Using Administrative Records and Survey Data to Construct Samples of Tweeters and Tweets

Adam G. Hughes[1], Stefan D. McCabe[2*], William R. Hobbs[3], Emma Remy[1], Sono Shah[1], and David M. J. Lazer[2]

[1]Pew Research Center, 1615 L St NW #800, Washington, DC, 20036

[2] Network Science Institute, Northeastern University, 10th floor, 177 Huntington Avenue, Boston, MA 02135

[3] Department of Human Development, Cornell University, T231 Martha Van Rensselaer Hall, 37 Forest Home Drive, Ithaca, NY14853

* Corresponding author

Email addresses: aghughes@gmail.com, mccabe.s@northeastern.edu, hobbs@cornell.edu, eremy@pewresearch.org, sshah@pewresearch.org, d.lazer@northeastern.edu

Word Count: 6,198

ADAM G. HUGHES was associate director of the Data Labs team at Pew Research Center, Washington, DC, USA when the study was conducted. STEFAN D. McCABE is a PhD. student at the Network Science Institute, Northeastern University, Boston, MA, USA. WILLIAM R. HOBBS is an assistant professor in the Department of Human Development at Cornell University, Ithaca, NY, USA. EMMA REMY is a data science analyst at Pew Research Center, Washington, DC, USA. SONO SHAH is a computational social scientist at Pew Research Center, Washington, DC, USA. DAVID M. J. LAZER is a professor at the Network Science Institute, Northeastern University, Boston, MA, USA. The first and second author contributed equally to this work. The authors thank the audience and discussants at the 2020 Annual Meeting of the American Association for Public Opinion Research, the 2020 annual meeting of the American Political Science Association, and the 2020 International Conference on Computational Social Science, Aaron W. Smith, and three anonymous reviewers for their feedback. The authors also thank Nir Grinberg, Lisa Friedland, and Kenny Joseph for earlier technical work on the development of the voter file sample. DL acknowledges support from the Hewlett Foundation for this research. The authors declare they have no conflicts of interest.

*Address correspondence to Stefan D. McCabe, Network Science Institute, Northeastern University, 177 Huntington Ave, 10th floor, Boston, MA 02115, USA; email: mccabe.s@northeastern.edu. REPLICATION DATA AND DOCUMENTATION are available at https://github.com/sdmccabe/poq-constructing-samples-replication.

Abstract

Social media data can provide new insights into political phenomena, but users do not always represent people, posts and accounts are not typically linked to demographic variables for use as statistical controls or in subgroup comparisons, and activities on social media can be difficult to interpret. For data scientists, adding demographic variables and comparisons to closed-ended survey responses have the potential to improve interpretations of inferences drawn from social media—e.g., through comparisons of online expressions and survey responses, and by assessing associations with offline outcomes like voting. For survey methodologists, adding social media data to surveys allows for rich behavioral measurements, including comparisons of public expressions with attitudes elicited in a structured survey. Here, we evaluate two popular forms of linkages—administrative and survey—focusing on two questions: how does the method of creating a sample of Twitter users affect its behavioral and demographic profile? What are the relative advantages of each of these methods? Our analyses illustrate where and to what extent the sample based on administrative data diverges in demographic and partisan composition from surveyed Twitter users who report being registered to vote. Despite demographic differences, each linkage method results in behaviorally similar samples, especially in activity levels; however, conventionally sized surveys are likely to lack the statistical power to study subgroups and heterogeneity (e.g. comparing conversations of Democrats and Republicans) within even highly salient political topics. We conclude by developing general recommendations for researchers looking to study social media by linking accounts with external benchmark data sources.

Social scientists rely upon social media data to measure the prevalence of misinformation (Grinberg et al. 2019; Guess et al. 2019), behavioral political polarization and selective exposure (Bakshy, Messing, and Adamic 2015; Bail et al. 2018), political discussion (Barberá and Rivero 2015; Jaidka, Zhou, and Lelkes 2019), and many other phenomena—in some contexts, even economic confidence (Pasek et al. 2018) and election outcomes (Beauchamp 2017). Yet researchers that rely on social media data—especially data collected from Twitter—face a fundamental generalizability problem: who uses social media and how do they use it? We argue that linking surveys or administrative data with social media data can help answer these questions and improve research designs for both data scientists and for survey methodologists. We provide here a comparison of the posting behaviors and demographic composition of Twitter accounts linked to administrative data and to surveys, comparing these samples to registered voters and survey respondents who use Twitter and/or report being registered to vote whether or not they provide us with Twitter handles.

For data scientists whose research focuses primarily on social media data, representativeness is often hard to systematically assess. Estimating race, age, gender, and partisanship based upon profile information is fraught: imputing this information raises ethical concerns (Hamidi, Scheuerman, and Branham 2018; Raji et al. 2020) as well as the potential for measurement error (Wu et al. 2019). Although social media data have many advantages for social science research, some research questions carry well-known representativeness concerns when studied on social media: the data-generating processes underlying social media posts can vary substantially over time and across platforms, and using online behavior as a measure of public opinion and other offline outcomes on its own can be particularly difficult and potentially misleading (Tufekci 2014; Klašnja et al. 2018; Barberá and Steinert-Threlkeld 2020). Adding demographic variables and comparisons to closed-ended survey responses have the potential to improve interpretations of inferences drawn from social media through use of demographic variables for subgroup analysis and weighting, through comparisons of online expressions and closed-ended survey responses, and by assessing associations with offline outcomes like voting.

For survey methodologists, the provision of behavioral data both improves measurement quality and provides new kinds of measurements. Self-reports of social media use may be unreliable (Ernala et al. 2020; Henderson et al., forthcoming); observed posting and following behavior can provide better estimates. And the opportunity to collect public expressions—unconstrained by survey response—allows for the analysis of new forms of behavior, such as misinformation sharing (Grinberg et al. 2019) or political discussion (Hughes et al. 2019). Indeed, social media posts, likes, and following decisions serve as a high-resolution form of panel data that supplements cross-sectional surveys when the two are linked.

Existing guidance about how social media data complements survey research (Japec et al. 2015; Nagler and Tucker 2015; Schober et al. 2016; Klašnja et al. 2018) points out that representativeness is often a significant concern, and that representativeness concerns vary widely by target population and by research question (Barberá and Steinert-Threlkeld 2020). What's more, people who post—let alone, post about particular topics—are not the same as all people who use a platform. Because social media platforms do not elicit opinions from users in a consistent way, understanding the representativeness of expressed opinions on social media requires additional legwork.

As a result, recent research has focused on identifying demographic attributes of social media users (Barberá and Rivero 2015; Grinberg et al. 2019; Guess et al. 2019), which would permit both a descriptive analysis of who posts particular content alongside the potential for applying weights to social media data (for a similar approach, see Wojcik et al., forthcoming). This kind of analysis also allows researchers to expand their analytic strategy; while Twitter metadata can be used to estimate ideological position (Barberá 2015) or location (Nagler and Tucker 2015), access to additional attributes provides an avenue for more sophisticated models and important subgroup analyses, perhaps aided with multilevel regression and poststratification (Ghitza and Gelman 2013; Gelman et al. 2016).

We present two approaches for linking information about U.S. adult Twitter users with their social media accounts. The first relies upon administrative data: the voter file. By matching names in the voter file with Twitter accounts, and blocking by geographic location, researchers can link individuals' basic demographic attributes with their social media data. This approach avoids survey nonresponse and provides massive scale—but is limited by the kinds of descriptive information contained in political administrative data. Further, the matching process—which relies on the statistical (in)frequency of names and identification of geographic location of a user based on their profile information— yields its own kinds of biases, as we discuss below. The second strategy relies upon surveys linked with social media data, in which respondents volunteer their social media usernames and researchers separately collect posts. This provides a richer and potentially more representative attitudinal and demographic portrait of those on the site, though it is costly and suffers from more limited sample size and from nonresponse bias. We test both approaches for U.S. adults on Twitter, focusing on two questions: how does the method of creating a Twitter sample affect its behavioral and demographic profile? And, second, what are the relative advantages of each of these methods?

In our analyses that follow, we first assess the demographic compositions of a voter file-based sample and two survey-based samples. To understand the sources of differences across samples, we determine which demographic differences existed in the source data (i.e., the demographics of all voter file registrants compared to all self-reported registered survey respondents) and then evaluate how the linking process for each recruitment method affected demographic composition. We then proceed to compare activity levels and tweet content across samples. This behavioral analysis compares tweeting and liking rates, follower and following numbers, skewed activity levels within samples (e.g., what fraction of users accounts for what proportion of tweets), and rates of mentions of select high-salience keywords.

Matching Twitter account and post data with individual-level information (either via surveys or via administrative data) presents important ethical challenges for researchers (Sloan et al. 2020). To preserve subjects' privacy, we stored all Twitter data in secure databases with restricted access. We also exclude users with private accounts from all tweet collection. For those with public accounts, we adopt two different approaches. Among respondents linked with the voter file, we cannot obtain consent to participate directly. Instead, we focus on minimizing harm or risk by protecting any personally identifiable information (PII), restricting access to data, and analyzing attributes of users and tweets in the aggregate.

A central virtue of survey samples is that individual users are directly asked for permission to view their Twitter data; we provide examples of this language in Online Appendix A3 (see also Sloan et al. 2020). Evidence from a convenience sample of Twitter users (Fiesler and Proferes 2018) suggests that permission is an important consideration for respondents. However, across both methods examined here, any PII about particular users is largely obscured throughout data collection and analysis. While there is evidence that the vast majority of users are unconcerned about their tweets being used for research purposes, 90% prefer that any personal information is not made publicly available as a result of that research (Williams et al. 2017). Accordingly, we do not release any PII, including tweet text, in the replication materials; this approach is also consistent with the privacy-protecting principles of Twitter's Terms of Service.

Voter File Data

For our first sample, we link public Twitter accounts to state voter records compiled by TargetSmart. Voter registration records are a matter of public record in the United States and have been used in political science to better understand individual-level voting behavior (Ansolabehere and Hersh 2012). We provide an overview of the matching process here, and provide more details in Online Appendix A4.[1] An earlier version of this data set is described in Grinberg et al. (2019).

To create this data set, we begin with a list of all users appearing in a 10 percent "Decahose" sample of Twitter between January 2014 and March 2017 (290 million accounts). Of these users, we focus on those with identifiable names and U.S. locations. For users for whom we are able to extract a name and location,[2] and whose name is unique within that location, we include them in our dataset. Although this is a relatively strict requirement, especially compared to probabilistic approaches (Enamorado, Fifield, and Imai 2019), we were able to match 1.5 million accounts (about 3% of all U.S. adult users, using estimates from Perrin and Anderson 2019). We note that this approach is limited in its ability to study low-frequency tweeters, but this is not a general feature of such administrative record matching. Users who never post of course never take part in conversations on Twitter, however it is possible that they account for a nontrivial portion of the audience for content on Twitter (Lazer 2020). In contrast, some strict requirements in name matching may be unavoidable for this form of matching, and we evaluate how unique names affect the demographics of such samples in the following analysis.

Survey Data

[1] Code that enables this record linkage process is available in the replication materials.

[2] For reference, 43% of respondents to the KnowledgePanel survey and 50% of respondents to the American Trends Panel survey had apparently valid names and locations listed in their profiles.

The second methodology we use to create a Twitter sample is to simply ask people for their Twitter handles. Using surveys to study social media behavior helps ensure that samples of users are representative—or at least that bias is measurable. However, survey-based self-reports of online behavior may not be reliable or accurate (Henderson et al., forthcoming). As an alternative, researchers can ask survey respondents to provide their social media account data or usernames, and separately collect and analyze their observable behavior. While some individuals prefer not to participate in this kind of passive data collection (Keusch et al. 2019; Kreuter et al. 2020), we can assess whether the demographic composition of those who supply handles substantially differs from those who do not (see results below and Online Appendix A1 for details).

We conduct two original surveys and link volunteered Twitter handles with data obtained from the Twitter API. The first survey was conducted via Ipsos' KnowledgePanel (KP) from Nov. 21–Dec. 17, 2018. Weights for the KnowledgePanel survey were created by raking to estimated population totals, rather than known population totals. These estimates came from a contemporaneous probability survey (a Nov. 2018 wave of Pew Research Center's American Trends Panel). We filtered that sample to self-identified Twitter users, and then used the weighted estimates from that survey as targets. Ipsos provides base weights for KnowledgePanel samples that account for the probability of selection; these were then raked to match the demographic parameters taken from the 2018 American Trends Panel survey.[3] The survey invitation was sent to 7,850 likely Twitter users (based upon the vendor's previous data

[3] See Online Appendix A6 for further information about weighting. The contemporaneous American Trends Panel survey was in the field Nov. 7–16, 2018.

collection on each sampled user), of whom 4,829 responded. Of that set, 3,649 confirmed that they used Twitter and 3,293 agreed to provide a Twitter handle.[4] In this survey, providing a handle was a qualification for survey completion: as a result, only three respondents who confirmed that they used Twitter broke off before entering a handle, and the share of invalid handles was substantially higher than in the second survey.

We evaluated the apparent validity of handles by comparing Twitter profile information with demographic information provided in the survey. All accounts belonging to organizations, institutions, international entities, and public figures are excluded from the analysis (see Online Appendix A1 for details of the validation process). Overall, 2,791 respondents (76% of those who provided any handle) provided a valid handle, according to these criteria. A total of 2,369 of these respondents had public accounts; that set is analyzed here.

The second survey was conducted via an online probability panel from Pew Research Center, the American Trends Panel (ATP). The survey was in the field between October 29 and November 11, 2019, and did not target Twitter users specifically (note that this wave was separate from the 2018 wave used for weighting targets for the first survey). The survey was sent to all panelists (regardless of Twitter use): a total of 14,412 respondents, of which 12,043 responded. Among those who responded, 2,561 reported using Twitter, and 1,518 provided a handle. We reviewed this set of handles; 1,346 appeared valid (89% of those who provided any handle), based upon

[4] See Online Appendix A1 for information about the demographic attributes of those who did and did not volunteer a handle. 62% of invited panelists responded.

the same criteria described above. Of those handles, 87% were public, resulting in an analysis dataset of 1,188 accounts.[5]

For comparison, we also examine estimates from a separate random-digit dial survey conducted in January 2019 by Pew Research Center to compare with the two survey samples. That survey did not include an opportunity for respondents to volunteer handles, but does provide demographic benchmarks that we use to evaluate the composition of the other survey samples (details about this survey are also included in Online Appendix A6).

Data for Sample Comparisons

We analyze demographic composition and behavioral data on Twitter to compare the three samples. Demographic variables come from self-reports in all of the samples, where in the voter file these self-reports are made during voter registration.[6] Behavioral data comes in two forms: metadata, which are largely cumulative statistics that can be collected at a single point in time; and tweet data, which must be collected continuously to capture a user's most recent tweets.

Before comparing each data source, we checked for overlap across the samples as a rough assessment of coverage in the voter file data. About 5% of accounts in each survey sample

[5] See Online Appendix A6 for information about weighting the 2019 American Trends Panel sample of Twitter users.

[6] The vendor draws from self-reports where available—that is, in areas that record registration by race to comply with the Voting Rights Act—and falls back on other data sources otherwise.

overlapped with the voter file sample; Online Appendix A5 contains details on this procedure as well as a comparison of self-reported demographics and those drawn from the voter file.

Demographic Composition

We begin the sample comparison by focusing on registered voters who are linked with Twitter accounts: while this group of Twitter users is not representative of all U.S. adults on Twitter, it is the clearest comparison point for the data obtained from the voter file. We compare the voter file-linked accounts with respondents from each survey, which included the same question about voter registration. In this comparison, we define registered voters within the surveys as all respondents who say that they are *sure* they are registered to vote. 74% of ATP panelists who provided handles report being registered, while 71% of KnowledgePanel members say the same. Agreement between self-reported voter registration and voter file-reported registration is generally quite high, and disagreements are not necessarily indicative of social desirability bias; disagreements can also be accounted for by state-level variation in maintaining the voter file (Berent, Krosnick, and Lupia 2016). For the American Trends Panel survey, the figure includes demographic information for both the full set of registered respondents who say that they use Twitter, as well as the set of registered respondents who both say they use Twitter and provided researchers with access to their behavioral data.

Figure 1 shows that in terms of age, the voter file has a larger share of 18-29 year-olds than the survey samples: a difference of 8 percentage points relative to the KnowledgePanel ($p < 0.01$) and 12 points relative to both American Trends Panel samples ($p < 0.01$). It is unclear why the samples depart so much, but as discussed below, there is some evidence that all self-identified registered voters in the ATP tend to be older than individuals in the voter file. There is also a notable difference when it comes to self-reported gender: the voter file has a larger share of women than either survey source. The difference is 8 percentage points (compared with KnowledgePanel, $p < 0.01$) and 12-13 percentage points in comparison with the American Trends Panel ($p < 0.01$ for both comparisons). As discussed below, this difference is driven by the matching algorithm: women are more likely to have unique names than men do (thus facilitating a match with the voter record).

[INSERT FIGURE 1 ABOUT HERE]

Turning to racial composition, we find that white Americans are substantially more likely to appear in the voter file sample: a 22 point larger share of respondents are white, relative to KP. The difference is 19 points relative to all Twitter users on the ATP, and 16 points relative to just those ATP respondents who provided handles. In addition, the voter file has fewer Hispanic users than the survey-based approaches (11 percentage points less than KP, 8-9 points less than the ATP samples). However, this percentage matches the overall distribution of Hispanic voters in the voter file (see below, Table 1).

[INSERT TABLE 1 ABOUT HERE]

To understand whether the racial differences we observe are robust, and given that only Voting Rights Act states[7] consistently collect self-reported race, we analyze the race/ethnicity numbers within those states specifically. While estimates of the fraction of Black users in each sample align much more closely (a difference of 1 point for KP, 2–5 points for ATP), other differences persist. In Figure 2, we see no reduction in the discrepancy between the administrative and survey approaches for white users: the difference between the voter file and KP is 29 points, while the difference relative to the ATP samples are between 14 and 20 points. This suggests that caution, and perhaps reweighting (depending on the scientific objective), is warranted when analyzing the online behavior of non-white Twitter users who are matched with administrative data, especially because the racial composition of both survey samples align closely with a separate RDD poll (see Figure 3).

[INSERT FIGURE 2 ABOUT HERE]

[INSERT FIGURE 3 ABOUT HERE]

When it comes to partisanship, 40% of all voter-file-sampled users are registered with either party, 23 percentage points less than the share of partisan identifiers in any of the survey samples. To evaluate the source of this discrepancy, we examine the subset of states where party is consistently recorded—specifically, states with closed primaries, as well as those which allow unaffiliated voters to participate but prohibit votes from the opposing party (which we will

[7] We analyze states entirely under preclearance between 1975 and 2012: AL, GA, LA, MS, SC, VA, AK, AZ, TX.

simply label “closed” in the figure).[8] When we look at this subset, the fraction of individuals with an identified partisanship is substantially higher for the voter-file sample than the surveys, i.e., presumably self-identifying independents with a partisan lean tend to register with that party. Overall, we see greater proportions of Democrats using Twitter though the survey samples are no more Democratic than the RDD phone poll (see below, Figure 3). The voter record Twitter sample leans somewhat more Republican than the survey samples: 6 points relative to KP ($p = 0.09$) and 7-8 points relative to ATP ($p < 0.01$).

Next, we compare the overall estimates of the registered voter population across the full voter file and the broader sample of American Trends Panel respondents, changing our population from registered Twitter users to all registered voters. Because the American Trends Panel is a national survey (subset to Twitter users for much of our analysis), it provides a reasonable benchmark to compare against the voter file sample of registered voters. This allows us to evaluate whether our starting samples were different to begin with, or if the differences were more likely caused by data processing procedures and/or survey non-response.

Table 1 shows that the voter record had a higher proportion of registered voters that are white, female, and Democratic, and a notably smaller proportion who are Asian-American or who identify with other racial groups. The high proportions of white and female registered voters in the voter record, in particular, help explain the discrepancies in Figure 1.

[8] Closed (and partially closed) primary states used in this analysis include: CT, DE, FL, KS, KY, ME, MD, DC, NE, NM, NY, PA, WY. The list is drawn from the National Conference of State Legislatures: https://www.ncsl.org/research/elections-and-campaigns/primary-types.aspx

To further explain discrepancies in Figure 1, we also calculate in the full voter file the number of unique names within a state by age and gender. We then multiply the fraction of unique names for those characteristics by the survey numbers to estimate what those numbers would be given the same non-unique name exclusion rules. We find that half of the difference in gender composition between the American Trends Panel and the voter record sample can be explained by more unique names for women (56%) compared to men (43%) in the voter record sample. Removing the same fractions of women and men from the American Trends Panel (57% men, 43% women in the registered voter subset) as from the voter record based on non-unique name combinations would shift those survey estimates to roughly 51 percent men and 49 percent women. The higher fraction of women in the voter record sample is further explained by the initial difference in gender between the voter record and the survey of all registered adults (Table 1). The large number of people under 30 in the voter record Twitter sample is not as well explained by unique names among younger people, but, as shown in Table 1, the voter record did have a larger share of people under 30 than those who self-reported voting in American Trends Panel to begin with.

Finally, we examine the demographic characteristics of respondents to both online surveys alongside estimates from a separate, RDD survey of U.S. adult Twitter users. None of these comparisons are restricted to registered voters. Across all three samples, age, income, education, and urbanicity[9] are also very similar. The American Trends Panel sample is slightly more likely to include men than the other samples; as Online Appendix A1 shows, this difference is

[9] 2013 NCHS Urban-Rural Classification Scheme for Counties: https://www.cdc.gov/nchs/data_access/urban_rural.htm

attributable to the fact that women were less likely to volunteer handles in that survey (57% for men versus 43% for women). Partisanship is consistent across the samples, though the phone poll has a smaller share of Republicans.

Behavioral Data

We next analyze two separate kinds of behavioral data: metadata and tweets. The former describes behavior in general terms across the lifetime of particular accounts, while the latter require researchers to select a timeframe for analysis. In the following section, metadata are collected as of January 2020 for all samples, while tweet data include all tweets posted within the window of October 1, 2019–November 30, 2019. Researchers obtained all Twitter data using the Twitter API, and extracted relevant text and quantities of interest from the JSON returned by a consistent set of queries.

Account metadata includes the age of all accounts, the number of tweets they created, the number of accounts they follow, the number of accounts that follow them, and the number of tweets they favorited. These analyses are the easiest to align across the data sets because they do not rely on continuous data collection over time.

Overall, the metadata-based behavioral statistics are similar across the four kinds of behavior examined here (Figure 4), with some notable exceptions. Unsurprisingly, the Twitter data linked to voter records features relatively few users who have only posted once; this particular sample was developed from Twitter IDs of posts which had shown up in the Twitter Decahose.[10] In addition, the American Trends Panel sample has many more likes than the other data, while the KnowledgePanel sample has many more people who have never liked content. Otherwise, all three samples have similar activity levels.

[INSERT FIGURE 4 ABOUT HERE]

We also calculate the share of tweets—based upon metadata—that come from the top decile of active users. Using the voter file source, we estimate that the 10% most active tweeters (based upon lifetime tweets) generate 81% of all tweets. In the KnowledgePanel sample, the top 10% generate 83% of all tweets, while that figure is 73% for the American Trends Panel. The implication for studies of tweets—examined separately from users—is that a small number of accounts is responsible for most observed tweeting behavior. To the extent that the demographic composition of frequent tweeters on a specific topic does not mirror the population of U.S. adults talking about the topic on Twitter, research based upon large samples of tweets containing keywords (Barberá and Rivero 2015; Beauchamp 2017) may miss the contributions of users who rarely tweet.[11]

The metadata also allows for a comparison of coverage among rare tweeters. We define "lurkers" as U.S. adults on Twitter with fewer than 10 tweets across the lifetime of their account.

[10] It is possible that we undercount systematically different users as a result of the Decahose-based sampling scheme. We present an assessment in the Online Appendix A10, along with a possible weighting scheme to address sampling concerns.

[11] Here, for example, the top 0.1% of users (who are the top 1% of those tweeting the topic) most frequently tweeting or retweeting "impeach" are more likely to be white (93% in VRA states) or Republican (36% in closed primary states) than users tweeting "impeach" at all (88% and 20% respectively).

In the KnowledgePanel sample, 21% of users fall into this category; the rate is 16% for the American Trends Panel. By contrast, in the voter file sample, the share of users with fewer than 10 lifetime tweets is just 5.4%. This would point to the desirability, when matching to administrative data, of developing methods to identify low-activity Twitter accounts, as in the random-sampling-of-IDs approach of Barberá et al. (2019).

The second kind of data, actual tweets, sheds light on researchers' ability to extract meaning from social media posts and following behaviors. Across the three samples of tweets, we examine the timeframe October 1, 2019–November 30, 2019. We include retweets and replies in this analysis. The amount of tweets that researchers can collect is limited by the scale of accounts that researchers select. Due to limitations in the Twitter API, collecting all posted tweets for a very large number of accounts is impossible. For smaller samples—including both survey samples discussed here—we collected all tweets, on an ongoing daily basis, across the study timeframe. For the voter file sample, we collect tweets from each user in approximately six-week intervals.

To begin, we looked at basic descriptive statistics for each group. First, a large share of accounts created no new tweets during the time window: for the Voter File, 60% had no tweets, while for the KnowledgePanel and American Trends Panel, the shares are 51% and 32%, respectively. Because providing a handle was not a screener question in the American Trends Panel survey, it is possible that a larger share of active users remembered and volunteered handles, resulting in a lower rate of inactive tweeters. The volume of tweets obtained during the two-month period varies according to the scale of each sample: from the voter file sample, we obtained 46,539,343 tweets (a mean of approximately 30 per account); from KnowledgePanel, 118,238 tweets (mean 50), and from the American Trends Panel, 96,514 tweets (mean 80).

Consistent with the analysis of account metadata above, analyzing tweets themselves also show that the most prolific users generate most content on the platform: for the voter file, 70% of tweets come from the 10% most active, 74% of tweets in the KnowledgePanel sample come from the top 10%, and 61% of tweets in the ATP sample are from the most prolific 10%. All three samples show very similar results when it comes to retweets and replies: for the voter file sample, 35% of tweets are retweets and 29% are replies. For KnowledgePanel and the ATP, 33% are retweets in each case, while 35% (KP) and 40% (ATP) are replies.

Next, we examined a series of keywords to compare estimates of overall popularity. The keywords we selected focused on political discussion ("impeach", "Trump," "Republicans," "Democrats") and one examine of a popular hashtag ("#TBT", meaning "throw-back Thursday").

The analysis shows that overall estimates of the share of users that ever used any of these terms during the selected timeframe varies across the samples, with ATP respondents using political terms more often. For example, 13% of users in the voter file sample used "impeach," similar to the 12% of KnowledgePanel respondents who used the term. But 18% of American Trends Panel members did the same. A similar pattern holds for "Republican," with 10% of voter-file-sampled users mentioning the term, compared with 9% of KnowledgePanel respondents and 18% of ATP panelists. Table 2 reports the results for the other terms.

[INSERT TABLE 2 ABOUT HERE]

Results in Table 2 would appear to suggest that ATP respondents are more politically interested than the other samples, and the followed-accounts analysis appears to support this conjecture (see Online Appendix A11).[12] However, among the respondents who used each term at least once, the mean and median rates of tweeting the term is largely consistent. And, one non-political keyword ("#TBT") has similar use patterns across all sources.

In addition to the higher rates of political keyword use, this analysis also reveals a fundamental limitation of survey-based approaches: the number of respondents who used particular terms is often small. Indeed, for even a highly salient keyword like "impeach", however, the two surveys have a much smaller number of respondents mentioning the word over even two months. Although we are able to estimate the fraction of respondents using the word, the surveys have limited sample size to further explore discussions surrounding the topic. So while it is true that a larger share of users used the word "impeach" in both survey samples, the raw number of respondents who did so is quite small: 208 for KnowledgePanel and 178 for the ATP. The "poll"-wide (maximum) margin of error for estimates based upon each of these subsets is 11 percentage points (for each).

[12] The sample of Twitter users who volunteered handles were also slightly more likely to say that they regularly follow the news and are following the 2020 U.S. Presidential election, compared with Twitter users who opted not to provide handles.

In Online Appendix A8, we further compare content posted by the voter file users to US-based users who do not appear in the sample. This analysis makes use of the Twitter "Decahose", a random 10% sample of tweets. The findings suggest that non-sampled users were substantially more likely to post about pornography and conspiracy theories, and may be more likely to engage in hashtag manipulation. Sampled users, in contrast, were more likely to discuss the COVID-19 pandemic and (support for) social justice movements, especially Black Lives Matter.

Finally, we selected a small number of political Twitter handles to measure the proportion of followers in each sample, and for these compared their followers against our user samples. These results are shown in Online Appendix A11; in general there was agreement between the voter file and KP samples, with ATP showing a higher level of political interest.

Discussion

Although social media data provide rich behavioral measurements that can complement survey work, inferences based on social media alone have the potential to mislead. The process of linking individuals to accounts can help provide a relatively well-defined and interpretable sample.[13] Linkages permit evaluations of demographic representativeness, analyses of subgroups, and validation of measurements drawn from online posting behavior by comparing to survey responses. In addition, these methods help avoid the bias associated with drawing samples of users from narrow time windows: by developing samples linked to external benchmarks and

[13] For example by excluding bots, which may account for a large share of Twitter behavior (Wojcik et al. 2018), as well as organizations, celebrity pets, and the many other creatures in the social media menagerie.

whose members are followed over time, researchers can assess whether the users who tweet about a particular topic of interest at a particular point in time reflect the broader population of U.S. adults on the platform.

Our research is intended to assist both data scientists and survey methodologists considering how to best link social media accounts with external data. How accounts are matched to individuals (and various individual attributes) is potentially quite consequential, and conducting such matches demand significant time and resources. To help guide that planning, this paper compares the demographic and behavioral biases introduced by two survey-based and one administrative data-based approaches to sample construction. During the period examined, tweeting behavior was highly concentrated across all three samples, with large shares of accounts never tweeting, and the share of all tweets generated by the 10% most active users for accounts that tweeted at least once is (conservatively) above two-thirds of tweets. More generally, behaviors as captured by these different approaches were fairly similar. Since our samples include only US adult users, and not accounts in general, we provide evidence that the general tendency towards highly skewed activity distributions on Twitter (previously documented in boyd, Golder, and Lotan 2010; and Chalmers et al. 2011; among others) holds for subpopulations of real users and is not an artifact of bots or institutional accounts.

However, there are some important ways that these approaches diverge. For example, the matches to voter records are more likely to include white adults and women than the surveys. We infer this is likely due to both differences in the administrative data and issues in the matching process—for example, there are more whites in the administrative data compared to survey self-reports of voter registration and women are more likely to have unique names within a given geographic area. We did not observe similar biases in the linking process on the survey side—when we compare demographic estimates with benchmarks from a random-digit dial survey, we find almost no differences across the samples. This suggests that asking survey respondents to provide a handle does not materially damage demographic representativeness. On the other hand, behaviorally, even surveys that did not differ demographically were markedly and significantly different for some tweeted content and following behavior. This suggests that studying within-sample changes to estimate shifts in content may be more fruitful than attempting to study the prevalence of different topics of conversation by demographic re-weighting alone.

Unsurprisingly, the matched administrative data offer statistical power that is infeasible for a survey—even with groups that are relatively underrepresented. This statistical power allows for compositional analysis that is impossible in surveys with fewer respondents (Foucault Welles 2014). Even for relatively salient topics—such as tweeting about President Trump's impeachment and following particular accounts—surveys provide limited sample size for comparing online behaviors with external benchmarks. What's more, the top 0.1% of Twitter users, in terms of posting behavior, might be quite important; that 0.1% would constitute a subsample of 1,500 accounts in the matches to the voter records, and just 2 or so accounts in the surveys examined. Given meaningful selection concerns about who posts on any given topic—especially over time—analyses of *within-conversation* variation (e.g. subsetting to political conversations on Twitter rather than counting all sampled users equally in analyses) may need to restrict inferences to a more limited population, such as “Twitter users who talked about politics on Twitter just prior to the 2020 election”, and then compare this subset to the characteristics and behaviors of the excluded members of a sample.

These findings suggest complementary strengths of the two approaches. Survey-based approaches offer potential advantages in terms of representativeness, as well as capturing behavior in the bulk of the distribution. In addition, survey-based approaches would be necessary for linkage to information that can only be accessed via survey questions. Matched administrative data offer particular power with respect to subgroups—even those subgroups that may be underrepresented in the data. For example, minorities are underrepresented in the voter file sample, but there are still 111,495 Black and 67,890 Hispanic users in the sample; as compared to 85 and 282 in the American Trends Panel; and 200 and 252 in the KnowledgePanel.

These results suggest several avenues for further research. Can data collection be combined in ways that leverage their relative strengths? For example, could administrative data guide a behavior-targeted survey, allowing for robust surveys of subpopulations? And could survey-based approaches guide improvement of matching strategies that would reduce the biases in linked samples? How might account-linked survey questions best guide interpretation of online activities? To what extent do samples of accounts in existence at one point in time need to be continuously updated to accurately track social media populations? Finally, can a survey-based approach allow behaviorally driven inferences of the likely survey responses in a larger, administratively based, sample (for one example, see Barberá 2016)?

References

Ansolabehere, Stephen, and Eitan Hersh. 2012. "Validation: What Big Data Reveal About Survey Misreporting and the Real Electorate." *Political Analysis* 20 (4): 437–59. https://doi.org/10.1093/pan/mps023.

Bail, Christopher A., Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. "Exposure to Opposing Views on Social Media Can Increase Political Polarization." *Proceedings of the National Academy of Sciences* 115 (37): 9216–21. https://doi.org/10.1073/pnas.1804840115.

Bakshy, Eytan, Solomon Messing, and Lada A. Adamic. 2015. "Exposure to Ideologically Diverse News and Opinion on Facebook." *Science* 348 (6239): 1130–32. https://doi.org/10.1126/science.aaa1160.

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

———. 2016. "Less Is More? How Demographic Sample Weights Can Improve Public Opinion Estimates Based On Twitter Data." Working Paper.

Barberá, Pablo, Andreu Casas, Jonathan Nagler, Patrick J. Egan, Richard Bonneau, John T. Jost, and Joshua A. Tucker. 2019. "Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data." *American Political Science Review* 113 (4): 883–901. https://doi.org/10.1017/S0003055419000352.

Barberá, Pablo, and Gonzalo Rivero. 2015. "Understanding the Political Representativeness of

Twitter Users." *Social Science Computer Review* 33 (6): 712–29. https://doi.org/10.1177/0894439314558836.

Barberá, Pablo, and Zachary Steinert-Threlkeld. 2020. "How to Use Social Media Data for Political Science Research." In *The SAGE Handbook of Research Methods in Political Science and International Relations*, edited by Luigi Curini and Robert Franzese, 404–21. Thousand Oaks, CA: SAGE Publications.

Beauchamp, Nicholas. 2017. "Predicting and Interpolating State-Level Polls Using Twitter Textual Data." *American Journal of Political Science* 61 (2): 490–503. https://doi.org/10.1111/ajps.12274.

Berent, Matthew K., Jon A. Krosnick, and Arthur Lupia. 2016. "Measuring Voter Registration and Turnout in Surveys." *Public Opinion Quarterly* 80 (3): 597–621. https://doi.org/10.1093/poq/nfw021.

boyd, danah, Scott Golder, and Gilad Lotan. 2010. "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter." In *Proceedings of the 43rd Hawaii International Conference on System Sciences*, 1–10. IEEE. https://doi.org/10.1109/HICSS.2010.412.

Chalmers, Dan, Simon Fleming, Ian Wakeman, and Des Watson. 2011. "Rhythms in Twitter." In *Proceedings of the Third International Conference on Privacy, Security, Risk and Trust*, 1409–14. IEEE. https://doi.org/10.1109/PASSAT/SocialCom.2011.226.

Enamorado, Ted, Benjamin Fifield, and Kosuke Imai. 2019. "Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records." *American Political Science Review* 113 (2): 353–71. https://doi.org/10.1017/S0003055418000783.

Ernala, Sindhu Kiranmai, Moira Burke, Alex Leavitt, and Nicole B. Ellison. 2020. "How Well

Do People Report Time Spent on Facebook?: An Evaluation of Established Survey Questions with Recommendations.” In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. New York: ACM Press. https://doi.org/10.1145/3313831.3376435.

Fiesler, Casey, and Nicholas Proferes. 2018. “‘Participant’ Perceptions of Twitter Research Ethics.” *Social Media + Society* 4 (1). https://doi.org/10.1177/2056305118763366.

Foucault Welles, Brooke. 2014. “On Minorities and Outliers: The Case for Making Big Data Small.” *Big Data & Society* 1 (1): 1–2. https://doi.org/10.1177/2053951714540613.

Gelman, Andrew, Sharad Goel, Douglas Rivers, and David Rothschild. 2016. “The Mythical Swing Voter.” *Quarterly Journal of Political Science* 11 (1): 103–30. https://doi.org/10.1561/100.00015031.

Ghitza, Yair, and Andrew Gelman. 2013. “Deep Interactions with MRP: Election Turnout and Voting Patterns among Small Electoral Subgroups.” *American Journal of Political Science* 57 (3): 762–76. https://doi.org/10.1111/ajps.12004.

Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. “Fake News on Twitter During the 2016 U.S. Presidential Election.” *Science* 363 (6425): 374–78. https://doi.org/10.1126/science.aau2706.

Guess, Andrew, Kevin Munger, Jonathan Nagler, and Joshua Tucker. 2019. “How Accurate Are Survey Responses on Social Media and Politics?” *Political Communication* 36 (2): 241–58. https://doi.org/10.1080/10584609.2018.1504840.

Hamidi, Foad, Morgan Klaus Scheuerman, and Stacy M. Branham. 2018. “Gender Recognition or Gender Reductionism?: The Social Implications of Embedded Gender Recognition

Systems.” In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13. New York: ACM Press. https://doi.org/10.1145/3173574.3173582.

Henderson, Michael, Ke Jiang, Martin Johnson, and Lance Porter. Forthcoming. “Measuring Twitter Use: Validating Survey-Based Measures.” *Social Science Computer Review*. https://doi.org/10.1177/0894439319896244.

Hughes, Adam, Brad Jones, Alec Tyson, Emma Remy, and Aaron Smith. 2019. “National Politics on Twitter: Small Share of U.S. Adults Produce Majority of Political Tweets.” Washington, D.C.: Pew Research Center. https://www.people-press.org/2019/10/23/national-politics-on-twitter-small-share-of-u-s-adults-produce-majority-of-tweets/.

Jaidka, Kokil, Alvin Zhou, and Yphtach Lelkes. 2019. “Brevity Is the Soul of Twitter: The Constraint Affordance and Political Discussion.” *Journal of Communication* 69 (4): 345–72. https://doi.org/10.1093/joc/jqz023.

Japec, Lilli, Frauke Kreuter, Marcus Berg, Paul Biemer, Paul Decker, Cliff Lampe, Julia Lane, Cathy O’Neil, and Abe Usher. 2015. “Big Data in Survey Research: AAPOR Task Force Report.” *Public Opinion Quarterly* 79 (4): 839–80. https://doi.org/10.1093/poq/nfv039.

Keusch, Florian, Bella Struminskaya, Christopher Antoun, Mick P. Couper, and Frauke Kreuter. 2019. “Willingness to Participate in Passive Mobile Data Collection.” *Public Opinion Quarterly* 83 (S1): 210–35. https://doi.org/10.1093/poq/nfz007.

Klašnja, Marko, Pablo Barberá, Nick Beauchamp, Jonathan Nagler, and Joshua A. Tucker. 2018. “Measuring Public Opinion with Social Media Data.” In *The Oxford Handbook of Polling and Survey Methods*, edited by Lonna Rae Atkeson and R. Michael Alvarez, 555–82. New York: Oxford University Press.

Kreuter, Frauke, Georg-Christoph Haas, Florian Keusch, Sebastian Bähr, and Mark Trappmann. 2020. "Collecting Survey and Smartphone Sensor Data With an App: Opportunities and Challenges Around Privacy and Informed Consent." *Social Science Computer Review* 38 (5): 533–49. https://doi.org/10.1177/0894439318816389.

Lazer, David. 2020. "Studying Human Attention on the Internet." *Proceedings of the National Academy of Sciences* 117 (1): 21–22. https://doi.org/10.1073/pnas.1919348117.

Nagler, Jonathan, and Joshua A. Tucker. 2015. "Drawing Inferences and Testing Theories with Big Data." *PS: Political Science & Politics* 48 (1): 84–88. https://doi.org/10.1017/S1049096514001796.

Pasek, Josh, H. Yanna Yan, Frederick G. Conrad, Frank Newport, and Stephanie Marken. 2018. "The Stability of Economic Correlations over Time." *Public Opinion Quarterly* 82 (3): 470–92. https://doi.org/10.1093/poq/nfy030.

Raji, Inioluwa Deborah, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. 2020. "Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 145–51. New York NY USA: ACM Press. https://doi.org/10.1145/3375627.3375820.

Schober, Michael F., Josh Pasek, Lauren Guggenheim, Cliff Lampe, and Frederick G. Conrad. 2016. "Social Media Analyses for Social Measurement." *Public Opinion Quarterly* 80 (1): 180–211. https://doi.org/10.1093/poq/nfv048.

Sloan, Luke, Curtis Jessop, Tarek Al Baghal, and Matthew Williams. 2020. "Linking Survey and

Twitter Data: Informed Consent, Disclosure, Security, and Archiving." *Journal of Empirical Research on Human Research Ethics* 15 (1–2): 63–76. https://doi.org/10.1177/1556264619853447.

Tufekci, Zeynep. 2014. "Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls." In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, 505–14. Palo Alto, CA: AAAI Press.

Williams, Matthew L., Pete Burnap, Luke Sloan, Curtis Jessop, and Hayley Lepps. 2017. "Users' Views of Ethics in Social Media Research: Informed Consent, Anonymity, and Harm." In *Advances in Research Ethics and Integrity*, edited by Kandy Woodfield, 2:27–52. Emerald Publishing.

Wojcik, Stefan, Avleen Bijral, Richard Johnston, Juan Miguel Lavista, Gary King, Ryan Kennedy, Alessandro Vespignani, and David Lazer. Forthcoming. "Survey Data and Human Computation for Improved Flu Tracking." *Nature Communications*. https://j.mp/2X10j2U.

Wojcik, Stefan, Solomon Messing, Aaron Smith, Lee Rainie, and Paul Hitlin. 2018. "Bots in the Twittersphere." Washington, DC: Pew Research Center.

Wu, Patrick Y., Walter R. Mebane Jr., Logan Woods, Joseph Klaver, and Preston Due. 2019. "Partisan Associations of Twitter Users Based on Their Self-Descriptions and Word Embeddings." Working Paper. http://www-personal.umich.edu/~wmebane/partisanassociations_wumebanewoodsklaverdue_apsa2019.pdf.

List of Figures

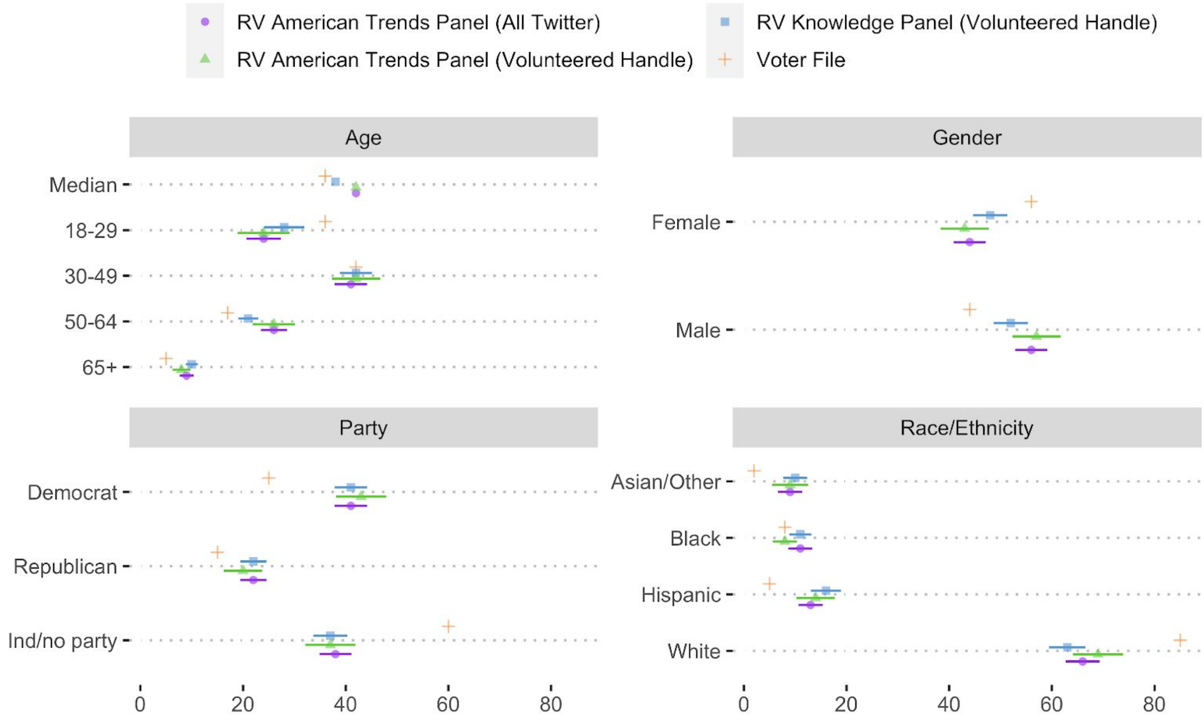

Figure 1. Demographic Attributes of Twitter Samples. The attributes of Twitter users are broadly consistent across sources. Several major differences in the survey versus voter file demographics are due to differences in registered voters (as recorded in administrative data) and U.S. adults in the surveys—see Table 1 and Figure 2. 95% confidence intervals shown. Full results available in Online Appendix A7. All estimates in this figure other than median age are percentage points. Party estimates from the voter files are based on party registration, while the surveys use self-reports of party identification.

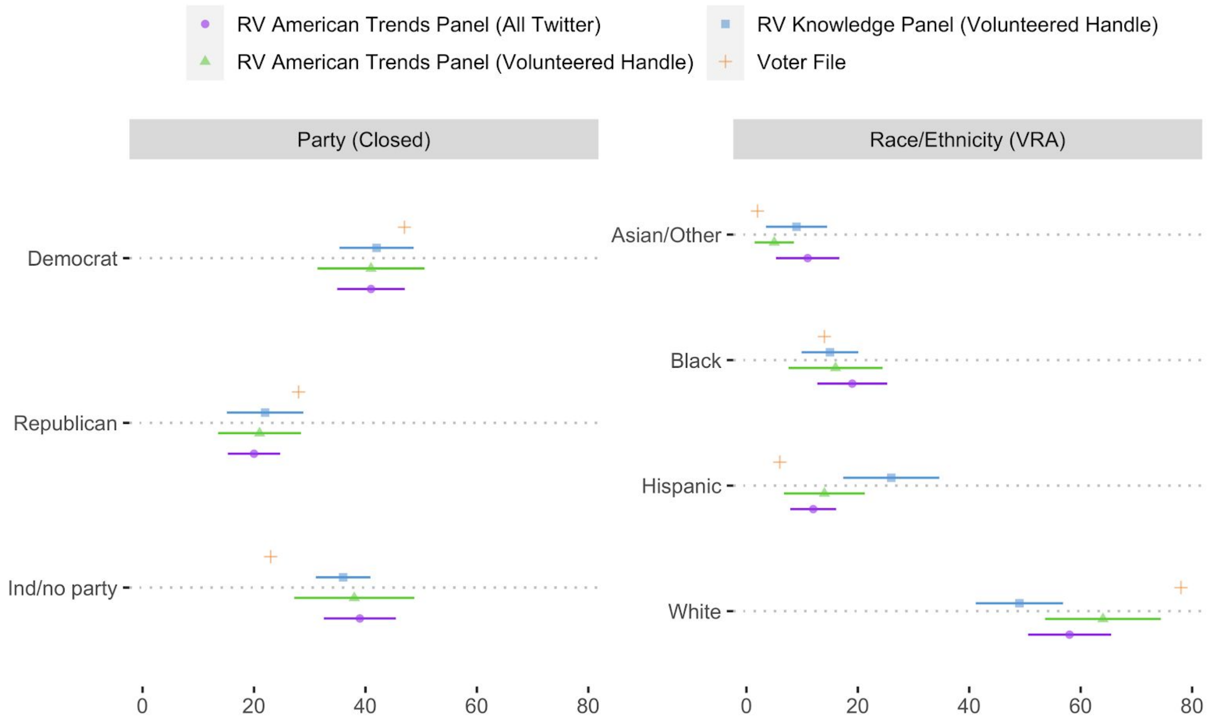

Figure 2. Demographic Attributes of Twitter Samples (Subset). Close party registration states and states included in the VRA show smaller differences across sources. Here, party estimates from the voter files are based on party registration, while the surveys use self-reports of party identification. 95% confidence intervals shown. Full results available in Online Appendix A7.

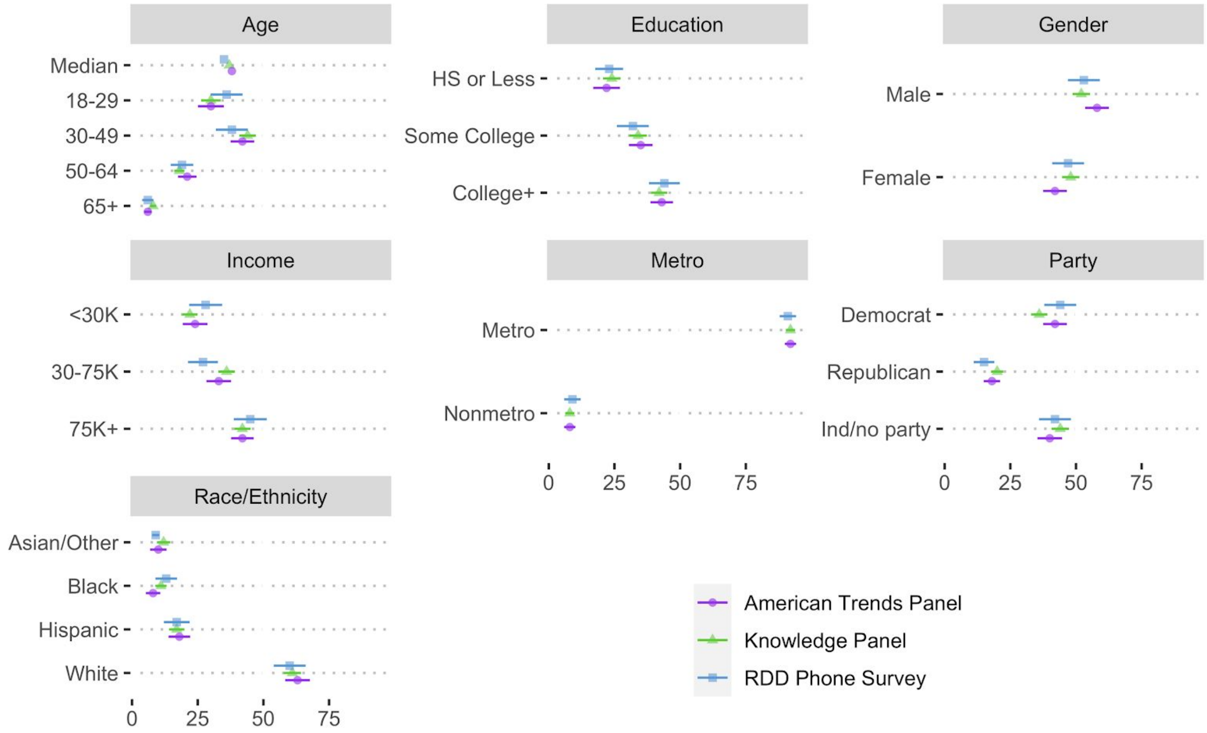

Figure 3. Demographic Attributes of Survey Samples and RDD Phone Poll. The attributes of Twitter users are highly consistent across two samples from online panels that volunteered handles and a separate RDD phone poll. 95% confidence intervals shown. Full results available in Online Appendix A7.

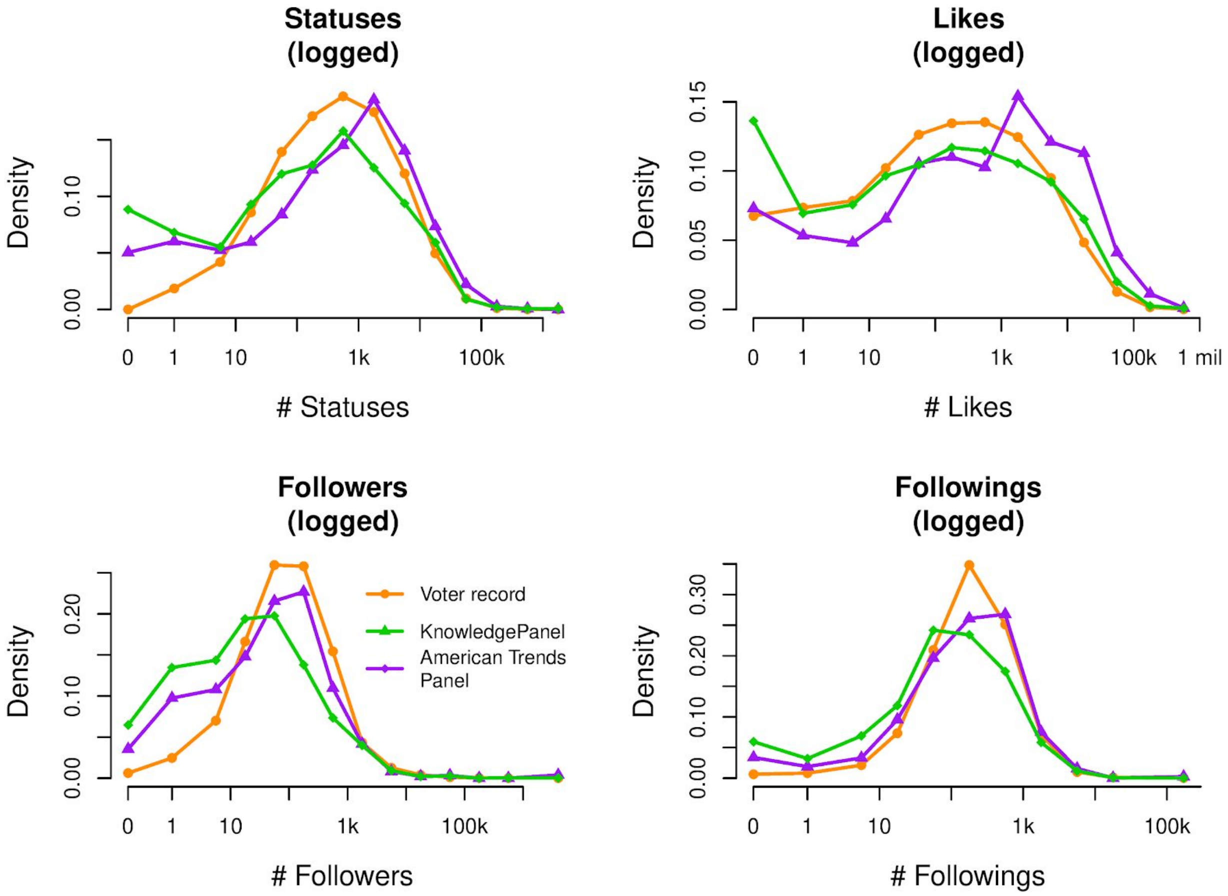

Figure 4. Activity Distributions of Twitter Samples. The x-axis uses a log-10 scale and the y-axis is the proportion of individuals for a given sample in an activity bin.

Tables

Population target: registered voters					
Demographics	**Voter File Sample** (RV)			**American Trends Panel** Weighted November 2019 (all RV)	
N	202,647,162[14]			10,063	
Age		*Unique name, state*			
Median	*50*	*48*		*53*	
18-29	17	20		12 (0.6)	
30-49	32	32		31 (0.7)	
50-64	27	25		29 (0.7)	
65+	23	23		27 (0.7)	
Race/Ethnicity	*All*	*VRA*	*VRA, unique*	*All*	*VRA*
White	78	69	71	70 (0.7)	59 (1.8)
Black	11	18	19	11 (0.5)	21 (1.6)
Hispanic	8	11	7	11 (0.5)	12 (1.2)
Asian/Other	2	2	3	7 (0.4)	6 (0.9)
Gender		*Unique name, state*			
Male	47	40		48 (0.8)	
Female	53	60		52 (0.8)	
Party	*All*	*Closed primary*	*Closed, unique*	*All*	*Closed*
Democrat	23	44	45	34 (0.7)	35 (1.4)
Republican	17	31	24	31 (0.7)	30 (1.3)
Ind/no party	60	24	30	35 (0.7)	35 (1.5)
Urbanicity					
Metro	84			86 (0.6)	
Nonmetro	16			14 (0.6)	

Table 1: Demographic Characteristics of Starting Samples (all US adults registered to vote). This table compares the demographic composition of each of our starting samples,

[14] This number likely includes duplicate registrations that are *later* removed by our unique name within geographic location restriction. We include estimates for demographics after that restriction under the sub-column “Unique name, state.”

before subsetting to Twitter users, for which we have voter registration information. The closed primary sub-column in this table includes states which allow unaffiliated voters to participate but prohibit votes from the opposing party. All values are percentages; standard errors for survey estimates presented in parentheses.

Keywords:	Voter File Sample	KnowledgePanel	American Trends Panel
“Impeach”	Users: 13% N: 78,428 Median: 2 Mean: 18	Users: 12% N: 208 Median: 2 Mean: 14	Users: 18% N: 178 Median: 3 Mean: 11
“Trump”	Users: 28% N: 165,318 Median: 3 Mean: 34	Users: 20% N: 323 Median: 3 Mean: 39	Users: 33% N: 310 Median: 4 Mean: 27
“Republican”	Users: 10% N: 62,713 Median: 2 Mean: 8	Users: 9% N: 164 Median: 2 Mean: 11	Users: 18% N: 165 Median: 2 Mean: 6
“Democrat”	Users: 13% N: 74,986 Median: 2 Mean: 11	Users: 10% N: 172 Median: 2 Mean: 11	Users: 20% N: 162 Median: 2 Mean: 11
“#TBT”	Users: 3% N: 17,331 Median: 1 Mean: 2	Users: 1% N: 16 Median: 1 Mean: 2	Users: 1% N: 10 Median: 1 Mean: 1

Table 2: Keyword prevalence comparison of voter file and survey samples. Median and mean are conditional on 1 or more mentions of each keyword. Keyword searches are case-insensitive.