Supplementary Material for "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

Contents:

Online Appendix A1: Demographic Attributes of Compliers and Non-compliers

For both surveys, researchers asked respondents to provide their handles (question wording appears in Online Appendix A3. We reviewed all submitted handles, removing those that did not appear valid. The following criteria were used when making this determination:

1. The handle is not an obvious refusal or invalid answer (such as "no" or "none")

2. The account is accessible (meaning it exists and has not been suspended)

3. The account appears to belong to someone located in the U.S.

4. The account appears to correspond with the age and gender reported by the respondent (and, if age and gender was not apparent, state of residence)

5. The account is not an institutional or organizational account

American Trends Panel	Provided a **valid** Twitter handle	Provided an **invalid** Twitter handle	Uses Twitter but did **not** provide any Twitter handle
N	1,345	172	1,044
Age 18-29 30-49 50-64 65+	 18 44 27 11	 15 44 27 15	 21 39 28 12

Race White Black Hispanic Asian/Other	 63 7 24 6	 54 11 27 6	 62 9 21 7
Gender Male Female	 51 49	 46 54	 48 52
Party Democrat Republican Independent/Other	 49 20 32	 40 20 40	 43 23 34
Education HS or Less Some College College+	 8 28 64	 17 35 48	 10 27 63
Income <30K 30-75K 75K+	 16 29 54	 24 34 42	 16 30 49
Urban Metro Nonmetro	 93 7	 93 7	 91 9

Table A1. *American Trends Panel demographic comparison of Twitter handle volunteers* (unweighted percentages)

American Trends Panel	Provided a **valid** Twitter handle	Provided an **invalid** Twitter handle	Uses Twitter but did **not** provide any Twitter handle
Share of Respondents	51 (1.6)	7 (0.8)	42 (1.6)
Age 18-29 30-49 50-64	 32 (2.4) 42 (2.2) 20 (1.6)	 19 (5.3) 45 (5.4) 23 (4.2)	 34 (2.6) 39 (2.4) 20 (1.7)

65+	6 (0.7)	13 (3.2)	6.7 (0.8)
Race White Black Hispanic Asian/Other	 62 (2.3) 8 (1.3) 18 (2.0) 11 (1.7)	 51 (5.5) 18 (4.4) 21 (5.2) 6 (2.1)	 57 (2.6) 13 (1.9) 19 (2.3) 10 (1.6)
Gender Male Female	 57 (2.2) 43 (2.2)	 56 (5.2) 43 (5.2)	 53 (2.5) 47 (2.5)
Party Democrat Republican Independent/Other	 42 (2.2) 18 (1.5) 40 (2.3)	 39 (5.5) 20 (4.1) 41 (5.3)	 37 (2.3) 21 (1.9) 42 (2.5)
Education HS or Less Some College College+	 22 (2.5) 36 (2.2) 41 (2.0)	 33 (5.9) 39 (5.2) 28 (4.1)	 27 (2.8) 32 (2.3) 40 (2.2)
Income <30K 30-75K 75K+	 25 (2.3) 32 (2.2) 43 (2.1)	 28 (5.1) 37 (5.5) 35 (4.9)	 22 (2.3) 33 (2.5) 38 (2.2)
Urban Metro Nonmetro	 92 (1.1) 8 (1.1)	 90 (3.5) 10 (3.5)	 90 (1.6) 10 (1.6)

Table A2. *American Trends Panel demographic comparison of Twitter handle volunteers* (weighted percentages, with standard errors)

The KnowledgePanel sample included only 3 respondents who said they use Twitter, but did not provide a handle; that group is excluded.

KnowledgePanel	Provided a **valid** Twitter handle	Provided an **invalid** Twitter handle

N	2,808	485
Age 18-29 30-49 50-64 65+	 11 39 33 16	 13 41 28 18
Race White Black Hispanic Asian/Other	 75 9 11 6	 72 11 9 8
Gender Male Female	 45 56	 42 58
Party Democrat Republican Independent	 40 25 35	 36 23 41
Education HS or Less Some College College+	 11 30 59	 14 30 56
Income <30K 30-75K 75K+	 16 34 50	 21 35 44
Urban Metro Nonmetro	 90 10	 88 12

Table A3. *KnowledgePanel demographic comparison of Twitter handle volunteers* (unweighted)

Online Appendix A2: Supplementary Behavioral Statistics

Behaviors (Metadata)	Voter File Sample	KnowledgePanel	American Trends Panel
Median Metadata Estimates	Tweets 431 Favs: 173 Followers: 92 Following: 192	Tweets: 182 Favs: 113 Followers: 27 Following: 93	Tweets: 573 Favs: 457 Followers: 58 Following: 169
Share of All Tweets from top 10% of Tweeters (metadata)	69%	83%	73%
Account age (Median Days)	2930	2821	2716

Table A4. *Metadata behavioral comparison.* Activity numbers for each of the samples using metadata rather than active tweet collection.

Online Appendix A3: Twitter Handle Question Wording

KnowledgePanel:

We would like to better understand the role of Twitter in society. In order to do that it would be very helpful if you would share your Twitter handle with us. The handle is the username you have selected your Twitter account. Handles will be used for research purposes only. We won't use it to contact you and we won't share it with anyone for marketing purposes.

Please list your Twitter handle in the box below.

American Trends Panel:

We would like to better understand the role of Twitter in society; for example, how often people use Twitter and what they tweet about. To help us with this research, we hope you will share your Twitter handle with us. Your handle is the username you have selected for your Twitter account (like @yourhandle). We will treat your handle with the same care we take with all of your survey responses - it will be used only for research purposes, and we will never share any of your tweets or any profile data that can be linked back to you. We expect that this research will be concluded within 18 months.

Please list your Twitter handle in the box below.

If at any point you wish to opt out of research related to your Twitter account, contact info@pewresearch.org. For more information about our privacy practices, please see our privacy policy [link].

Online Appendix A4: Linking Twitter users with the voter file

To match Twitter accounts to the voter file, we expand upon a method first presented in (Grinberg et al. 2019). The basic idea is that, if an account provides name and location information, we should be able to reference that information against the voter file.

To do this, we begin with a dataset of 290 million user profiles collected from the Twitter Decahose between January 2014 and March 2017. We first extract the name and location from the user profiles, and discard accounts that list clearly non-US locations. The name used for matching can be drawn from either the profile's account name or the display name, and must be matched to the voter file exactly. The location is drawn from the location field of the profile; if there is no location information for the account, it cannot be matched.

The complete US voter file was obtained from TargetSmart and delivered in October 2017. To make the age estimates consistent with the other two samples analyzed, we added two years to everyone's age, bringing them forward to fall 2019. In general, we relied on attributes drawn from the voter file directly; where attributes are missing TargetSmart may also draw on commercial data vendors to supplement.

For each individual in the voter file, we generate a list of "candidate matches" of Twitter accounts with the same name. If there were zero or more than ten such candidate matches, we disallow a match. If there are between one and ten candidate matches, we then examine the locations of the candidate matches. First, we try to parse the location using U.S. Census gazetteer

data, and if there is no match, we then check each profile for the city and state of the candidate matches as listed in the voter record. To see if we can narrow down candidates to only one match, we then use a classifier trained on profile features to identify and exclude accounts that are very likely outside of the United States. This step allows us to exclude accounts with blank locations from interfering with an otherwise unique match in the United States. After this location matching, if there is a unique match at the city level, we allow the match. After considering the city level, we perform the same evaluation at the state level and match individuals with unique state-name combinations. Finally, we removed matches where the matching procedure produced duplicate Twitter accounts for a single voter record.

For classification, we used a conditional inference tree (Hothorn, Kornik, and Zelles 2006), as implemented in the R library *party*. The specific features used in the classifier are: account language, most common tweet language, time zone offset, account age, verified status, protected account status, number of followers, number of accounts followed, and number of Tweets. The classifier had an AUC of 0.84; we experimented with simpler methods like Naive Bayes, which had comparable but slightly lower AUCs (0.82).

To motivate this matching, consider a few fictitious examples.

1. There are 100 John Does on Twitter, none of whom report their location. In the voter file, we encounter a John Doe in Wyoming. Because there are more than ten candidate matches, we are unable to perform a match.
2. There are three Jane Does on Twitter; one resides in Boston, Massachusetts; one resides in Amherst, Massachusetts; and the last resides in New York City. In the voter file we see a Jane

Doe in Amherst, and we match it to the Twitter account, because the matching is performed first at the city level, then the state level.

3. There are still three Jane Does in Boston, Amherst, and New York City. In the voter file we see a Jane Doe in New York City, and because the name-state combination is unique, we match that record in the voter file to the Twitter account.

One challenge when performing this matching is the use of informal language on Twitter. We used a gazetteer to map abbreviations and casual references to cities. So, for example, “NYC” is mapped to “New York, New York”; “Philly” is “Philadelphia, PA”; and so on.

Online Appendix A5: Analysis of self-report/voter file sample overlap

In this appendix, we provide summary statistics on the users who appeared in both the KnowledgePanel or the American Trends Panel and in the voter file sample (overall N = 182). An important caveat to this analysis is that, where the two approaches disagree, we are unable to adjudicate the disagreement by manual inspection of an individual's Twitter account. This is precluded by data-sharing agreements between the organizations who have collected the data. In order to protect the personally identifiable information provided by respondents, members of the research team who collected the survey data did not have access to account information for members of the voter file sample, and vice versa.

Survey	N	Size of intersection
KP	2,369	109
ATP	1,188	73

Table A5: *Size of the intersections between voter file sample and each survey.*

ATP / Voter File	M	F	Unknown
M	33	2	4
F	0	31	2

KP / Voter File	M	F	Unknown
M	39	2	1
F	1	59	1

Unknown	0	0	5

Table A6: *Confusion matrix: self-reported gender versus administrative data*.

ATP / Voter File	Black	White	Hispanic	Other	Unknown
Black	2	1	0	1	1
White	0	46	1	1	1
Hispanic	0	2	7	0	1
Other	0	3	0	1	2

KP / Voter File	Black	White	Hispanic	Other	Unknown
Black	3	3	0	0	1
White	1	83	0	0	2
Hispanic	0	3	4	0	0
Other	1	4	0	1	2

Table A7: *Confusion matrix: self-reported race versus administrative data*.[1]

ATP / Voter File	Dem	Rep	Ind	Other	Unknown
Dem	13	1	16	3	0
Rep	1	7	7	1	0
Ind	3	1	9	2	3
Other	0	2	4	0	0

[1] If we reproduce this analysis for VRA states, as in the body of the paper, the numbers are too small to report separately for each survey. Pooling the ATP and KP samples, we observe 22 individuals in the intersection. Of these, from the perspective of the voter file, 19 are white, 1 is Hispanic, and 2 are Black. The survey results differ only in finding that one of the Black respondents self-identified as White.

KP / Voter File	Dem	Rep	Ind	Other	Unknown
Dem	20	1	25	0	2
Rep	2	10	8	0	3
Ind	3	5	18	1	1
Other	0	2	5	0	2

Table A8: *Confusion matrix: self-report party identification versus party registration in administrative data*.

We also evaluate the degree to which survey respondents would be likely to be matchable with the voter file by evaluating whether each respondent's profile contained an apparently valid name and location. Overall, 43% of respondents to the KnowledgePanel survey and 50% of respondents to the American Trends Panel survey had apparently valid names and locations, though the fact that many of these names are commonly held means that we cannot estimate the exact number who could be successfully matched. We also examined racial disparities among these potential matches: among those identifying as Black, 41% used apparently valid names and locations in the American Trends Panel and 34% did the same in the KnowledgePanel survey. Similarly, 40% of Hispanics in the ATP had apparently valid names and locations, while 39% of Hispanics in the KP survey did the same. These rates are lower than the rates observed for white respondents: 51% (ATP) and 40% (KP).

Online Appendix A6: Survey Weighting and Response Rates

Both surveys included in this analysis are weighted to approximate a national sample of U.S. adults on Twitter. However, since there is no administrative data that provides population-level estimates for this group, we took two separate approaches when weighting the data.

The sample drawn from Ipsos' KnowledgePanel was raked to match population totals estimated using Wave 39 of Pew Research Center's American Trends Panel (ATP) which was conducted in November 2018 and featured a question asking respondents if they used Twitter. Methodological details for that survey can be found in the accompanying Pew Research Center report (see: Pew Research Center 2018). Population totals for adult Twitter users in the U.S. were estimated by filtering down to self-identified Twitter users and calculating weighted estimates for age, sex, education, race/Hispanic origin, census region, metropolitan status, voter registration, and party identification. Using the Ipsos provided base weight as a starting point, the KP sample was then raked to match these estimated population parameters.

The second sample is from a separate wave of the American Trends Panel, conducted Oct. 29–Nov. 11, 2019. Weighting parameters for the full sample of U.S. adults come from the 2017 American Community Survey (gender, age, education, race/Hispanic origin, country of birth among Hispanics, and home internet access), the 2018 CPS March supplement (region by metropolitan status), the 2017 CPS Volunteering and Civic Life Supplement (volunteerism), the 2016 CPS Voting and Registration supplement (voter registration), and an average of the three

most recent Pew Research Center telephone surveys (self-reported party identification). For the analysis of Twitter users within this sample, we filter to those who report using the platform and retain the original weights.

The cumulative response rate (accounting for both nonresponse and attrition) for the American Trends Panel wave used to collect Twitter handles is 4.5%. The cumulative response rate for the KnowledgePanel survey is 4.6%. The cumulative response rate for the American Trends Panel wave that was used to create estimated population targets for the purposes of weighting the KnowledgePanel sample is 3.7%.

The 2019 phone poll was conducted by Abt Associates on behalf of Pew Research Center. Its response rate (AAPOR RR3) was 8.1% for the landline sample and 6.0% for the cell sample.

Online Appendix A7: Detailed Tables

Population target: US Twitter users registered to vote				
Demographics	**Voter File Sample** Registered Voters	**KnowledgePanel** Registered Voters on Twitter (providing handle)	**American Trends Panel** Registered Voters on Twitter (regardless of handle provision)	**American Trends Panel** Registered Voters on Twitter (providing handle)
N	1,496,433	2,162	2,138	1,018
Age				
Median	*36*	*38*	*42*	*42*
18-29	36	28 (2.0)	24 (1.7)	24 (2.6)
30-49	42	42 (1.6)	41 (1.6)	42 (2.4)
50-64	17	21 (1.0)	26 (1.3)	26 (2.1)
65+	5	10 (0.6)	9 (0.7)	8 (0.9)
Race/Ethnicity	*All* *VRA*	*All* *VRA*	*All* *VRA*	*All* *VRA*
White	85 78	63 (1.8) 49 (4.0)	66 (1.7) 58 (3.8)	69 (2.5) 64 (5.3)
Black	8 14	11 (1.1) 15 (2.6)	11 (1.2) 19 (3.2)	8 (1.2) 16 (4.3)
Hispanic	5 6	16 (1.5) 26 (4.4)	13 (1.2) 12 (2.1)	14 (1.9) 14 (3.7)
Asian/Other	2 2	10 (1.2) 9 (2.8)	9 (1.2) 11 (2.9)	9 (1.8) 5 (1.8)
Gender				
Male	44	52 (1.7)	56 (1.6)	57 (2.4)
Female	56	48 (1.7)	44 (1.6)	43 (2.4)
Party	*All* *Closed*	*All* *Closed*	*All* *Closed*	*All* *Closed*
Democrat	25 47	41 (1.6) 42 (3.4)	41 (1.6) 41 (3.1)	43 (2.5) 41 (4.9)
Republican	15 28	22 (1.3) 22 (3.5)	22 (1.3) 20 (2.4)	20 (1.9) 21 (3.8)
Ind/no party	60 23	37 (1.7) 36 (2.5)	38 (1.6) 39 (3.3)	37 (2.5) 38 (5.5)

Urbanicity				
Metro	89	93 (0.8)	90 (1.0)	91 (1.3)
Nonmetro	11	7 (0.8)	10 (1.0)	9 (1.3)

Table A9: *Demographic characteristics of voter file users and survey respondents who use Twitter and are registered to vote.* This table compares the demographic composition of the voter-file sample approach to the survey-based approach, including comparisons for both Twitter use and handle provision responses in the survey. Standard errors for survey estimates appear in parentheses.

Population target: US adult Twitter users			
Demographics	**Knowledge Panel** November 2018	**American Trends Panel** November 2019	**RDD Phone Poll Subset to Twitter Users** January 2019
N	2,369	1,188	1,502 (327 Twitter users)
Age			
Median	*37*	*38*	*35*
18-29	30 (1.9)	30 (2.5)	36 (3.1)
30-49	44 (1.6)	42 (2.3)	38 (3.1)
50-64	18 (1.0)	21 (1.8)	19 (2.2)
65+	8 (0.6)	6 (0.7)	6 (1.1)
Race/Ethnicity			
White	61 (1.7)	63 (2.4)	60 (3.1)
Black	11 (1.1)	8 (1.4)	13 (2.1)
Hispanic	17 (1.5)	18 (2.1)	17 (2.5)
Asian/Other	12 (1.3)	10 (1.6)	9 (0.7)
Gender			
Male	52 (1.7)	58 (2.3)	53 (3.1)
Female	48 (1.7)	42 (2.3)	47 (3.1)
Party			
Democrat	36 (1.6)	42 (2.3)	44 (3.1)
Republican	20 (1.2)	18 (1.6)	15 (2.0)
Ind/no party	44 (1.7)	40 (2.4)	42 (3.1)
Education			

HS or Less	24 (1.7)	22 (2.6)	23 (2.7)
Some College	34 (1.7)	35 (2.3)	32 (3.1)
College+	42 (1.6)	43 (2.2)	44 (3.0)
Income			
<30K	22 (1.5)	24 (2.4)	28 (3.2)
30-75K	36 (1.6)	33 (2.4)	27 (2.9)
75K+	42 (1.6)	42 (2.2)	45 (3.2)
Urbanicity			
Metro	92 (0.9)	92 (1.1)	91 (1.6)
Nonmetro	8 (0.9)	8 (1.1)	9 (1.6)

Table A10: *Demographic characteristics of survey respondents who provided Twitter handles and Twitter users from RDD phone survey.* This table compares the demographic composition of the two online survey samples with a sample from an RDD phone poll. Standard errors for survey estimates appear in parentheses.

Online Appendix A8: Comparison of Random Tweets with Voter Record Linked Tweets

In this section, we compare the hashtags appearing in the Twitter "Decahose"—the 10% random sample of tweets—for US-based accounts to hashtags from the US-based accounts linked to voter records. We use Twitter's "profile geo"[2] enrichment for determining that an account is US-based. We further compare two sets of data 1) all US-based accounts, and 2) US-based accounts with first and last names appearing (separately) in the voter record linked data—i.e. accounts with "real" names. "Real" names here are first and last names appearing more than 10 times in the voter file sample[3]—"John Smith" is a real name if both "John" and "Smith" appear in the list of voter file names appearing more than 10 times. The goal of the real name analysis is to limit comparisons to accounts that appear to represent people rather than organizations, online personalities such as famous pets, or clearly anonymous accounts.

For 3 months starting May 7, 2020 (our first full day of access to the Decahose), we count the number of distinct users posting any given hashtag. Using the number of distinct users for accounts not in the voter file sample versus those included in the sample, we calculate the ratio of distinct users by hashtag for not-in-sample versus in-sample and in-sample versus not-in-sample. For reference, on the last day of this analysis (August 7, 2020), 5% of users appearing in the Decahose with US-based location and "real" name were in the sample, and 3% of users appearing in the Decahose with US-based location were in the sample. 20% of US-based

[2] https://developer.twitter.com/en/docs/tweets/enrichments/overview/profile-geo
[3] We use a minimum of 10 here to avoid selecting on particularly distinctive names.

users in the Decahose had "real" names, and 40% of sampled users in the Decahose had names in the list of common "real" names[4].

Below, in Tables A11 and A12, we show the hashtags with the largest ratios for appearing in one set but not the other. These hashtags are limited to those with mutual information scores[5] higher than 0.01 (for "real" name) and 0.02 (for all US-based accounts).[6]

These distinctive hashtags suggest that tweets appearing outside the sample are more likely to post about video games and entertainment, pornography[7], conservative or polarizing topics, and conspiracies, as well as alternative hashtags to the official #blacklivesmatter hashtag. The prevalence of conservative topics perhaps suggests some far-right or conservative manipulation of hashtags on Twitter, but it is also possible that conservatives are more likely to post polarizing topics using anonymous accounts or with accounts not using their real name or as it appears in a voter record. We further show in the overall comparison that "BTS" (a popular Korean boy band) was a top hashtag outside the sample, and K-pop fans were reported to have flooded anti-protest and anti-Black Lives Matter hashtags in support of the Black Lives Matter movement.[8]

---

[4] This analysis also used a faster name parsing method (splitting the Twitter name field on spaces, and removing symbols) than the original matching to ensure that this paper's revision was completed before the resubmission deadline.

[5] Manning, Christopher D., Prabhakar Raghavan and Hinrich Schütze. 2008. Introduction to Information Retrieval. NY: Cambridge University Press.

[6] This ensures that a minimal number of people in the sample have posted distinctive hashtags there. For the "real" name comparison, for example, all distinctive sample hashtags were posted more than 200 times. The higher value for US-based accounts limits the number of K-pop and pornography-related hashtags.

[7] Filtering on a smaller mutual information score, most to all of the distinctive hashtags from users in the non-sample subsetsample are related to pornography.

[8] https://www.bbc.com/news/world-asia-52996705

Conversely, users in the sample appear more likely to post about the COVID-19 pandemic, sweepstakes and commerce, and, among several liberal or social justice related topics, (support for) the Black Lives Matter movement—despite the sample being more White and no more liberal or conservative than Twitter overall according to the panel surveys and their RDD phone survey comparison. Over shorter time frames (e.g. 1 month), we further see everyday life and holiday-related hashtags appear more often in the sample (e.g. "#MemorialDay", "GraduateTogether") than outside it.

Not in Sample	**Ratio: Not Sample / Sample**	**In Sample**	**Ratio: Sample / Not Sample**
#onlyfans	5.03	#pandemic	1.54
#RT	2.31	#shopmycloset	1.51
#TheGreatAwakening	2.02	#covid19	1.45
#AntifaTerrorists	1.98	#JohnLewis	1.43
#PS4live	1.93	#SCOTUS	1.37
#Anonymous	1.91	#WearAMask	1.37
#NewProfilePic	1.85	#Juneteenth	1.33
#BenghaziAintGoingAway	1.8	#Covid19	1.32
#2A	1.79	#SayHerName	1.32
#TheMoreYouKnow	1.73	#blackouttuesday	1.3
#Qanon	1.72	#realestate	1.26
#Flynn	1.7	#ObamaDayJune14th	1.25

#KAG	1.7	#giveaway	1.25
#QAnon	1.68	#COVID19	1.24
#WWG1WGA	1.68	#DACA	1.23
#Trump2020Landslide	1.62	#COVID	1.23
#OBAMAGATE	1.62	#NYC	1.22
#BillClintonIsAPedo	1.58	#Texas	1.22
#PS5	1.56	#LGBTQ	1.22
#KillerCuomo	1.56	#BreonnaTaylor	1.19

Table A11: *Comparison of randomly sampled tweets with voter record linked tweets ("real" names subset).* This table compares distinctive hashtags in the Twitter "Decahose" for US-based accounts appearing in the voter-file-linked sample versus not. This table is restricted to US-based accounts posting under first and last names that appear (separately) in the voter file linked sample.

Not in Sample	**Ratio: Not Sample / Sample**	**In Sample**	**Ratio: Sample / Not Sample**
#horny	11.36	#covid19	1.66
#nsfw	8.61	#Covid19	1.57
#onlyfans	6.8	#COVID	1.48
#gay	5.44	#COVID19	1.47
#MTVHottest	5.41	#DefundThePolice	1.41
#방탄소년단	5.31	#Juneteenth	1.41
#JIMIN	5.26	#TrumpIsNotWell	1.35
#BTS	4.36	#coronavirus	1.34
#RT	2.81	#Covid_19	1.31

#Anonymous	2.51	#TrumpVirus	1.27
#PS4live	2.25	#Biden2020	1.26
#NewProfilePic	1.92	#SayHerName	1.24
#BlackLivesMatters	1.8	#Trump	1.24
#PS4share	1.75	#BREAKING	1.19
#BLACK_LIVES_MATTER	1.62	#BreonnaTaylor	1.17
#NintendoSwitch	1.6	#BlackOutTuesday	1.14
#BlackIsKing	1.57	#Coronavirus	1.13
#PS5	1.56	#blacklivesmatter	1.08
#ACNH	1.52	#Minneapolis	1.06
#AnimalCrossing	1.47	#GeorgeFloyd	1.06

Table A12: *Comparison of randomly sampled tweets with voter record linked tweets (all US-based accounts).* This table compares distinctive hashtags in the Twitter "decahose" for US-based accounts appearing in the voter-file-linked sample versus not. This table uses a slightly higher mutual information score cutoff (0.02). The lower value of 0.01 returns distinctive hashtags for the non-sampled accounts that are all either about pornography or K-pop.

Online Appendix A9: Comparison of Twitter Users in Survey Samples by Voter Registration

Demographics	**KnowledgePanel** Twitter Users (providing handle)	**KnowledgePanel** Registered Voters on Twitter (providing handle)	**American Trends Panel** Twitter users (providing handle)	**American Trends Panel** Registered Voters on Twitter (providing handle)
N	2,369	2,162	1,188	1,018
Age				
Median	*37*	*38*	*38*	*42*
18-29	30 (1.9)	28 (2.0)	30 (2.5)	24 (2.6)
30-49	44 (1.6)	42 (1.6)	42 (2.3)	42 (2.4)
50-64	18 (1.0)	21 (1.0)	21 (1.8)	26 (2.1)
65+	8 (0.6)	10 (0.6)	6 (0.7)	8 (0.9)
Race/Ethnicity				
White	61 (1.7)	63 (1.8)	63 (2.4)	69 (2.5)
Black	11 (1.1)	11 (1.1)	8 (1.4)	8 (1.2)
Hispanic	17 (1.5)	16 (1.5)	18 (2.1)	14 (1.9)
Asian/Other	12 (1.3)	10 (1.2)	10 (1.6)	9 (1.8)
Gender				
Male	52 (1.7)	52 (1.7)	58 (2.3)	57 (2.4)
Female	48 (1.7)	48 (1.7)	42 (2.3)	43 (2.4)
Party				
Democrat	36 (1.6)	41 (1.6)	42 (2.3)	43 (2.5)
Republican	20 (1.2)	22 (1.3)	18 (1.6)	20 (1.9)
Ind/no party	44 (1.7)	37 (1.7)	40 (2.4)	37 (2.5)
Urbanicity				
Metro	92 (0.9)	93 (0.8)	90 (1.0)	91 (1.3)
Nonmetro	8 (0.9)	7 (0.8)	10 (1.0)	9 (1.3)

Table A13: *Comparison of demographic attributes of all Twitter users in each survey sample who provided handles with those who provided handles and reported being registered to vote*. Values are percentages; standard errors for survey estimates appear in parentheses.

Online Appendix A10: Addressing sampling concerns

Low-activity accounts. It is possible that we undercount systematically different users as a result of the Decahose-based sampling scheme. That is, if a user has only one tweet they have a lower probability of appearing in the sample of candidate matches than users with 1,000 tweets. We present results below suggesting that while users with five or fewer tweets may be underrepresented, they do not meaningfully differ from users with 10 or 20 tweets.

Note that we can estimate the probability of a missed tweet and a missed account (we appreciate the suggestion from an anonymous reviewer). For example, an account posting only one tweet has a 90% probability (0.9^1) of being missed by this sampling scheme, and an account posting three tweets a 72.9% chance (0.9^3). One can use the inverse of these probabilities to adjust the sample of matches.

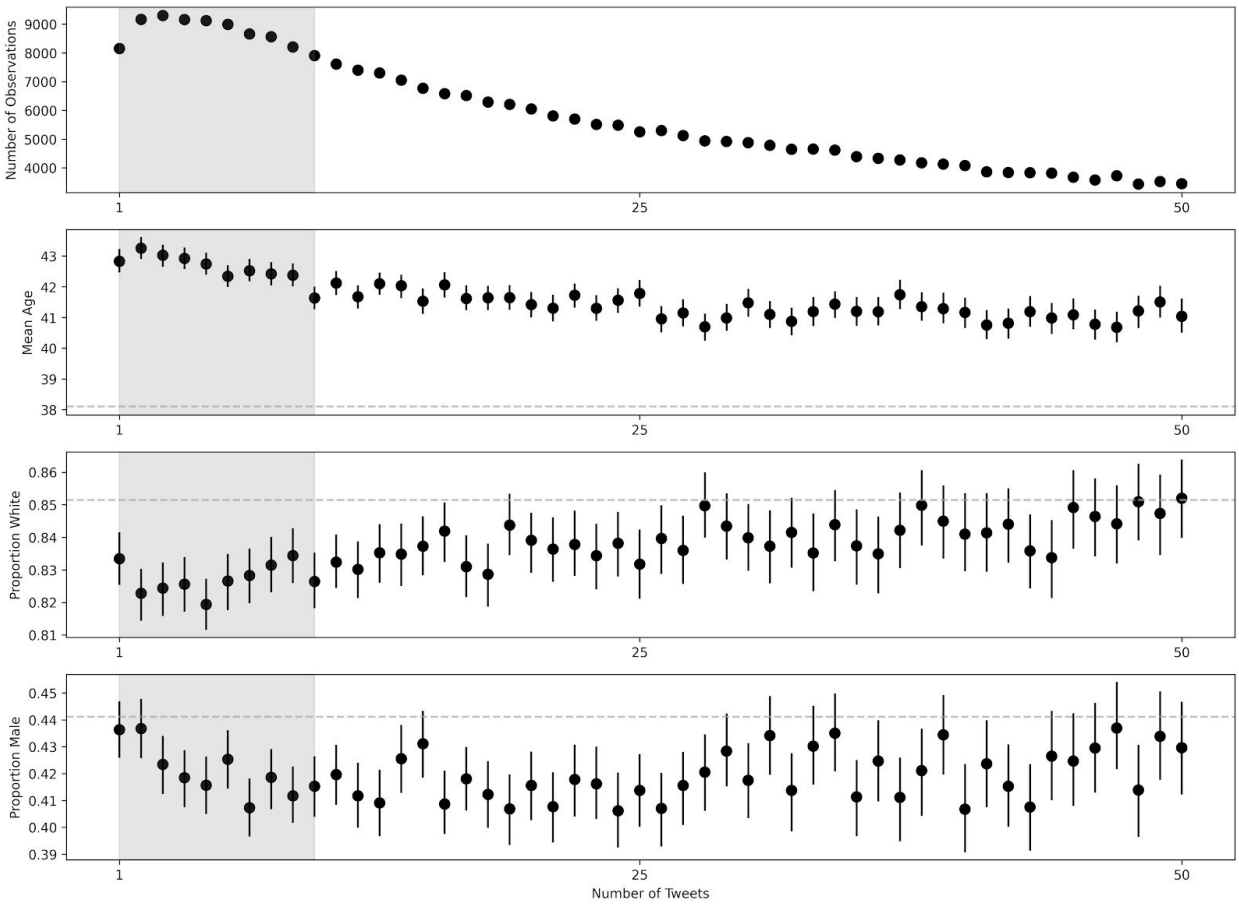

Figure A1: Number of users and select demographic information for low-activity accounts. Because the Decahose is a 10% sample of Twitter, users with fewer than 10 Tweets (shaded region) may have a lower probability of being included. If we assume that the number of tweets in the overall population is described by some form of power-law distribution, then users with fewer than five tweets likely are being underrepresented. However, while low-activity users may differ systematically, especially in age, from the overall population (dashed horizontal lines), there do not appear to be major differences between users with one tweet and nine tweets.

Data-collection periods. For the voter file sample, we collect tweets from each user in approximately six-week intervals. We simulate the consequences of this limitation by examining only accounts in each sample with a maximum of 3,200 tweets from October 1, 2019–November 30, 2019. We also exclude accounts with zero tweets. For the voter file sample, this results in the exclusion of 893,016 accounts (due to a lack of any tweets) and 1,510 accounts (0.1% of matched accounts) due to exceeding the 3,200 tweet limit. For the KnowledgePanel survey, the

removal of accounts with zero tweets decreases the sample size from 2,369 to 1,153 (leaving 49% of the original sample), while the removal of accounts with more than 3,200 tweets results in the exclusion of an additional 16 users (removing 1.4%). For the ATP, dropping those with no tweets brings the total from 1,188 to 806 accounts (leaving 68% of the original sample). In this sample, no users had more than 3,200 tweets.

We can approximate the consequences, in terms of data loss, of using a six-week interval by examining the metadata of tweets gathered in each data collection period. If the metadata reports a change in a users' tweets of 5,000, and we only collected 3,200, we can infer that we have lost 1,800 tweets. Our six-week intervals do not line up perfectly with the two-month timeframe presented in this paper, so our estimates of data loss are likely to be over-estimates. In the worst case, we are missing around 31,000 tweets for the most prolific tweeters and 1.2 million tweets overall. This is a substantial loss but does not affect many of our conclusions; for example our estimate of the Gini coefficient shifts by only 0.02.

Online Appendix A11: Following Comparison

Following political accounts provides another measure of political behavior on Twitter—and one that is insensitive to tweet frequency. As Figure A2 shows, a similar share of Twitter users follow most of the political accounts examined here—though members of the American Trends Panel are substantially more likely to follow Alexandria Ocasio-Cortez. More than twice as many of these users follow Alexandria Ocasio-Cortez compared with the other two data sources, and a slightly larger proportion of users follow Ted Cruz.

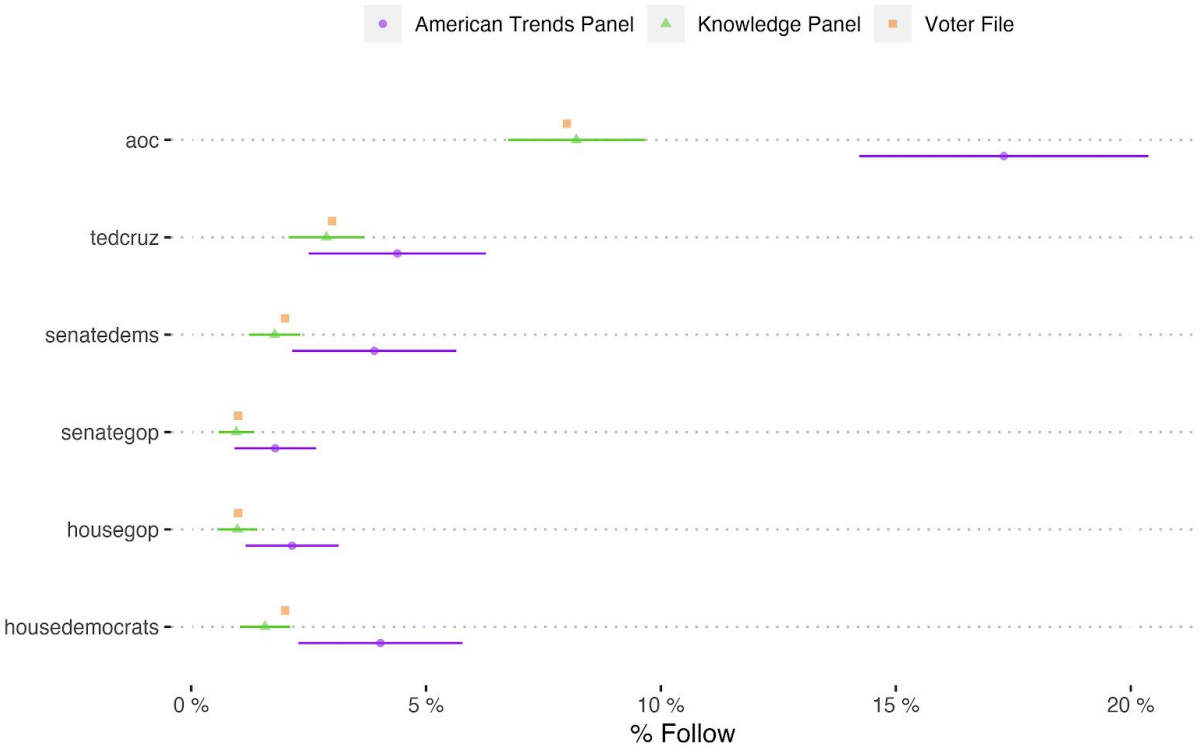

Figure A2: Following behavior comparison of voter file and survey samplespanels.

However, the raw sample size among users following particular accounts of interest is substantially different across data sources. Using the voter file nets more than 25,000 accounts following @senatedems, while in both surveys fewer than 70 users follow the same account. In

concrete terms, the subsample from KnowledgePanel following @senatedems has a within subset “poll”-wide (maximum) margin of error of 22.9 percentage points; for the American Trends Panel, that margin of error is 15.3. Surveys therefore provide very limited opportunities for the analysis of who follows relatively low-salience accounts.