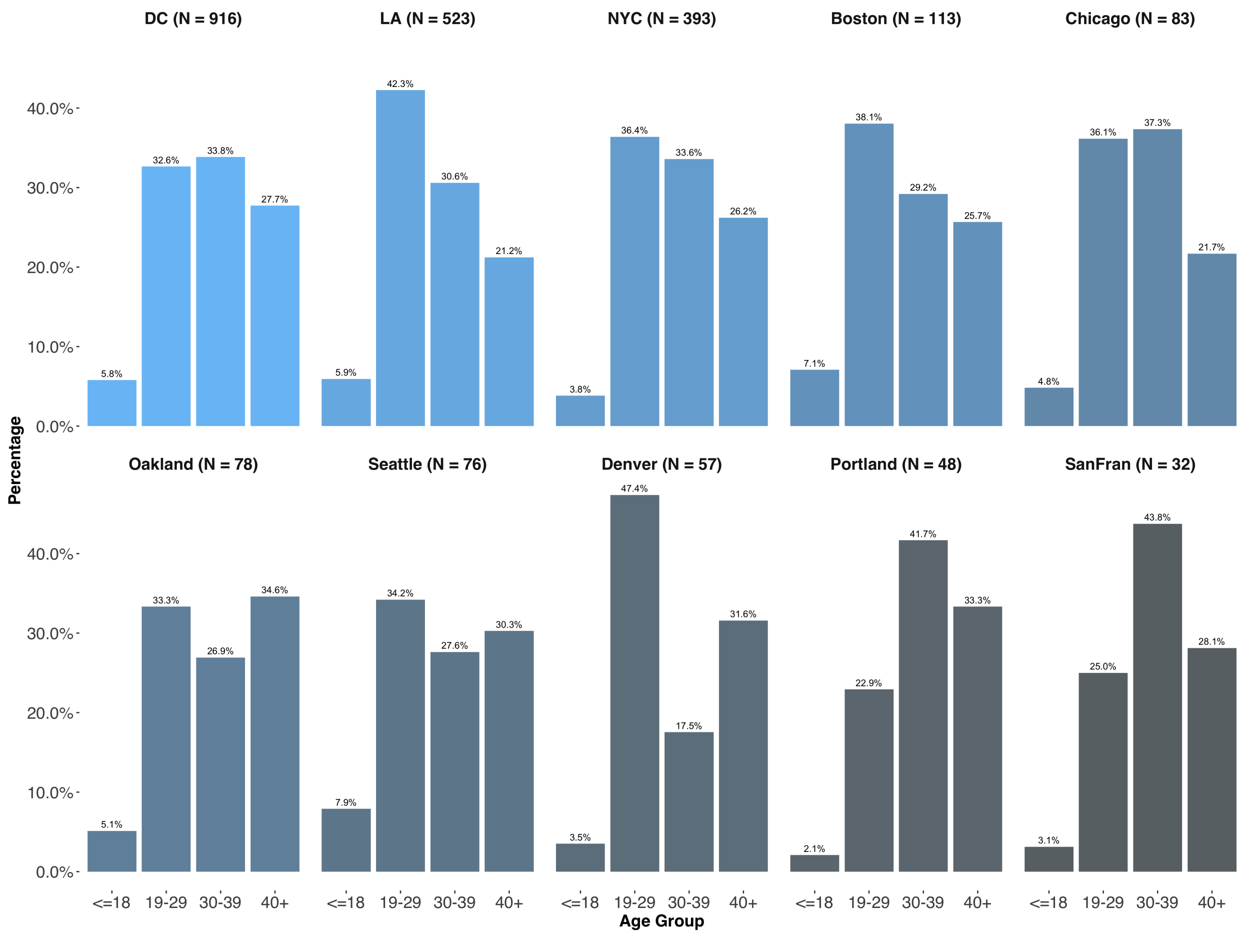


Figure 4: Age distributions in the Twitter-based samples across ten cities in the 2017 US Women's Marches from geolocated users.

protest surveys were conducted by Fisher et al. (2017) and Heaney (2018). We focus first on the ideology estimates and second on demographic estimates. For the first, the in-protest survey asks participants to place themselves on an ordinal ideological scale, from "Very Left" (1) to "Very Right" (7) in Fisher et al. (2017) and from "To the "left" of strong liberal" (1) to "To the "right" of strong conservative" (9) in Heaney (2018). While in-protest surveys are subject to their own biases, they nonetheless represent the gold standard for obtaining systematic data on protest participation. For this reason, we use these surveys as a benchmark for our own Twitter-based estimates.[22]

In addition to our protestor-users geolocated to Washington D.C., we now incorporate our two other Twitter samples for these comparisons. The first is our photo-coded sample of protestors at the march in D.C.; the second is our random sample of users filtered by hashtag alone. Note that the users in this second sample could be tweeting from any location and may or may not have attended the D.C. protest—inclusion was based solely on their having tweeted with the `#WomensMarch` hashtag on the day of the protests. We then compare the Twitter-based estimates of ideology distributions to survey results in Fisher et al. (2017) and Heaney (2018). We visualize the distributions of ideology scores for the in-protest survey and Twitter samples in the upper panel of Figure 5.[23]

The Twitter-based ideology estimates are already standardized to follow a normal distribution with mean 0 and standard deviation 1; that is, a user with score -1 is to be understood as one standard deviation to left of the "average" user (Barberá et al., 2015a). For the purposes of comparison, we centre the ideology scales of the in-protest surveys such that a score of 0 represents the middle category of each respective ordinal scale before standardizing by

[22]As we go on to describe below, these independent surveys also produced estimates for ideology and demographic distributions that closely correspond to each other. The refusal rates for both surveys were relatively low (7.5% for Fisher et al. (2017) and 20% for Heaney (2018)), and they both employed similar crowd sampling strategies. As such, we claim that these surveys constitute a valid and high quality point of comparison.

[23]We only use observations with complete records for the purposes of comparison. The number of observations for each sample therefore represents observations for which we have complete records for age, gender, and ideology.

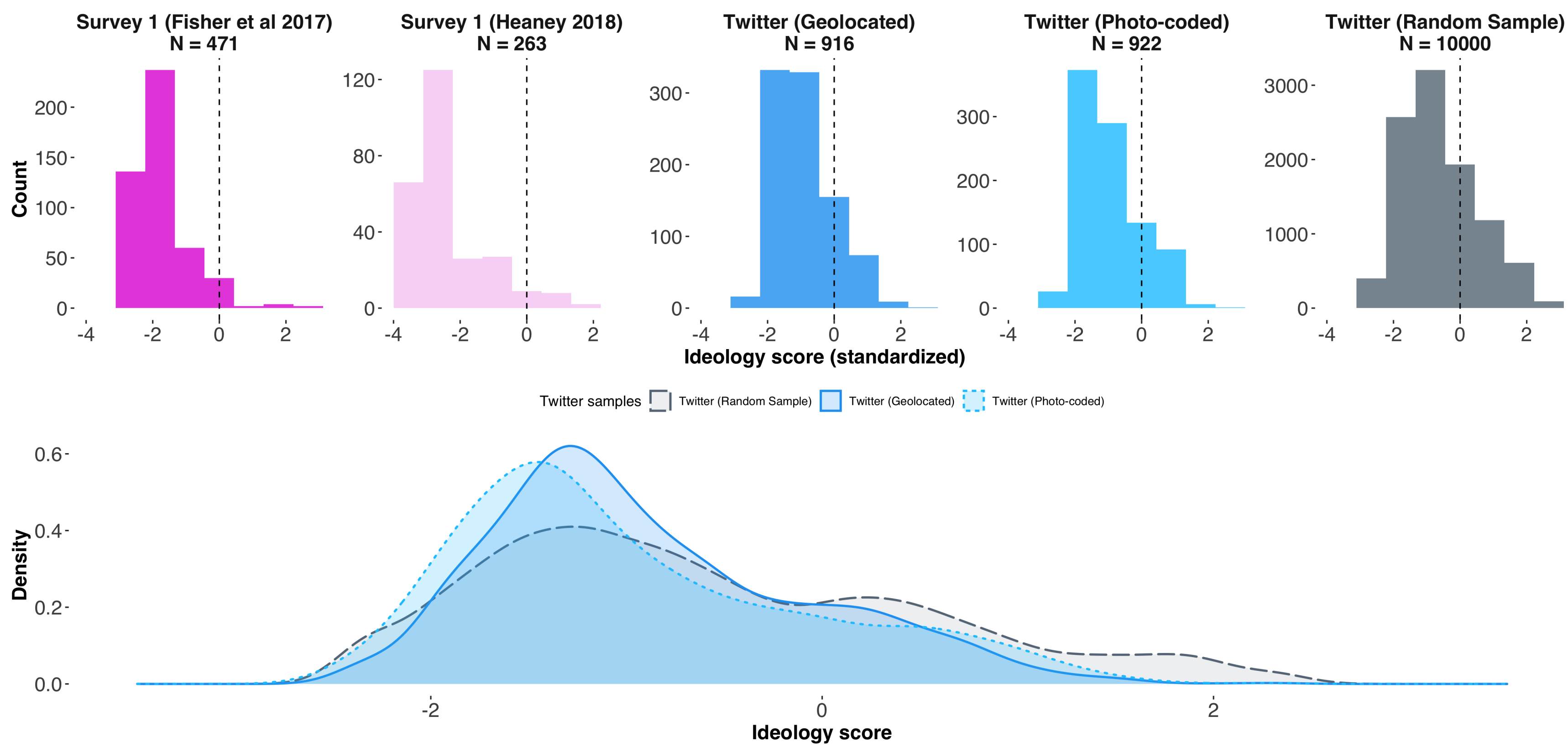


Figure 5: *Upper panel:*Ideology score distributions in survey- and Twitter-based samples for Washington, D.C.; *Lower panel:* Comparison of ideology distributions in protestor Twitter sample and random sample of all accounts using the hashtag `#WomensMarch`.

dividing by one standard deviation.[24] We see that individuals surveyed in-protest are relatively more left-wing than our Twitter-based geolocated and photo-coded samples. Our Twitter-based samples of identified protestors nonetheless do have ideology distributions that are similarly right-skewed, peak to the left of zero, and have negative modal values.

It is important to note that, despite the different sampling strategies, both Twitter samples of geolocated and photo-coded protestors provide highly similar estimates of crowd ideology. To assess whether this similarity is merely a feature of the underlying data, we compare our Twitter-based ideology estimates with the ideology estimates obtained from our random sample of hashtag users. In the lower panel of Figure 5 we overlay the ideology distributions for our geolocated and photo-coded users on the distribution for the random sample. We see that the estimates for those users we identify as protestors on march routes have greater density to to the left of zero than our estimates for the random sample. This is initial evidence that simply using hashtags to identify protestors is insufficient for capturing the ideology distributions of actual protestor crowds, and suggests that both geocoding and photo-coding methods identify similar users as protestors.[25]

Next, we compare the demographic estimates from our Twitter-based samples to those derived from the in-person surveys (Figure 6). We can see that in both the geolocated and photo-coded Twitter samples, similar to the in-person surveys, there is a preponderance of women making up the crowd. The geolocated and photo-coded Twitter-based samples are highly similar across both age and gender composition; compared to the in-protest surveys, however, they feature substantially more male participants, with male users making up 28.5% and 29.5% of the Twitter samples versus 14.1% and 16.2% respectively for the Fisher et al. (2017) and Heaney (2018) samples. Age differences between the online and in-protest samples are more pronounced. The modal age group in the geolocated and photo-coded

[24]The middle categories for each of the in-protest surveys are: [5]"moderate" in Heaney (2018) and [4] "Moderate, middle of the road" in Fisher et al. (2017).

[25]Note that only 46 users are in both the geolocated and photo-coded Twitter samples. This means the similarity between both samples is not due to considerable overlap in users who geolocated themselves at the protest march, and users who uploaded a tweet containing a photo from within the protest.

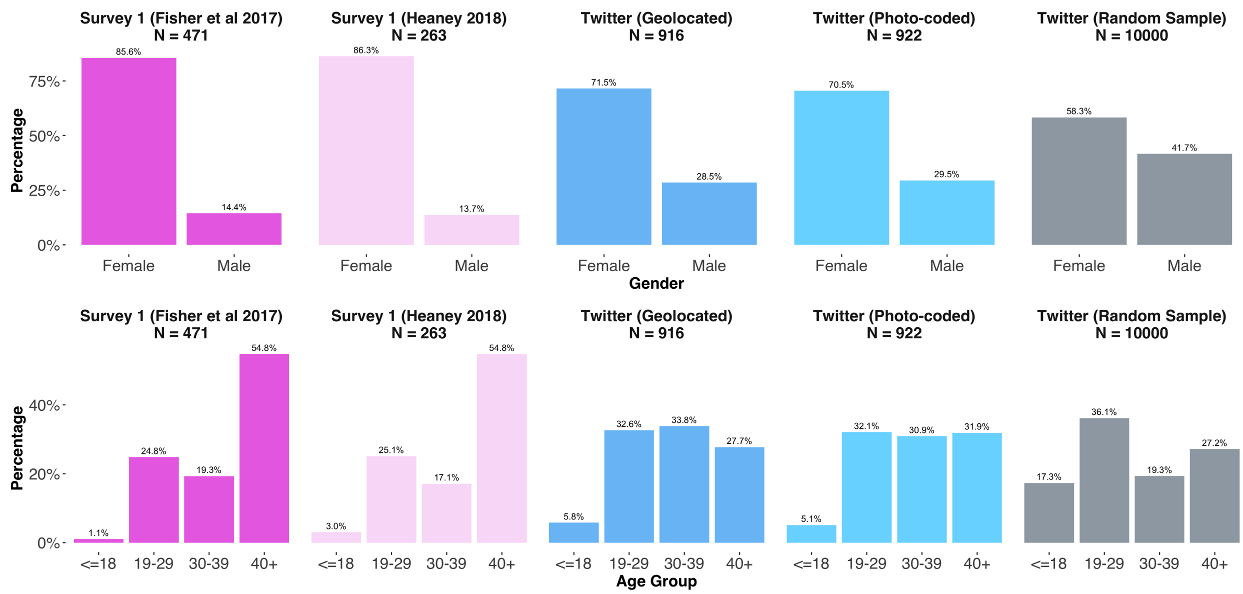


Figure 6: Gender and age distributions in survey- and Twitter-based samples for Washington, D.C.

Twitter samples is 19-29, for example, whereas for the in-protest sample it is the 40+ group. Still, across both age and gender, the geolocated and photo-coded samples of Twitter users look alike and improve on the gender estimates derived from the random Twitter sample.

5 Discussion and Conclusion

The use of digital trace data to make inferences about crowds has, to date, largely focused on the estimation of crowd size (Botta et al., 2015; Sobolev et al., 2020). This paper represents the first test of the viability of using online digital traces to estimate demographic and ideological characteristics of protestor crowds. For this, we rely on the availability of two in-protest surveys against which to compare the Twitter-based estimates. We find that Twitter can provide approximations of the ideological and gender composition of crowds but there remain considerable biases. Irrespective of sampling strategy, we are unable to recover accurate estimates of crowd age demographics.

What explains these differences? One explanation could be differences in the type of person likely to post about protest participation online. Geolocating to a particular event entails privacy costs, which digitally literate users may be less likely to accept. It is nonetheless worth noting that our geocoding and photo-coding techniques for identifying protestors do give very similar estimates despite the different sampling procedures used. This suggests that both methods do well to capture protestors on the ground who are also active online. Another explanation is difference in measurement context. It may be that in the politically charged environment of a protest, individuals are more likely to place themselves further to the extremes of an ideology scale than they otherwise would have. Alternatively, bias may result from the inferential procedure used to derive ideology scores from follow networks. Some individuals follow only a few relevant accounts, meaning their ideology scores can only be estimated with error. That said, removing accounts who only follow a few elite accounts does not substantively alter the distribution of ideology scores in our protestor crowds (see Section 5 in the Appendix).

In the case of age and gender, bias may result from measurement error in the automated procedure used to infer these demographic characteristics. Importantly, this measurement error may be systematic. For example, younger or more digitally literate users may be more likely to use an avatar in place of a photograph of themselves, thereby limiting the accuracy of algorithmic age and gender prediction. Then again, the large majority of our photo-coded sample used profile photos that did not appear to hide their real identity.

While we do not discount the above sources of potential bias, we suggest that the majority of the difference between our survey-based and Twitter-based demographic estimates most likely comes not from our sampling strategy or from classification error but from biases in the type of individual who is active on Twitter. After all, Twitter is not a representative sample of the general population. In the United States, the average Twitter user is younger, more likely to be male, and wealthier than the US populace (Blank, 2017). What is more, political discussions tend to feature men more than women and disproportionately include more educated users and users from urban or metropolitan areas (Barberá and Rivero, 2015). And the differences between our Twitter-based and survey-based demographic estimates map closely onto these sources of bias.

This notwithstanding, the findings do point to the potential future use of digital trace data as a source of information on the composition of protestor crowds. As connectivity and online platform usage increase over time, it is possible that these sources will become more representative of general populations (Pew, 2019). What is more, we know that, even if the average Twitter user has a different demographic profile compared to the general population, they are nonetheless very similar on various attitudinal measures (Wojcik and Hughes, 2019). This insight accords with our own findings above, which show that, as a source of information on aggregate ideological preferences, Twitter provides estimates that approximate those from surveys on the ground.

As for the viability of this method in other contexts, we are less optimistic. In many respects, the Women's March protests represent one of the most-likely cases for recovering representative samples of protest crowds from digital trace data. After all, these were very large protests in a democratic setting with high connectivity. In other contexts, low connectivity will likely mean insufficient sample sizes. Further, in non-democratic political contexts individuals may be less willing publicly to signal dissent online. A growing body of work is nonetheless making use of digital trace data, and Twitter in particular, for the study of movement campaigns outside of Western or liberal-democratic contexts (Budak and Watts, 2015; Kubinec and Owen, 2021; Pan and Siegel, 2020). Validating the offline representativeness of users sampled online will require benchmarking to in-protest surveys conducted in these contexts (e.g., Beissinger 2013; Berman 2019; Rosenfeld 2017; Tufekci and Wilson 2012.

Several limitations of the technique presented in this article do highlight possible avenues for extending and refining the approach. First, the technique we propose uses data from only one platform. For future implementations of the basic method, our technique is by no means limited to Twitter, however. Gathering information on the ideologies of users requires only that the researcher can access relevant information on the accounts followed by any given user.[26] As for collecting information on the gender and age of a user, this can be achieved using a neural architecture that relies only on limited user-level information, all of which would be accessible across diverse platforms.

Second, our method relies only on information that has been made publicly available by the user (i.e., their tweets, who they follow, their photo, user name, screen name, and account description). Naturally, this limits the amount of information the researcher is able to glean from any one individual. One future direction for the sampling method we outline would involve sending online questionnaires to sampled users. In order to shed further light on the sources of difference between offline and online samples, in-protest surveys might also ask for the Twitter handle of protestors. Researchers could then link the survey and Twitter data to determine the correlates of online presence and activity in the context of protest. These methods would likely encounter high refusal rates, however, and have associated privacy concerns (Al Baghal et al., 2020; Clark et al., 2019). We are also inferring age and gender algorithmically in the approach we outline, which entails measurement error—particularly for age (Han et al., 2015). An alternative would involve manual annotation by individual researchers, users themselves, or crowd-sourced online workers (see e.g., Huang et al. 2020).

Still, while our sample of online protestors is not representative of crowds on the ground, it does allow for within-platform comparisons. Digital trace data is "always on" (Salganik, 2018), enabling researchers to construct longitudinal panels after the initial sampling frame has been established (Budak and Watts, 2015). Differentiating between users who do and do not participate in protests also allows researchers to make use of a ready-made comparison group against which to benchmark their findings. Using digital traces to identify protest participants can thus help us understand how protestors' activity on social media platforms differs from other users, and can shed light on whether participation in a protest changes online behaviour over time.

[26] On Facebook, this is equivalent to an account "liking" the page of a particular prominent individual; Instagram and Sina Weibo have a following option very similar to Twitter; VKontakte provides information on the "Groups" and "Public pages" of which any given user is member; and on both TikTok and YouTube, the equivalent would be subscriptions.

Overall, this article provides a first validation test for using Twitter to "survey" protestors from afar. By locating users to the march route on the day of the protest, we identify protest participants on Twitter and compare their ideological and demographic composition to estimates from two separate in-protest surveys. Our method considerably improves on a random sample of all users tweeting about the `#WomensMarch`, and can recover an approximation of the ideological and demographic profile of protest crowds. Still, important differences remain between online and offline protestors: in line with general discrepancies between Twitter and the US populace, online protestors tend to feature a higher share of young and male participants. By signalling the capabilities and limitations of Twitter data for protest research, our results provide an important reference point for researchers wishing to study offline mobilization with online digital trace data.

References

Al Baghal, Tarek, Luke Sloan, Curtis Jessop, Matthew L. Williams, and Pete Burnap. 2020. "Linking Twitter and Survey Data: The Impact of Survey Mode and Demographics on Consent Rates Across Three UK Studies." *Social Science Computer Review* 38(5):517–532. doi:10.1177/0894439319828011.

Azure, Microsoft. 2018. "Microsoft Face API v1.0."

Barberá, Pablo. 2015. "Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data." *Political Analysis* 23(1):76–91. doi:10.1093/pan/mpu011.

Barberá, Pablo, John T. Jost, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. 2015a. "Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber?" *Psychological Science* 26(10):1531–1542. doi:10.1177/0956797615594620.

Barberá, Pablo, Ning Wang, Richard Bonneau, John T. Jost, Jonathan Nagler, Joshua Tucker, and Sandra González-Bailón. 2015b. "The Critical Periphery in the Growth of Social Protests." *PLOS ONE* 10(11):e0143611. doi:10.1371/journal.pone.0143611.

Barberá, Pablo and Gonzalo Rivero. 2015. "Understanding the political representativeness of twitter users." *Social Science Computer Review* 33(6):712–729. doi:10.1177/0894439314558836.
URL https://doi.org/10.1177/0894439314558836

Barnes, Samuel and Max Kaase. 1979. *Political Action, Mass Participation in Five Western Democracies*. London: Sage.

Beissinger, Mark R. 2013. "The semblance of democratic revolution: Coalitions in Ukraine's Orange revolution." *American Political Science Review* 107(3):574–592.

Beissinger, Mark R., Amaney A. Jamal, and Kevin Mazur. 2015. "Explaining Divergent Revolutionary Coalitions: Regime Strategies and the Structuring of Participation in the Tunisian and Egyptian Revolutions." *Comparative Politics* 48(1):1–24.

Berman, Chantal. 2019. "When Revolutionary Coalitions Break Down: Polarization, Protest, and the Tunisian Political Crisis of August 2013." *Middle East Law and Governance* 11(2):136–179. doi:10.1163/18763375-01102003.

Berry, Marie and Erica Chenoweth. 2018. "Who made the women's march?" In David S. Meyer and Sidney Tarrow, editors, "The Resistance: The Dawn of the Anti-Trump Opposition Movement," pages 75–89. Oxford: Oxford University Press.

Biggs, Michael. 2015. "Has protest increased since the 1970s? How a survey question can construct a spurious trend." *British Journal of Sociology* 66(1):141–162.

Blank, Grant. 2017. "The digital divide among twitter users and its implications for social research." *Social Science Computer Review* 35(6):679–697. doi:10.1177/0894439316671698. URL https://doi.org/10.1177/0894439316671698

Borge-Holthoefer, Javier, Walid Magdy, Kareem Darwish, and Ingmar Weber. 2015. "Content and Network Dynamics Behind Egyptian Political Polarization on Twitter." *CSCW '15: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* pages 1–12. doi:http://dx.doi.org/10.1145/2675133.2675163.

Botta, Federico, Helen Susannah Moat, and Tobias Preis. 2015. "Quantifying crowd size with mobile phone and *Twitter* data." *Royal Society Open Science* 2(5):150162. doi:10.1098/rsos.150162.

boyd, danah and Kate Crawford. 2012. "Critical Questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon." *Information, Communication & Society* 15(5):662–679. doi:10.1080/1369118X.2012.678878.

Budak, Ceren and Duncan Watts. 2015. "Dissecting the Spirit of Gezi: Influence vs. Selection in the Occupy Gezi Movement." *Sociological Science* 2:370–397. doi:10.15195/v2.a18.

Chancellor, Stevie, Michael L. Birnbaum, Eric D. Caine, Vincent M. B. Silenzio, and Munmun De Choudhury. 2019. "A Taxonomy of Ethical Tensions in Inferring Mental Health States from Social Media." In "Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19," pages 79–88. Atlanta, GA, USA: ACM Press. doi:10.1145/3287560.3287587.

Clark, Karin, Matt Duckham, Marilys Guillemin, Assunta Hunter, Jodie McVernon, Christine O'Keefe, Cathy Pitkin, Steven Prawer, Richard Sinnott, Deborah Warr, and Jenny Waycott. 2019. "Advancing the ethical use of digital data in human research: Challenges and strategies to promote ethical practice." *Ethics and Information Technology* 21(1):59–73. doi:10.1007/s10676-018-9490-4.

Conover, Michael D, Emilio Ferrara, Filippo Menczer, and Alessandro Flammini. 2013. "The Digital Evolution of Occupy Wall Street." *PLOS ONE* 8(5):5.

Conway, Mike. 2014. "Ethical Issues in Using Twitter for Public Health Surveillance and Research: Developing a Taxonomy of Ethical Concepts From the Research Literature." *Journal of Medical Internet Research* 16(12):e290. doi:10.2196/jmir.3617.

Fisher, Dana R., Dawn M. Dow, and Rashawn Ray. 2017. "Intersectionality takes it to the streets: Mobilizing across diverse interests for the women's march." *Science Advances* 3(9). doi:10.1126/sciadv.aao1390.
URL https://advances.sciencemag.org/content/3/9/eaao1390

Fisher, Dana R., Kevin Stanley, David Berman, and Gina Neff. 2005. "How Do Organizations Matter? Mobilization and Support for Participants at Five Globalization Protests." *Social Problems* 52(1):102–121. doi:10.1525/sp.2005.52.1.102.

González-Bailón, Sandra, Javier Borge-Holthoefer, Alejandro Rivero, and Yamir Moreno. 2011. "The Dynamics of Protest Recruitment through an Online Network." *Scientific Reports* 1(1):197.

González-Bailón, Sandra, Ning Wang, Alejandro Rivero, Javier Borge-Holthoefer, and Yamir Moreno. 2014. "Assessing the bias in samples of large online networks." *Social Networks* 38:16–27. doi:10.1016/j.socnet.2014.01.004.

Han, H., C. Otto, X. Liu, and A. K. Jain. 2015. "Demographic Estimation from Face Images: Human vs. Machine Performance." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37(6):1148–1161. doi:10.1109/TPAMI.2014.2362759.

Heaney, Michael T. 2018. "Making Protest Great Again." *Contexts* 17(1):42–47. doi:10.1177/1536504218766550.

Huang, Xiaolei, Linzi Xing, Franck Dernoncourt, and Michael J. Paul. 2020. "Multilingual Twitter Corpus and Baselines for Evaluating Demographic Bias in Hate Speech Recognition." *arXiv:2002.10361 [cs]* Comment: Accepted at LREC 2020.

Jaech, Aaron and Mari Ostendorf. 2015. "What Your Username Says About You." *arXiv:1507.02045 [cs]* .

Jung, Soon-Gyo, Jisun An, Haewoon Kwak, Joni Salminen, and Bernard J Jansen. 2017. "Inferring Social Media Users' Demographics from Profile Pictures: A Face++ Analysis on Twitter Users." In "Proceedings of The 17th International Conference on Electronic Business," pages 140–145. Dubai.

Kalinin, Kirill and Sasha de Vogel. 2016. "Measuring Propensity of Individual Anti-Government Protest Behavior in Autocracies." *SSRN Electronic Journal* doi:10.2139/ssrn.2767663.

Kearney, Michael W. 2019. "rtweet: Collecting and analyzing twitter data." *Journal of Open Source Software* 4(42):1829. doi:10.21105/joss.01829. R package version 0.7.0. URL https://joss.theoj.org/papers/10.21105/joss.01829

Klandermans, B., J. van Stekelenburg, M.-L. Damen, D. van Troost, and A. van Leeuwen. 2014. "Mobilization Without Organization: The Case of Unaffiliated Demonstrators." *European Sociological Review* 30(6):702–716. doi:10.1093/esr/jcu068.

Klandermans, Bert, Jacquelien van Stekelenburg, Dunya Van Troost, Anouk Van Leuween, Stefaan Walgrave, Jeroen van Laer, and Ruud Wouters. 2010. "Manual for Data Collection on Protest Demonstrations. Caught in the Act of Protest: Contextualizing Contestation (CCC)." Technical report, VU University and University of Antwerp., Amsterdam and Antwerp.

Knowles, Rebecca, Josh Carroll, and Mark Dredze. 2016. "Demographer: Extremely Simple Name Demographics." In "Proceedings of the First Workshop on NLP and Computational Social Science," pages 108–113. Austin, Texas: Association for Computational Linguistics. doi:10.18653/v1/W16-5614.

Kubinec, Robert and John Owen. 2021. "When Groups Fall Apart: Identifying Transnational Polarization during the Arab Uprisings." *Political Communication* page 36. doi:10.31235/osf.io/wykmj.

Kuran, Timur. 1995. "The Inevitability of Future Revolutionary Surprises." *American Journal of Sociology* 100(6):1528–1551.

Le Bon, Gustave. [1896] 1947. *The Crowd: A Study of the Popular Mind*. London: Ernest Benn Limited.

Littman, Justin and Soomin Park. 2017. "Women's March Tweet Ids." doi:10.7910/DVN/5ZVMOR.
URL https://doi.org/10.7910/DVN/5ZVMOR

Nissenbaum, Helen. 2004. "Privacy as Contextual Integrity." *Washington Law Review* 79:41.

Opp, Karl-Dieter and Christiane Gern. 1993. "Dissident Groups, Personal Networks, and Spontaneous Cooperation: The East German Revolution of 1989." *American Sociological Review* 58(5):659. doi:10.2307/2096280.

Pan, Jennifer and Alexandra A. Siegel. 2020. "How Saudi Crackdowns Fail to Silence Online Dissent." *American Political Science Review* 114(1):109–125. doi:10.1017/S0003055419000650.

Pew. 2019. "Demographics of Social Media Users and Adoption in the United States."

Poole, Keith T. and Howard Rosenthal. 1985. "A Spatial Model for Legislative Roll Call Analysis." *American Journal of Political Science* 29(2):357. doi:10.2307/2111172.

Rafail, Patrick. 2018. "Nonprobability Sampling and Twitter: Strategies for Semibounded and Bounded Populations." *Social Science Computer Review* 36(2):195–211.

Rosenfeld, Bryn. 2017. "Reevaluating the Middle-Class Protest Paradigm: A Case-Control Study of Democratic Protest Coalitions in Russia." *American Political Science Review* 111(4):637–652.

Rudé, George. [1964] 2005. *The Crowd in History.* London: Serif.

Ruest, Nick. 2017. "#WomensMarch tweets January 12-28, 2017." doi:10.5683/SP/ZEL1Q6. URL https://doi.org/10.5683/SP/ZEL1Q6

Salganik, Matthew J. 2018. *Bit by Bit: Social Research in the Digital Age.* Princeton: Princeton University Press.

shakti franzke, aline, Anja Bechmann, Michael Zimmer, Charles Ess, and Association of Internet Researchers . 2020. "Internet Research: Ethical Guidelines 3.0." Technical report, Association of Internet Researchers.

Sobolev, Anton, M. Keith Chen, Jungseock Joo, and Zachary C. Steinert-Threlkeld. 2020. "News and Geolocated Social Media Accurately Measure Protest Size Variation." *American Political Science Review* 114(4):1343–1351. doi:10.1017/S0003055420000295.

Steinert-Threlkeld, Zachary C. 2017. "Spontaneous Collective Action: Peripheral Mobilization During the Arab Spring." *American Political Science Review* 111(2):379–403. doi:10.1017/S0003055416000769.

Tufekci, Zeynep and Christopher Wilson. 2012. "Social Media and the Decision to Participate in Political Protest: Observations From Tahrir Square." *Journal of Communication* 62(2):363–379. doi:10.1111/j.1460-2466.2012.01629.x.

Walgrave, Stefaan and Ruud Wouters. 2014. "The Missing Link in the Diffusion of Protest: Asking Others." *American Journal of Sociology* 119(6):1670–1709. doi:10.1086/676853.

Walgrave, Stefaan, Ruud Wouters, and Pauline Ketelaars. 2016. "Response Problems in the Protest Survey Design: Evidence from Fifty-One Protest Events in Seven Countries *." *Mobilization: An International Quarterly* 21(1):83–104. doi:10.17813/1086/671X-21-1-83.

Wang, Zijian, Scott Hale, David Ifeoluwa Adelani, Przemyslaw Grabowicz, Timo Hartman, Fabian FlÃ¶ck, and David Jurgens. 2019. "Demographic inference and representative population estimates from multilingual social media data." In "The World Wide Web Conference," WWW '19, page 2056–2067. New York, NY, USA: Association for Computing Machinery. doi:10.1145/3308558.3313684.
URL https://doi.org/10.1145/3308558.3313684

Wang, Zijian and David Jurgens. 2018. "It's going to be okay: Measuring Access to Support in Online Communities." In "Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing," pages 33–45. Brussels.

Williams, Matthew L, Pete Burnap, and Luke Sloan. 2017. "Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users'

Views, Online Context and Algorithmic Estimation." *Sociology* 51(6):1149–1168. doi: 10.1177/0038038517708140.

Wojcik, Stefan and Adam Hughes. 2019. "U.S. adult Twitter users are younger and more likely to be Democrats than the general public. Most users rarely tweet, but the most prolific 10% create 80% of tweets from adult U.S. users." Technical report, Pew Research Center.

Zimmer, Michael. 2018. "Addressing Conceptual Gaps in Big Data Research Ethics: An Application of Contextual Integrity." *Social Media + Society* 4(2):205630511876830. doi: 10.1177/2056305118768300.

Appendix

Links to the original maps of march routes we used to create buffers and locate individuals to march routes are provided as web-archived links in Table .1. A visualization of our geolocation procedure to a 1km buffer of march routes is provided in Figure .7.

City	**Link**
DC	https://tinyurl.com/DCwmrmp
Boston	https://tinyurl.com/Bwmrmp
Chicago	https://tinyurl.com/CHIwmrmp
Denver	https://tinyurl.com/DEwmrmp
LA	https://tinyurl.com/LABAYwmrmp
NYC	https://tinyurl.com/NYCwmrmp
Oakland	https://tinyurl.com/OAKwmrmp
Portland	https://tinyurl.com/PRTwmrmp
Seattle	https://tinyurl.com/SEwmrmp
SF	https://tinyurl.com/SFwmrmp

Table .1: Women's March original map archived sources.

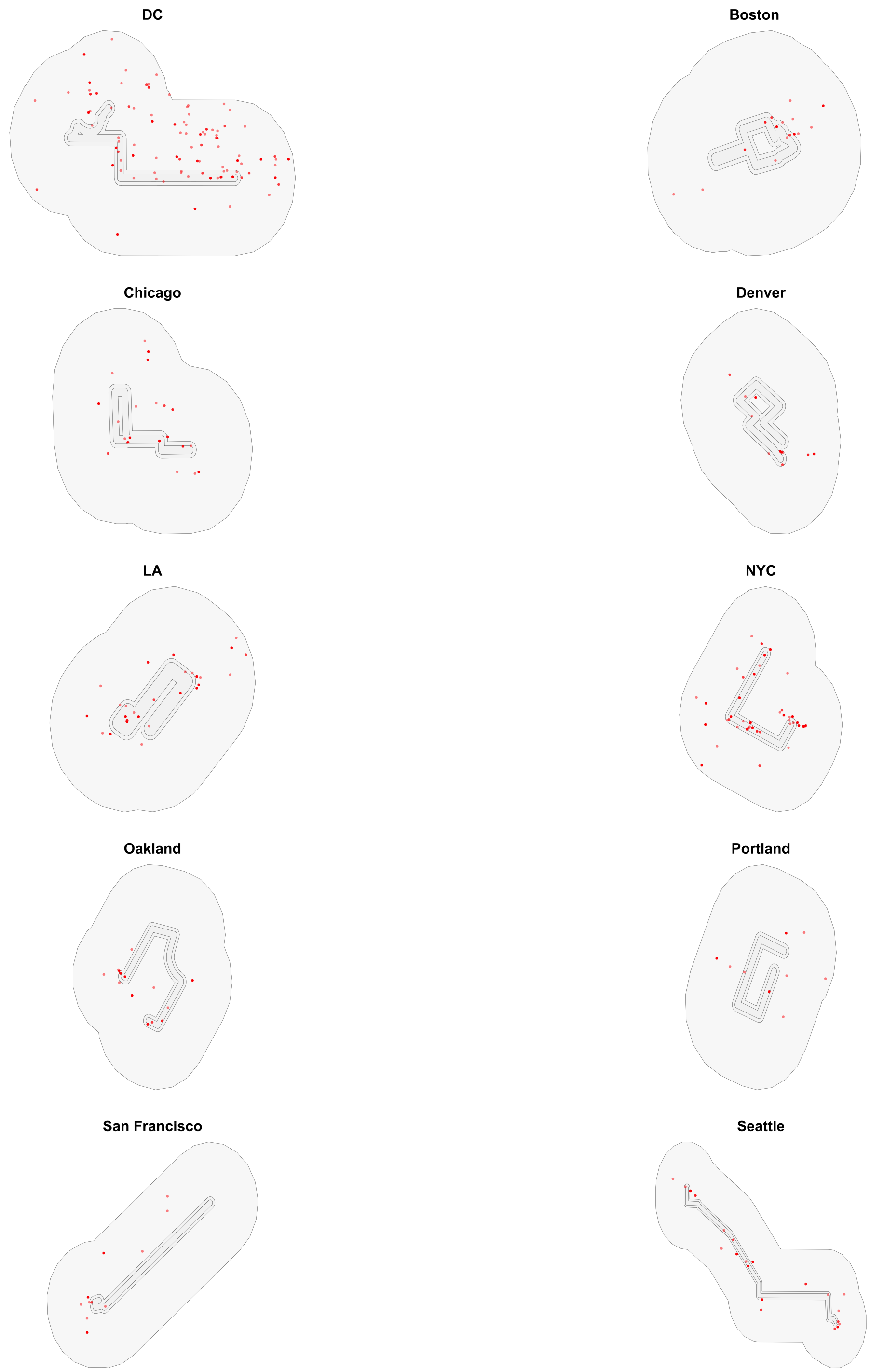


Figure .7: Georeferenced Women's March route maps and geolocated protestor Twitter users in ten US cities.

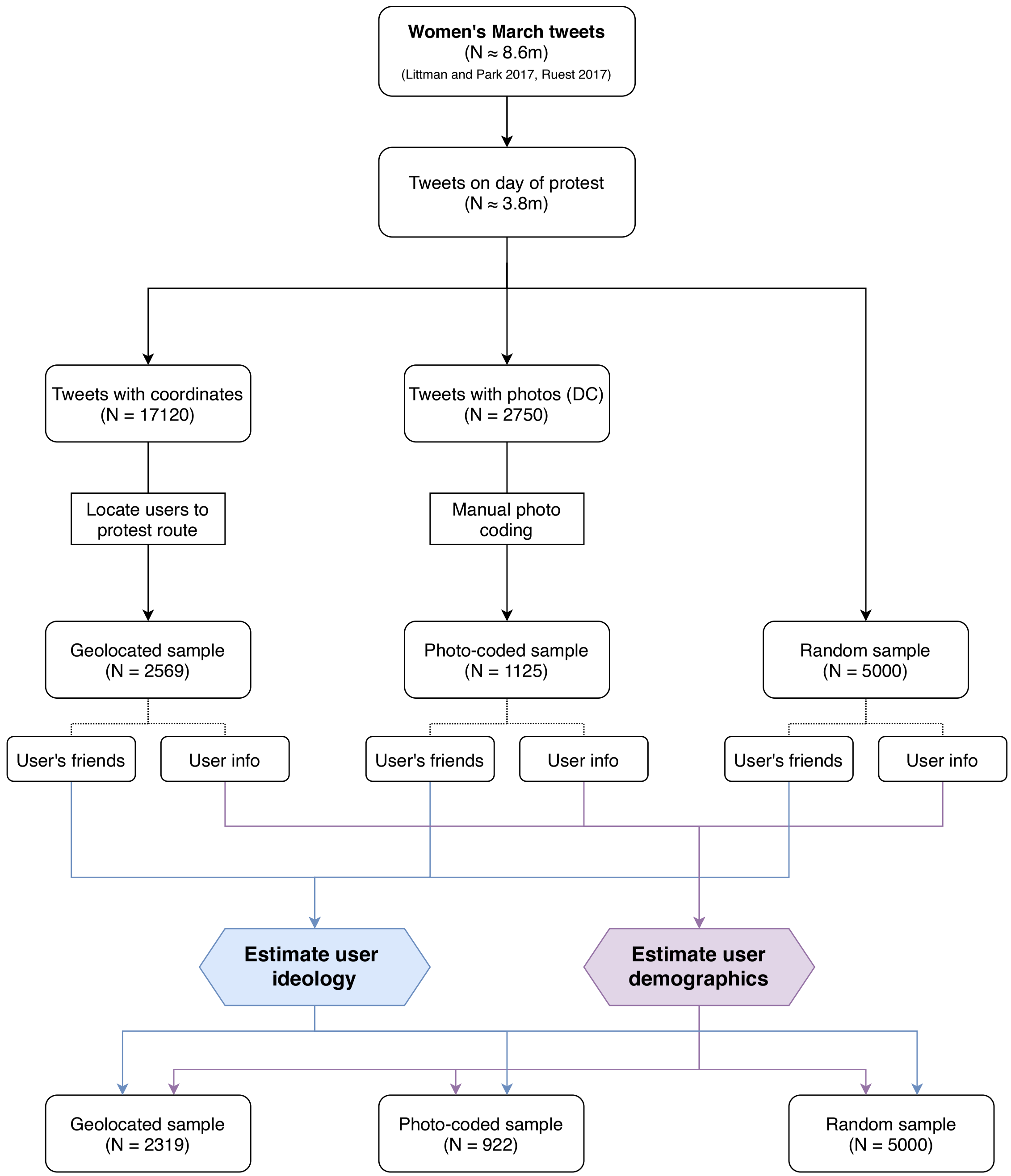


Figure .8: Visualization of data pre-processing and estimation procedure used in calculating crowd ideologies.

Ethical considerations

Twitter data is deemed public by most Institutional Review Boards, including the authors' home institutions.[27] Further, we did not stage any intervention when using these data. As a result, formal ethical review was not strictly required. We nonetheless sought the guidance of our Departmental Research Ethics Committee. The Tweet IDs we used were posted publicly by community organization Documenting the Now in accordance with Twitter's Terms of Service.[28]

Practitioners have questioned the exemption of public data from ethical review, and have noted tensions in the continued reliance on ethical review guidelines imported from a pre-digital age Chancellor et al. (2019); Conway (2014). A key reason for this skepticism derives from adherence to norms of contextual integrity (Nissenbaum, 2004). Using social media data for academic research removes the data from the original context of its intended reception, notwithstanding its public status (Zimmer, 2018). We believe our study *does* constitute a violation of contextual integrity, according to the nine principles set out by Nissenbaum (2004) and extended to social media research by Zimmer (2018). We have removed these tweets from the original context of their dissemination and likely intended audience (boyd and Crawford, 2012; Williams et al., 2017). What is more, we have processed these data in a way that facilitates the connection of an individual with a particular—and potentially sensitive—form of political behaviour. If users become disinclined to share information on associational activities such as protest as a result, there is a small chance that studies such as ours could harm civil society.

Based on these considerations, we determined 1) to elaborate an anonymization procedure for each analysis script to minimize the authors' exposure to identifying information; 2) that

[27] The (updated) guidance provided by the University of Oxford on Internet-Based Research is available at `https://researchsupport.admin.ox.ac.uk/files/bpg06internet-basedresearchpdf.` Given that this is a fast-moving domain and ethical considerations require constant updating, we also consulted, per the advice of the Central University Research Ethics Committee, the guidelines published by the Association of Internet Researchers, available here: `https://aoir.org/reports/ethics3.pdf`.

[28] See `https://catalog.docnow.io/` and `https://developer.twitter.com/en/developer-terms/agreement-and-policy`.

user-specific information on protestors—i.e., individuals we located to protest vicinities—would be stored in encrypted folders; and 3) that we will not include the identifiable geocoded or photo-coded protestor data subset in public replication files. While our approach does not eliminate the concerns we outline, we believe it helps significantly to mitigate them. In taking these measures, we follow best practice advice in the literature by aiming to reduce the likelihood that sensitive information will be traced back to individual users, by limiting researchers' own exposure to individual account information, and by taking responsibility for the custodianship of processed data (Clark et al., 2019; shakti franzke et al., 2020).

Twitter image coding criteria

To classify images in our Twitter dataset, we formulated a coding framework that followed a set of qualitative criteria. These criteria sought accurately to capture protestors on the ground by excluding: 1) photos not of protest; 2) photos accompanied by text indicating news reporting; and 3) photos that could be stock images. We provide the coding criteria used by both authors below.

Coding Criteria

We code as DC marchers tweets that include a photo, where:

1. The photo is taken from within the crowd, or on the way to the march
2. The caption accompanying the photo indicates that the user:
   (a) is at the march
   (b) is on the way to the march
   (c) is about to leave to go to the march
   (d) has just returned from the march
3. The text acompanying the photo describes the march, or uses popular hashtags
4. The text does not refer to a march outside of Washington, DC
5. The text and photo both indicate participation in the protest
6. The tweet does not indicate news reporting from official news media, and is not just purely descriptive and informational
7. The tweet is not just focusing on signs, but:
   (a) has to be accompanied by text indicating participation in the march, or

(b) taken from within the crowd, or

(c) has to be accompanied by text indicating that the user spotted this sign at the protest (e.g. one of my favourite signs I saw today)

8. The image is not curated or photoshopped

9. The photo is not of just children at march

10. The image is a selfie but includes text indicating that individuals are known to the tweeter and together at march.

Missingness

Both estimates of ideological and crowd composition suffer from some degree of missingness. Estimation of ideology scores relies on the following network of a Twitter user. If a user follows no elite accounts, their ideology score cannot be computed. This is the case for 111 observations (4.3%) in the Geo-located and 199 (17.7%) of the Photo-coded sample. In 7 further cases (5 for the Geolocated and 2 for the Photo-coded samples) no finite ideology estimate could be computed, so that the final missingness across both samples is 106 and 201. Estimates of demographic composition rely only on a user's profile information, and should in principle not include any missing data. Still, compiling errors and insufficient profile information results in 5.8% and 4.4% missing demographic data for the Geo-located and Photo-coded samples respectively. The considerable difference in missingness in our ideology estimates between the Geo-located and Photo-coded samples is due to differences in the timing of both estimations: the Photo-coded sample was added in the last stages of the writing process, to benchmark the findings of the geo-located sample against. Differences in missingness thus likely reflect the share of additional users who have made their profile private, deleted their Twitter profile, or had their account removed.

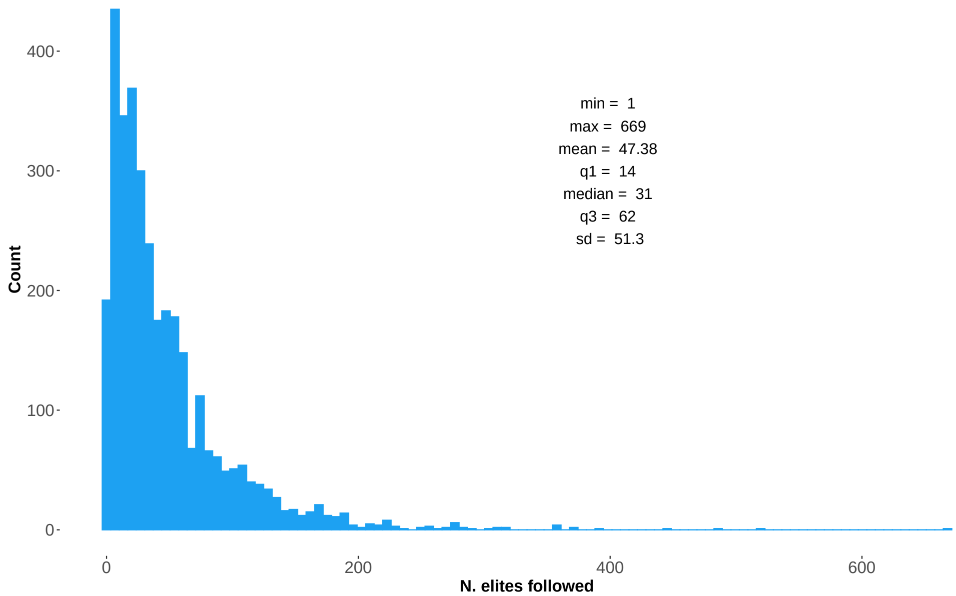
	Geo-located		Photo-coded	
	N	%	N	%
Ideology Score	116	4.5%	201	17.9%
Gender	148	5.8%	49	4.4%
Age Group	148	5.8%	49	4.4%
Sample size (incl. missing)	2519		1125	
Sample size (excl. missing)	2319		922	

Table .2: Missingness for ideology and demographic estimates for photo-coded and geolocated samples of Twitter protestors.

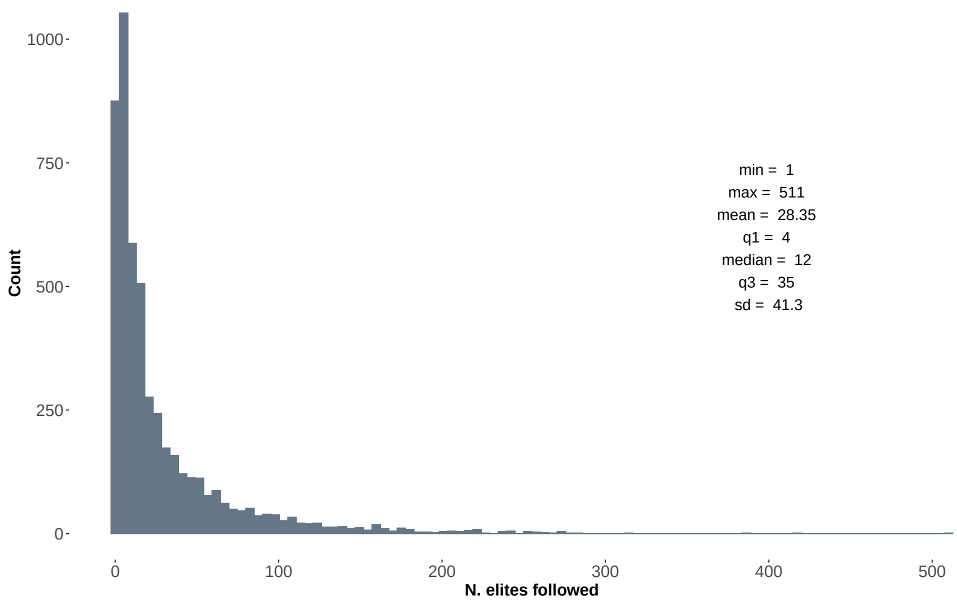
Following elite accounts

For our Geolocated and Photocoded samples, as well as our Random sample we also provide below descriptive statistics on the number of elite accounts they follow (see Figure .9). If a user follows only a small number of elite accounts, their ideology score will be estimated with larger error. Nonetheless, the mean and median number of elite accounts followed is reasonable: 47 and 31 respectively for our Geolocated/Photocoded users; 28 and 12 for our Random sample. If we understand the N. of elite follows as a proxy for political interest, the larger number of elite accounts followed by our protestors on the ground is another indication that hashtag sampling will return samples that also differ on this basic characteristic.

To assess whether the distributions we report are driven by users who follow only few elites (and whose ideological position are thus more imprecisely estimated) we repeat the analysis but only include users who follow at least 5 elite accounts. Figure .10 reveals that this has no impact on the ideological distribution of our three samples. In fact, the only sample whose distribution somewhat changes is the Random sample, which is likely due to us excluding users with a lower level of political engagement.



(a) Geolocated/Photo-coded samples



(b) Random sample

Figure .9: Number of elite accounts followed in Geolocated/Photo-coded and Random user samples

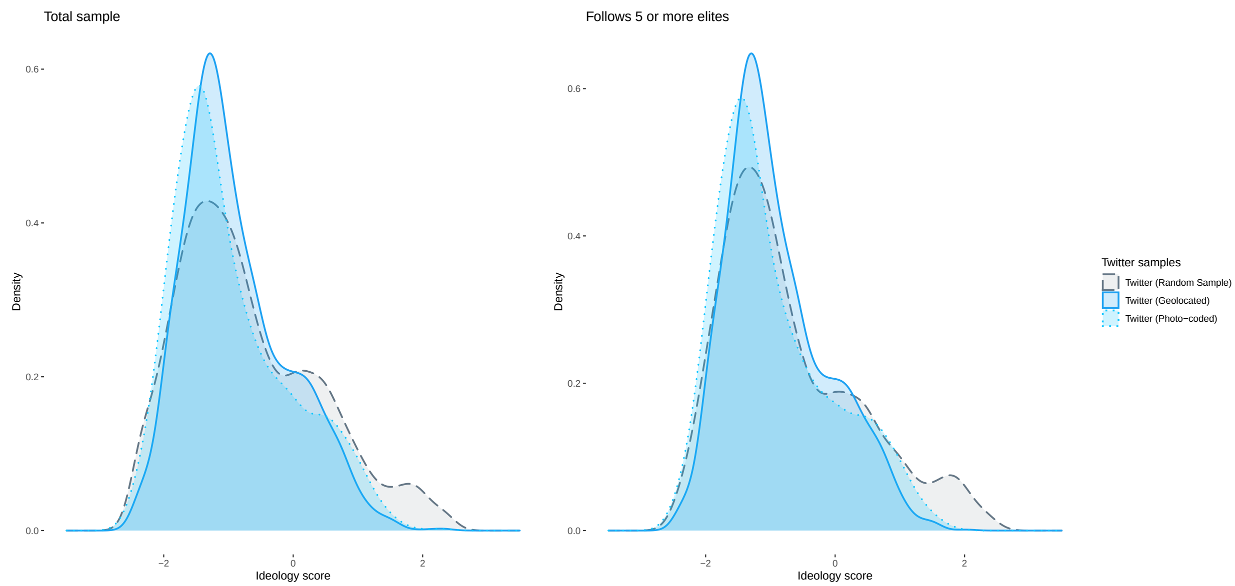


Figure .10: Comparison of ideology distributions of complete sample with ideology distribution of samples where users who followed less than 5 elites were excluded

Further sample characteristics

Sampling on hashtag alone, as previous research has done, risks including within the sample both commentators and opponents, as well as individuals who are only mobilizing online. By geolocating our users in our first "geolocated" sampling approach, we effectively remove the possibility of including protestors who are only active online as we are able to locate them to the route of the march. This approach, however, does not exclude the possibility that users in our sample geolocated to in or around the protest route might be either commentators (e.g., journalists, bloggers) or opponents. To scrutinize this, we first calculated the percentage of our geolocated samples in each of our ten cities who were "verified" users. The criteria for user "verification" has changed over time: it was initially used as a means to verify that accounts claiming to be celebrities or other individuals of public notoriety were genuine. It is also often granted to news organizations and journalists. Users granted "verified" status have a blue check mark against their user name. Indeed, the new verification guidelines (see: https://help.twitter.com/en/managing-your-account/twitter-verified-accounts) include "News organizations and journalists" as a category of user eligible for verification.

Table .3 displays the percentage of users in each of our samples who are verified users. These range from 0-7%.

N. verified	**N. non-verified**	**% verified**	**City**
0.00	83.00	0.00	Seattle
66.00	926.00	6.65	DC
14.00	417.00	3.25	NYC
44.00	544.00	7.48	LA
0.00	134.00	0.00	Boston
2.00	54.00	3.57	Portland
2.00	36.00	5.26	San Fran.
3.00	87.00	3.33	Oakland
2.00	89.00	2.20	Chicago
2.00	64.00	3.03	Denver

Table .3: Percentages of users "verified" in geolocated protestor samples

.

To further examine the potential size of any bias caused by the inclusion of commentators and/or opponents in our geolocated samples we manually coded a random sample of 500 of the geolocated tweets from the DC protest. In sum, twenty tweets (4%) were coded as being by news organizations or journalists (i.e., commentators) based on their user description and the content of the tweet. The criterion we used to judge the tweet content as being from a commentator was the same criterion used in the photo-coding criteria detailed above: i.e., "The tweet does not indicate news reporting from official news media, and is not just purely descriptive and informational." The account description was also used as indication: most often journalists will report their bylines, home publication, or broadcast news media organization.

We also coded the same random sample, on the basis of tweet text and user description, for potential oppositional content. One tweet (.2%) contained hostile sentiment toward the protest (though also mentioned gun rights as a women's rights issue, indicating the individual may have been a protestor but with an ideological outlook not congruent with the majority of other protestors).

In summary, while our geocoding procedure does not entirely filter out commentators on the ground, the size of any bias induced by their inclusion will be minimal. In other contexts, for example with smaller protests, the size of this bias will be non-negligible, meaning that other techniques such as the photo-coding procedure we outline may be preferable. It is also relatively straightforward to identify commentator accounts from their number of followers, verification status, user description, and tweet content.

As for opponents, this appears to be less of a concern, at least in the case of the Women's March. For more contentious protests, or protests that attract counter protests, it will be a larger task to parse supporters and opponents. In this case, filtering by geolocation and then manually coding a random subsample may be the best approach. To scale this approach, automated classification techniques may also be used to label users as opponents or supporters based on user-level and tweet-level characteristics (see e.g., Rafail 2018).