Supplementary Materials for

Fake News on Twitter During the 2016 U.S. Presidential Election

Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, David Lazer*

* Correspondence to: mailto:d.lazer@northeastern.edu

This PDF file includes:

Supplementary Materials

Contents

S.1 Linking Voting Records to Twitter Accounts

While difficulties with measurement on social media are well-established (*4*), these are particularly pernicious for studying fake news, where bots and governments alike engage in manipulative campaigns that include posing as ordinary people (*7, 28, 29*). In order to focus on the experiences of Americans on Twitter, we linked a sample of U.S. voter registration records to Twitter accounts. A similar method has been described by Barberá (*30*), who took locations from geolocated tweets instead of profile fields. The resulting panel contains the profile information and online activity of Twitter accounts associated with real people who live in the United States and are registered to vote. In this subsection, we provide details on the linking procedure. See SM S.8 for details on how we validated the accuracy of the matching process and SM S.2 for details on how we assessed the representativeness of the panel.

In order to perform this linking, we have developed a process that extracts names and locations from the text of Twitter profiles, then matches accounts to their voter registration records. As detailed below, we used two comprehensive datasets: Twitter accounts and voter records. We created a match across datasets if the names and locations matched and neither dataset recorded another person having the same name in that location. Location granularity was at the level of (U.S.) states; that is, we only matched people whose names appeared to be unique in their entire state. About 49% of voters met this criterion.

Twitter profiles and name extraction. We used a 10% sample of Twitter (also known as the Twitter Decahose) to gather a near-complete set of active Twitter accounts: the approximately 290 million users whose tweets appeared there at least once between Jan. 2014 and June 2015. From the profile fields for name and screen name (also known as the Twitter handle), we extracted a set of "name words" for each profile. These included the individual words from the name field, plus certain substrings of the screen name that could be parts of a person's name.

Among the Twitter profiles, only 70% had at least two words (of at least three characters each) in their name field, but this fraction increased to 95% when we combined these with words from the screen name field.

Candidate matches. From a national voter database, we used a sample of close to two million records whose names (first name, last name combination) were unique in their state. To try to match an individual voter, we first searched the Twitter data for accounts with matching names (that is, accounts that contained both the voter's first and last names among their "name words"). If this search returned between one and ten Twitter accounts ("candidate matches"), we then extracted locations from these Twitter profiles and examined the uniqueness of their names and locations. In the rare case that we could extract locations (details below) for all the candidate match Twitter accounts *and* exactly one had the same location as the voter record, then we declared the account to be a match.

Note that only a minority of Twitter profiles—39%—listed anything in their location field, and fewer provided text that yielded a geographic location. Those without locations were treated as "nuisances" among the candidate matches: they could not be matched to a voter record, but they could not be ruled out either. Approximately 32% of the voter records returned 1–10 candidate match Twitter accounts (with matching names). However, only around 4% of these (or 1% of the initial sample of voters) could successfully be matched to a Twitter account.

Location extraction. We extracted locations using the text in the location field of the profile. Previous work has found that self-reported locations are quite consistent with locations obtained through other methods, such as geo-tagged tweets or time zones (*31, 32*). For purposes of efficiency, we only attempted to extract locations from those Twitter accounts that were returned as candidate matches to voters. We mainly relied on a gazetteer approach (*33*), in which we matched the text field against lists of known domestic and foreign cities, states, countries, and

abbreviations. Prior to checking for known place names, the text field was parsed into components if it matched common patterns such as "city, state." For instance, we would extract the location "New York" (state) from profiles listing "New York", "greater NYC area" (common abbreviation), "Buffalo" (a major city), or "Hamilton, NY" (a small town not in the gazetteer, but recognized as a "city, state abbreviation"). On the other hand, "Brighton, UK", "Baja, California", and "Jakarta, Ind." would correctly be recognized as "foreign."

As an additional step to improve the number of locations extracted, we re-checked the Twitter account's location field against the voter record it was considered a candidate match for. If it contained the voter's city and/or state, we updated the inferred location. This second step enabled matches in cases where the location field alone was unrecognized or ambiguous. For example, if a voter registered in Hamilton, NY had a candidate match Twitter account with a location field of "Hamilton" or "originally from Arizona, now in upstate NY", the account's location would be updated to "New York" (state).

Among the Twitter accounts returned as candidate matches, we extracted a U.S. state for about 19%. About 12% of the location fields were recognized as foreign, and the remainder were either blank (57%) or contained no recognized location (12%).

Voter data and account filtering. As the outcome of the above process, we found matches for approximately 1% of the initial sample of voter registration data. The voter data we used for the paper was provided by TargetSmart, one of the leading companies in compiling and providing up-to-date U.S. voter records. The voter records included a variety of information, including each individual's name, address, age, gender, and inferred race. In this study, we used a 16,442-member panel: those individuals for whom the voter records provided age, gender and race[1], whose Twitter accounts were not protected, had not been compromised, followed at least

[1]For voter records with missing demographic data, we attempted to infer gender and race using name-based approaches with the `genderdata` and `wru` (*34*) packages in R, respectively. We removed accounts for which we

one other account, had sent at least one tweet and were exposed to at least one political URL during the study period (see below).

In order to further guarantee that the sample did not include bots, we identified accounts judged by BorOrNot as bot-like (*35*). For this determination, we used a threshold of scores $\geq .7$ for accounts having ≥ 50 tweets. This threshold flagged 141 accounts ($< 1\%$). We then manually investigated a subset of accounts and retained in the panel 15 we verified as controlled by the person to whom we had linked the account (see Section S.8 below). Finally, based on manual inspection of the most prolific accounts, we removed two profiles that appeared to be hijacked.

This left the 16,442 matches used in this study. In term of representativeness, the panel represents a 3.7% sample of the voters we expect to have Twitter accounts (based on two million voters we attempted to match and a recent estimate that 27% of Americans are on Twitter (*36*)). We used the Twitter API to collect tweets sent by these accounts during the 2016 election season (Aug. 1–Dec. 6, 2016) and to obtain their followers and their followees (accounts they followed) as of December 2016.

S.2 Panel Representativeness

The panel construction process identifies matches by leveraging name uniqueness within a state. The resulting sample therefore over-represents people with rare names and locations, compared to all voters or Twitter users. Further, additional biases may have been introduced by considering only those people who provided name and location in their Twitter profiles, and due to errors, duplicates, and other limitations of coverage of both the voter records and the Twitter data.

could not infer gender and race.

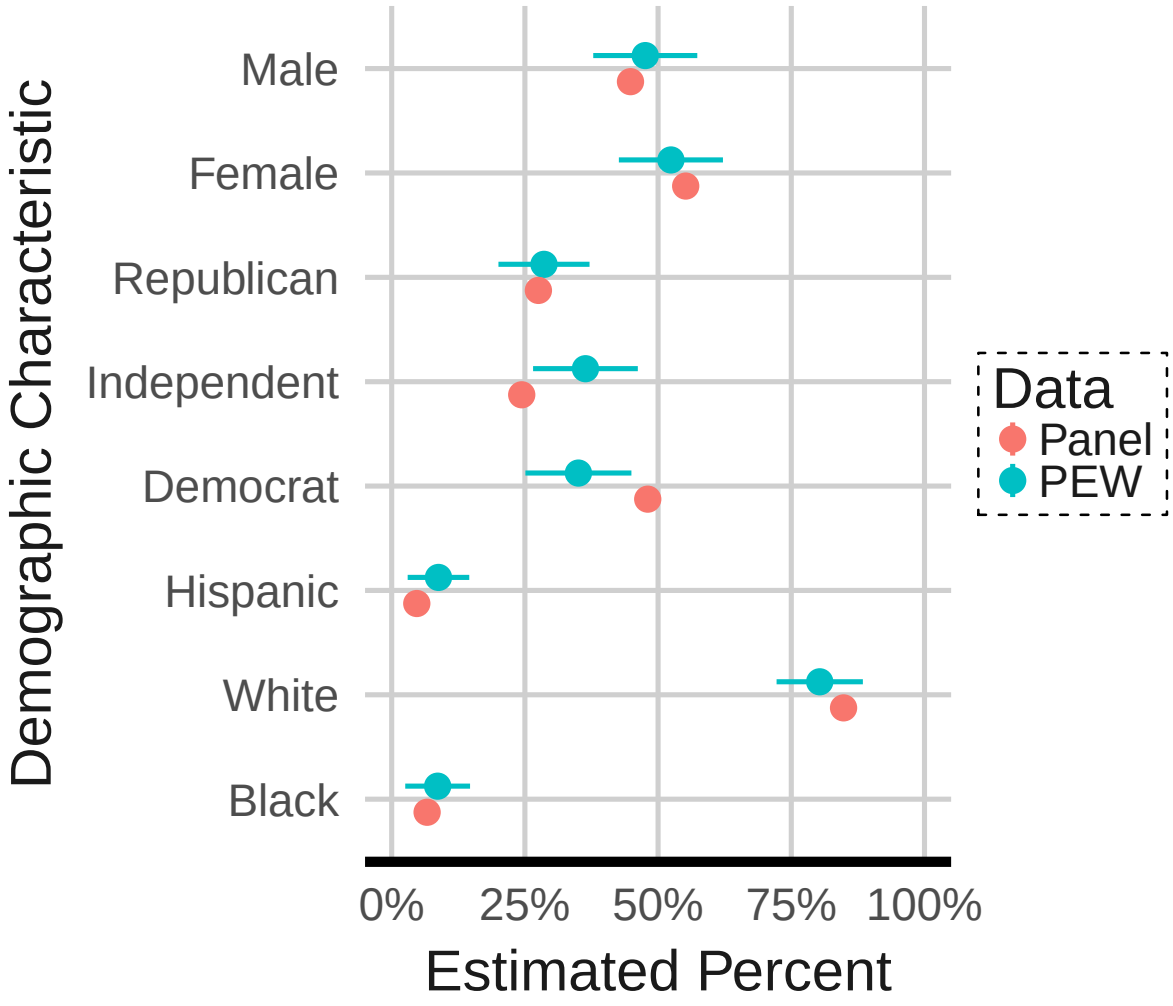

Figure S1: Demographics of panel compared to Pew survey of voters on Twitter. Confidence intervals for weighted survey data are calculated using the `survey` package in R (*37*); confidence intervals for our Twitter panel are computed using the approximate method in (*38*).

U.S. Voters on Twitter. To understand any population bias created by the panel construction method, we compare our sample to a representative sample of registered U.S. voters that are Twitter users. This sample was derived from a 2016 Pew Research survey of U.S. adults, which was weighted (by Pew) to be nationally representative (*14*). Of respondents who used the Internet, 26% (95% confidence interval of [23, 30%]) said they were on Twitter. Of those on Twitter, 71% [62.0%, 78.0%] said they were registered to vote. We compare the panel to the weighted set of 133 survey respondents who reported being in both these categories.

Figure S1 shows that our panel has little demographic bias compared to the survey respondents. Compared to survey respondents who listed their political affinity, our sample contains slightly more Democrats. As estimated from the survey data, somewhere between 25% and 45% of registered voters on Twitter are registered Democrats; our data contains 48% registered Democrats. As Figure S1 shows, this difference is due to a difference in the number of registered voters not aligned with a party, not an undersampling of Republicans. Additionally, although not shown in Figure S1, the average age of the panel (38.7 [38.5, 38.9]) is consistent with the average age of survey respondents (40.8 [37.6, 44]).

Prior work has shown that controlling for demographic variables eliminates most differences in political attention, values or behavior between users and non-users of social media (*39*). Based on these findings and the overall demographic resemblance of the panel to a Pew survey, we believe that the panel is reflective of the population of registered U.S. voters on Twitter.

Random Twitter accounts. We also examined how the Twitter accounts of panel members compare to the greater population of accounts on Twitter. We sampled a random set of accounts on Twitter from those that had been observed in Twitter's Decahose between Jan. 2014 and Aug. 2016. We collected basic account statistics in March 2017 from the public profile data of accounts still open at that time. Figure S2 shows how these account statistics differ between the

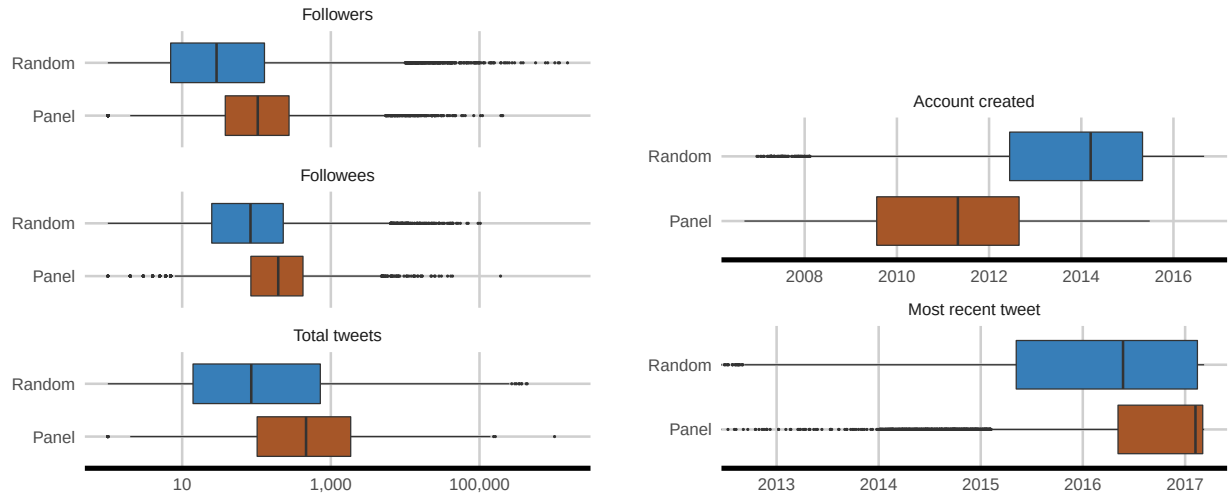

Figure S2: Account statistics of panel compared to a random sample of Twitter accounts, according to profile data from March 2017. Dates of most recent tweets are only visible for non-protected accounts.

panel used in the study and the random sample.

Overall, accounts in the panel were older, more active in terms of tweets and network size, and more recently active than the rest of Twitter. In addition, although exceedingly few accounts were "verified," panel accounts were around seven times as likely (0.5% [0.4%, 0.6%]) as random accounts (0.07% [0.06%, 0.09%]) to have this designation. Having earlier creation dates, as the panel accounts do, has often been shown to correlate with accounts being run by humans (*15, 40*), although in this case it might simply be due to the time period in which profile data was collected for matching. The other statistics show that the panel accounts are more active and more recently active than random accounts, which too is to be expected as a result of the selection process for this study (as described at the end of Section S.1 above).

S.3 Sharing of and Exposure to Political Tweets

As noted in the main text, we collected the panel's tweets, followers, and followees. We also retroactively estimated the content to which panel members were exposed. Examining exposure

to fake news sources is important because people on social media often scroll through lengthy feeds without any observable interaction with the content in the feed. This content may influence people's beliefs even in cases where individuals do not recall it or click through to view the full story (*41*). To study exposure on Twitter, we collected all tweets sent by followees of panel members within a historical dataset that contains 10% of all tweets (the Twitter Decahose). This allowed us to estimate the composition of the news feed of every panel member. We considered a potential exposure for a panel member to be any tweet shared by one of their followees. For example, a single tweet observed from an account followed by five panel members would be counted as five potential exposures, one for each panel member. The values we report for potential exposures (also shortened to "exposures") have been multiplied by ten to extrapolate to the complete data. Of course, panel members would only have seen a fraction of the tweets posted by accounts they follow. However, prior work suggests that for the vast majority of content shared on Twitter, 5% or more of potential exposures are actually seen (*16*). While this fraction is useful as a rough estimate for the amount of content seen, it should be interpreted with caution since the fraction is likely to vary considerably from one person to another based on actual use of the Twitter platform. We restricted our analysis to political tweets that contained a URL to a web page outside of Twitter. We identified political tweets using a logistic regression classifier which was trained to distinguish tweets that match political keywords from other tweets; we describe this method in detail immediately below. We validated the classifier with human judgments, and found that classifier predictions overwhelmingly agreed with human annotators on whether the tweet was politically relevant (see SM S.3 for details). For the URLs, we followed all redirects and took the URL of the final landing page. In total, we studied 89,875 shares of political URLs by panel members and 9.8 million political URLs that were posted by followees of the panel, amounting to over 640 million potential exposures to political URLs.

To identify political tweets, we have developed a classifier that works on a tweet by tweet

basis. Its basic architecture is as follows: first, we extract tweets that match a political keyword list, to use as "positives" for training. This is a high-precision set; the keywords have been selected to create minimal false positives. From the remaining tweets, we randomly sample an equal number to use as "negatives" for training. Finally, we apply the classifier to all tweets not already matched by keywords, to improve the recall. Such a method of training a classifier on imperfectly separated data is referred to as distant supervision and is widely used with social media data, where ground truth labels would be expensive to obtain at large scale (*42, 43*). Since word usage on Twitter is as dynamic as the news cycle, we train and apply a new classifier for each day of the study period.

Prior to any content-based filtering, we preprocess the tweets to expose as much content as possible. From the tweet's raw JSON representation, we concatenate several fields: the non-truncated text of the tweet, the quoted tweet, if applicable, any Twitter handles being retweeted or quoted, and the final landing pages of all URLs (after following redirects).

The political keyword list we developed, shown in Figure S3, was refined and updated from that in (*44*). It contains 111 words, phrases and regular expressions seen in the text, hashtags or Twitter handles of political tweets. While some terms are generic to elections, parties and officials, the majority are names, Twitter handles, or hashtags associated with candidates from either the primaries or the general election. Frequently occurring terms, mainly regarding Clinton and Trump, were manually vetted to keep false positive rates low.

Within a given tweet, we search for political keywords within words and URLs (respecting word boundaries), within hashtags (anywhere), and as exact Twitter handles. When a political keyword is found, we remove the entire word (or URL) containing it so that the classifier cannot obtain any signal from the keywords themselves. Tweets are tokenized into words (with punctuation removed), then stemmed. Additional tokens are created from URLs (by preserving the complete URL) and from hashtags (by splitting them into substrings based on capitalization).

General terms: election, debate, presdebate, VPdebate, liberal, conservative, republican, democrat, democratic, GOP, DNC, politics, political, president, voter, governor, congress, congressional, representatives, senate, senator, rep\., sen\.

Office holders: biden, harry reid, mitch mcconnell, boehner, paul ryan, pelosi, kevin mccarthy

Candidates: hillary clinton, HillaryClinton, VoteHillary2016, hillary2016, imwithher, clinton, hillary, tim kaine, timkaine, kaine, mike pence, GovPenceIN, mike_pence, pence, donald trump, realDonaldTrump, donaldtrump, donaldjtrump, trump2016, trump, the donald, MakeAmericaGreatAgain, maga\b, imwithhim, trumppence16, crookedhillary, lyinghillary, nevertrump, neverhillary,

bernie sanders, BernieSanders, SenSanders, FeelTheBern, Bernie2016, Sanders2016, jill stein, DrJillStein, jill2016, Ajamu Baraka, AjamuBaraka, gary johnson, GovGaryJohnson, johnsonweld, william weld, bill weld, GovBillWeld, evan mcmullin, evan_mcmullin, mindy finn, mindyfinn, jeb bush, JebBushforPres, ben carson, RealBenCarson, bencarson, chris christie, GovChristie, ChrisChristie, ted cruz, tedcruz, CruzCrew, carly fiorina, CarlyFiorina, jim gilmore, gov_gilmore, lindsey graham, LindseyGrahamSC, GrahamBlog, mike huckabee, GovMikeHuckabee, huckabee, john kasich, JohnKasich, kasich, marco rubio, marcorubio, SenRubioPress, TeamMarco, rick santorum, RickSantorum, TeamSantorum

Figure S3: Keywords used to identify tweets as political, for training the classifier. Matching is case insensitive.

The training set for the classifier consists of all tweets per day (up to 100,000) that contain a political keyword, and an equal number of tweets that do not. Each tweet is represented as a vector of word counts. Words occurring in fewer than 0.02% or more than 90% of tweets are ignored; this pruning of the feature set helps the model converge. We use a LASSO logistic regression model, with the tuning parameter λ chosen through cross-validation (*45*). The resulting classifier is applied to all tweets that did not contain political keywords. The political filter returns all tweets with classifier scores above a given threshold, plus those matching the political keyword list.

During development, we compared pairs of options by examining their respective performance on the training set, the separation of scores on the unlabeled test set, the number of tweets labeled positive, and words assigned the largest classifier coefficients. We also manually

annotated tweets on which the classifiers disagreed (up to 50 positives from each classifier). From these informal evaluations, we learned that the method is particularly sensitive to the initial keyword list, that plain word counts work better than tf-idf weights, and that a training set with 50% positives works well on this corpus in conjunction with a threshold of 0.8 on the classifier's predictions. Once finalizing the classifier with these parameterizations, we collected manual labels for 20,000 tweets in our sample in order to validate the classifier (see below for evaluation results).

Overall, the classifier judged 10–20% of tweets with URLs to be political during most days in the study period. The median day had 15% of tweets shared with URLs classified as political (IQR: 12–20%), and 11% of potential exposures with URLs classified as political (IQR: 10–14%). The fraction of political tweets peaked during the presidential debates and Nov. 8 and 9, reaching up to 41% of shares and 26% of exposures. The majority of these "positive" labels were triggered by the political keyword list, but in general 2% to 3% of all exposures or shares, respectively, were labeled as political due to the logistic regression model itself.

S.4 Evaluation of Political Classifier

In order to assess the accuracy of our political classifier and potential biases it might have in estimating levels of content from fake news sources, we randomly selected 20,000 tweets, stratified over the days included in the study, to hand-label. All tweets we selected for hand-labeling were annotated by at least two workers on Amazon's Mechanical Turk. Where Turkers could not agree ($N = 1,001$), authors agreed upon a final label.

Annotators were shown each tweet with unshortened and clickable URLs. Where we identified a tweet to be non-English,[2] we provided a translation from Google Translate. Annotators were asked to label each tweet as being either relevant to the 2016 U.S. election (8.4%,

[2]Using the `langid` package in python

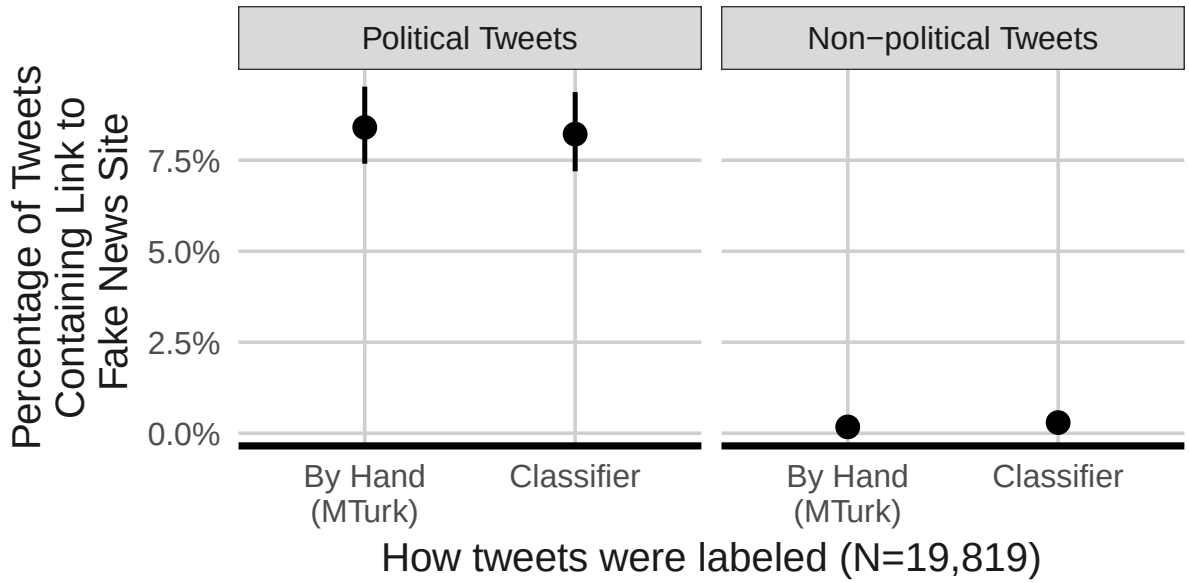

Figure S4: The percentage of content from fake news sources (y-axis) in annotated tweets for political (left panel) and non-political (right panel) tweets. The x-axis represents how a tweet was labeled, either using our classifier or by hand-annotation. Confidence intervals are 95% binomial confidence intervals computed using the method described in (*38*).

$N = 1,679$), relevant to U.S. politics writ large (4.9%, $N = 975$), about something else entirely (85.8%, $N = 17,165$), or "I don't know" (.9%, $N = 181$). Most "I don't know" answers were either tweets that were just posts of now-dead links or tweets that our language classifier misclassified as English. These tweets were removed from further analysis.

When combining the "U.S. election" and "U.S. politics" answers into a single "politics" class (vs. "something else"), the classifier had a precision of 83.1% [81.5,84.5] and recall of 76.9% [75.3, 78.6] for an F1 of 79.9 [78.7, 81.1]. Restricting the positive class to only "U.S. election" tweets, recall is boosted to over 95%; thus, we expect the classifier is able to capture nearly all tweets related to the 2016 election, and that tweets it captured not directly related to the election were still highly likely to be related U.S. politics in some other fashion.

While the classifier performs well overall, it is possible that it was still biased. It may have identified content from fake news sources as political at higher (or lower) rates than other types

of content, in which case estimates of the amount of content from fake news sources in the data might be inflated (or minimized). Figure S4 shows that such a bias is unlikely to have occurred. The figure presents a comparison of the percentage of content from fake news sources in the annotated sample for tweets labeled as political or not by the classifier, relative to those same tweets as labeled by hand. We find that 8.2% of the tweets labeled by the classifier as being political contained content from fake news sources, compared to 8.4% of tweets labeled as political by Turkers.[3] Similarly, 0.29% of the tweets classified as non-political by the classifier were fake, compared to 0.17% of tweets labeled non-political by hand. The latter difference is statistically significant, but in neither case is there a practical difference between estimates of content from fake news sources provided by the classifier versus by hand-labeled data.

S.5 Defining Fake News

Defining fake news is an extremely difficult task, not least because there is currently no consensus of what exactly fake news entails (*46*). As fake news remains a problematic binary label, we used a multiclass characterization of domains that were likely to share political misinformation during the campaign period due to poor journalistic practices. These are domains with all the trappings of legitimately produced news but without the underlying organizational processes. We included pre-existing lists of fake news domains (black domains) and constructed additional lists (red and orange domains).

Following the 2016 election, several organizations and individuals published lists of fake news websites and URLs. We used pre-existing domain lists both from trusted journalistic outlets and from prior academic work. First, we took 163 sources from Buzzfeed News's continuing series on fake news (*47–51*). We added Politifact's list of 200 fake news websites from May 2017 (*52*) and FactCheck.org's list of 56 fake news websites from July 2017 (*53*). There

[3]Note that this number is different from the 5.0% discussed in the text, because the text considers percentage of exposures, whereas we here consider only a sample of tweets and do not count exposure levels.

were two exceptions where pre-existing lists were not labelled as black. First, we mapped domains from Politifact's category of "parody" to our satire label, which were therefore excluded from analyses. Second, we mapped domains from Politifact's category of "some fake news" to our orange label, to reflect their diminished likelihood to spread fake news.

We combined these sources with a source list constructed concurrently by Guess et al. (*9*). Guess et al. (*9*) derived their list in part from Allcott and Gentzkow (*8*), who in turn relied on sources listed by both Snopes.com and Buzzfeed. Of the 92 domains on the list constructed by Guess et al. (*9*), 19 were in the list we had compiled from Buzzfeed, FactCheck and Politifact. An additional 22 of their domains were in the list we had derived from Snopes.com and hand-coded, described below. Of those, we had labeled 15 sites as red, 6 as orange, and 1 as green (not fake news). We classified the remaining 51 sites identified by Guess et al. (*9*) as black. The complete list of 382 black fake news sources we used—of which only 171 appeared anywhere in our data—is available at DOI *10.5281/zenodo.2483311*.

Inspired by work on source credibility (*54*) and in order to measure fake news more comprehensively, we expanded these existing lists with the help of articles from the fact-checking website Snopes.com. First, we scraped all articles on Snopes.com in May 2017 that were tagged under one or more of the categories "Political News," "Politics," "Fact Check," or "Fake News" and that were written about claims that were labeled as "[Mostly] false," "Incorrect," "Inaccurate," or "Unproven". From here, we then extracted the 9,202 URLs linking to external websites from the text of the Snopes.com articles. Intuitively, one might expect that websites often mentioned in Snopes.com articles with false claims, relative to those in articles with true claims, would be potential fake news sources. However, Snopes.com uses URLs both to show examples of false claims and to provide evidence that claims are false. Consequently, outlets like the Washington Post and even sites like Wikipedia routinely appeared on pages where claims were being proven false.

However, what Snopes.com typically does do is *archive*, using the archiving service archive.is, URLs that link to pages with examples of false claims. Evidence of this can be found in the fact that on Snopes.com web pages where the assessed claim was found to be "false", "incorrect", "inaccurate" or "unproven", there were a total of 1,051 archived URLs, compared to only 22 on pages where claims were "verified." We therefore extracted all 1,051 links to archive.is URLs, and then extracted the domain of the archived web page. This process led to a list of over 500 domains that were mentioned in Snopes.com articles about political news and were found containing false claims. We then further limited this set of domains to those that at least 1% of Democrats, Republicans, Independents, or other registered voters were potentially exposed to (as determined by our data drawn from the 10% sample) in their timeline during the three months leading up to the election. This filtering process resulted in 171 domains that we manually labeled.

For our manual labeling, we aimed to develop an objective scale for rating the likelihood of domains to spread inaccurate information. Four independent annotators agreed upon factors that were potentially indicative of a site's propensity to elicit fake news and misinformation. Annotators considered over 10 different dimensions of each site (e.g., author attribution, masthead, offering of corrections) as well as the severity and frequency of false claims documented on Snopes. When assessing the frequency and severity of fake news content, annotators reviewed the original Snopes.com articles, the website itself, and the website's archive. The red and orange labels reflect annotators' confidence in attributing the falsehoods found by Snopes to a flawed editorial process at the source. Two independent annotators evaluated each website and assigned the sites into one of six levels: green (reasonable and accountable journalism), yellow (low quality journalism), orange (negligent or deceptive), red (little regard for the truth), satire (self-described as satirical and affirmed as such by the annotators), and sites not applicable (for example, Amazon). The two annotators agreed on the labeling of 60% the sites when

using the six categories, and 83% of the time when collapsing the six categories into fake and non-fake categories. This suggests that annotators can separate fake from non-fake sites with high agreement, and that the more nuanced definitions are harder to distinguish.

To review the website's editorial process, annotators noted whether individual authors were attributed to each article, whether journalists had readily available information about journalism degrees, whether there was a masthead or listing of author biographies, and whether the site elicited corrections. Annotators also noted whether the content was sensationalist to the extent that the article becomes misleading or whether the coverage was particularly partisan. Other factors that were reviewed were whether the website was owned by a third party, whether the website had changed domain names, and whether there was a vested interest of hosting the site (e.g. featuring a store and/or promoting commercial enterprise).

After this initial round of coding was completed, annotators reconvened and discussed all coding conflicts. If annotators could not agree on a resolution, a third annotator was called to provide further judgement. Once annotators agreed upon the website's rating, a justification was noted which was converted into coded comments. Each website received multiple codes, with the worst transgression corresponding to the overall rating. For example, if a website was found having a "vested interest" and "mild and rare inaccuracies", it had a yellow rating. However, if the site had a "vested interest" and "major and frequent falsehoods", it was classified as red. Our coding scheme and source labels are publicly available at DOI *10.5281/zenodo.2483311*. For ease of access, Tables S1 and S2 below contain the ultimate rating of all red and orange sites included in analyses, respectively. The seven most popular sites by exposure are The Daily Caller (orange), The Gateway Pundit (red), Truthfeed (red), InfoWars (red), The Real Strategy (red), Zero Hedge (orange) and the Conservative Tribune (red).

Finally, as described in Section S.6, we also annotated an additional set of 827 websites accounting for 80% of all exposures to URLs from both fake and non-fake sites in our data. When

performing this annotation, we flagged seven websites - nutra-lifestyle.com, ecowatch.com, thecount.com, tacticalinvestor.com, bossip.com, therealstrategy.com and rawstory.com - as potentially fake news sources. We then carried out the annotation task described in this section with these websites. Of the seven flagged websites, we eventually determined that two of these sites - nutra-lifestyle.com and therealstrategy.com - fit our definition of red fake news sources and were added to our lists.

In total, then, we examined 171 websites identified by Snopes.com as producing articles with a false or unproven claim, plus seven additional websites identified during manual inspection of the top websites in our dataset. We also included in the orange category 18 domains labeled by Politifact as having "some fake news." Of the 171 websites identified by Snopes.com, annotators labeled 64 sites as red and 65 as orange. They characterized the remaining 42 domains as reputable news publishers, sites containing only mild inaccuracies, clearly satirical, or sites other than news; none of these were considered fake news sources.

S.6 Categorizing Non-fake News Websites

In order to get a better understanding of the range of political content voters were exposed to, we manually categorized the top 827 websites that accounted for 80% of all political URL exposures and a random sample of 200 websites from the remaining 20% of the exposure distribution. The sample of top sites was stratified by party, and constructed by taking the union of the top 475 websites that Democrats, Republicans, Independent, or other registered voters were exposed to, respectively.

Annotators labeled each website as one of 13 categories that most accurately describe the site. These categories included distinctions between political and non-political news sites, different entities behind a site (e.g. politician, commercial organization, governmental organization), and platforms of user-generated content. Two annotators coded each site, and a third

100percentfedup.com	louderwithcrowder.com
activistpost.com	myfreshnews.com
allenbwest.com	naturalnews.com
allenwestrepublic.com	newsrescue.com
americannews.com	nowtheendbegins.com
americantoday.news	nutra-lifestyle.com
americasfreedomfighters.com	observatorial.com
anonews.co	powderedwigsociety.com
anonhq.com	proudcons.com
barenakedislam.com	religiousmind.com
bipartisanreport.com	rightsidenews.com
channel-7-news.com	shariaunveiled.wordpress.com
collective-evolution.com	sourceplanet.net
conservativebyte.com	stateofthenation2012.com
conservativefiringline.com	superstation95.com
conservativeoutfitters.com	tacticalinvestor.com
conservativepost.com	theeventchronicle.com
dailystormer.com	thefreepatriot.org
dcclothesline.com	thegatewaypundit.com
downtrend.com	thelastamericanvagabond.com
eaglerising.com	thenewsclub.info
endtimeheadlines.org	therealstrategy.com
eutimes.net	trunews.com
fellowshipoftheminds.com	truthfeed.com
frontpagemag.com	truthuncensored.net
fury.news	usasupreme.com
getoffthebs.com	viralliberty.com
gopthedailydose.com	wearechange.org
gotnews.com	westernsentinel.com
infowars.com	whatdoesitmean.com
jookos.com	wnd.com
judicialwatch.org	worldtruth.tv

Table S1: Red fake news sites.

2ndvote.com	medicalkidnap.com
afa.net	mentor2day.com
ahtribune.com	nativeamericans.us
awarenessact.com	newcenturytimes.com
blackinsurancenews.com	onlysimchas.com
cannasos.com	palmerreport.com
chicksontheright.com	pamelageller.com
coed.com	qpolitical.com
concealednation.org	redflagnews.com
conservativetribune.com	regated.com
crooksandliars.com	rightwingnews.com
dailycaller.com	smag31.com
dailyheadlines.net	teaparty.org
dailynewsbin.com	thatviralfeed.net
dailypost.ng	thebigriddle.com
dailywire.com	theconservativetreehouse.com
davidwolfe.com	thefederalistpapers.org
defund.com	thehornnews.com
dennismichaellynch.com	themindunleashed.com
endoftheamericandream.com	thenationalmarijuananews.com
express.co.uk	thenationalpatriot.com
firstpost.com	thepoliticalinsider.com
healthnutnews.com	tmn.today
healthycareandbeauty.com	toprightnews.com
heatst.com	tribunist.com
ihavethetruth.com	trueactivist.com
ilovenativeamericans.us	urbanimagemagazine.com
impulsetoday.com	uschronicle.com
inquisitr.com	usuncut.com
iotwreport.com	welovenative.com
jewsnews.co.il	youngcons.com
joeforamerica.com	zerohedge.com
learnprogress.org	

Table S2: Orange fake news sites.

annotator was called to resolve conflicts. As noted above, in the process of the coding, annotators identified 7 websites that they suspected as fake news sources. We then assigned these websites, using the procedure described above, to either the red or orange fake news categories, or determined upon further assessment that the sites were not in fact fake news.

Despite the large number of categories, the two annotators agreed with each other on 88% of the labeled sites. After reviewing the annotations, we collapsed the thirteen categories into six larger ones: political news site or blog, non-political news site or blog, social media or user generated content, organization or government or politician's website, fact-checking website, or other. Figure S5 displays the percentage of all political exposures and shares that we categorize as fake news or satire (based on our fake news classification scheme) or into one of the six categories we identified for non-fake content. In order to estimate these quantities for exposures, we first compute the total amount of exposures for sites we hand-labeled in the top 80% of the exposure distribution. We then weight the remaining 20% of exposures for each category by its prevalence in the websites we sampled from the tail. A similar calculation is carried out for shares as well. Note, however, that because our fake news sampling strategy did not follow this approach, we do not carry out this estimate for fake news shares/exposures. The provided quantity is thus a slight under-estimate of the total given in the paper, where we simply compute a percentage of fake news out of all possible exposures, regardless of categorization.

S.7 Patterns in Overall Daily Trends

In this section we provide additional details of patterns in sharing and exposures aggregated across the entire panel. Figure S6 presents the daily percentage of shares of URLs from red, orange and black sources during the time period of interest. At its peak, content from fake news sources reached over 8% of exposures and over 18% of shares. Given that daily counts of shares from the panel are small (relative to exposures), the resulting values are noisier but

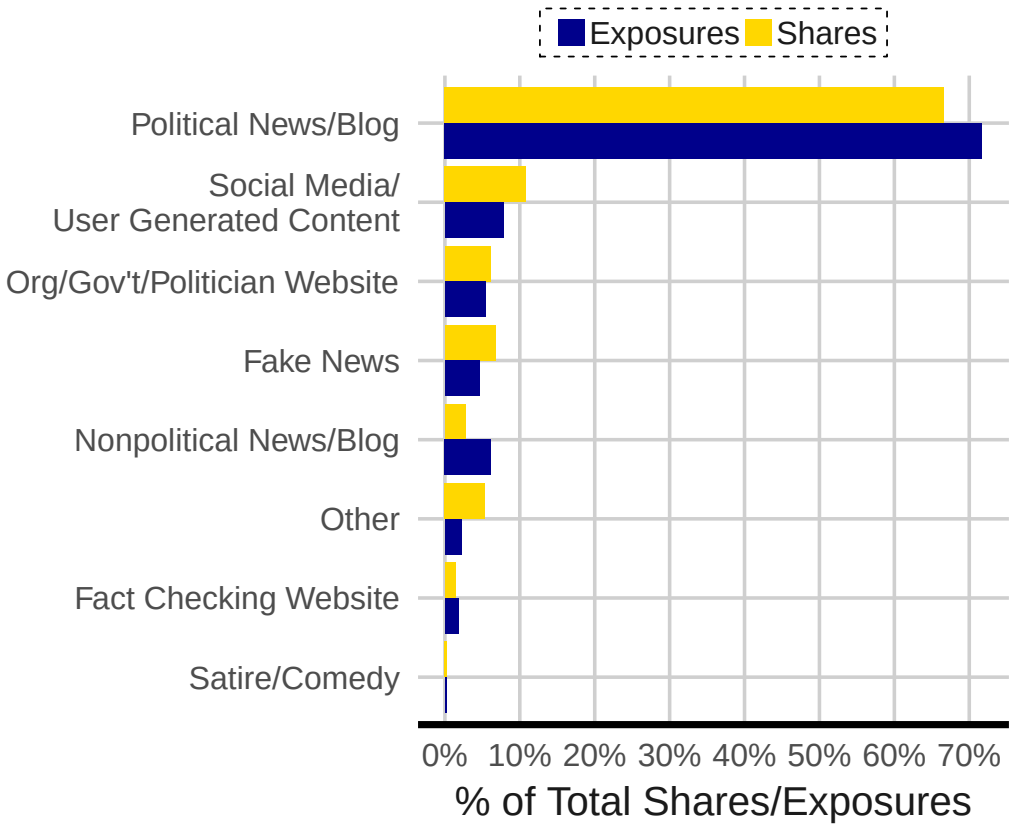

Figure S5: Estimates of the percentage of all exposures to (blue) and shares of (yellow) political links to different types of content.

exhibit similar qualitative trends to exposures. Figure S7 displays the same information as in Figure 1A in the main text, but using the sites on the Guess et al. list (*9*). Similar daily patterns are observed using their list, and the volume of content from fake news sources estimated by their list is somewhere in between the volume we estimate for black sites and orange sites.

We use two methods to validate that the percentage of both shares and exposures significantly increased in the weeks just before the election, and that similar patterns exist when using the Guess et al. list. Results are shown on the left half of Figure S8 provide the estimated trend in percentages of fake new exposures (top), shares (middle), and exposures with the Guess et al. list (bottom) using the Bayesian time-series model proposed in (*55*). Results shown that while

patterns for shares and exposures differ early in election season, both significantly increase in the period leading up to the election. We estimate the model with default priors and trend results are shown controlling for weekly patterns. Additionally, we estimate the model only from July 31st through election day, as the model could not be made to fit the rapid and sudden decrease in percentage of content from fake news sources shared in the days after the election.

On the right half of Figure S8, we display results from an analysis of changepoints in the exposure (top) and sharing (bottom) time series. We use the PELT algorithm, provided in the R package `changepoint` (*56*), and manually tune the penalty term (.001 for exposures, .01 for shares, .0005 for the Guess et al. list) until a changepoint is detected in the period before election day. We then assess whether or not the changepoint a) is near the election and b) displays a significant increase in percentage of exposures (shares). The figure provided shows this is indeed the case; changepoints are detected on October 10th and November 5th for exposures; for October 12th and November 5th for shares and on October 12th and November 6th for exposures using the Guess et al. list. In all cases, a significant increase in the mean value of the time series for the period between these two changepoints, relative to means on the two ends of the time series. We thus find additional evidence of a significant uptick in the percentage of exposures to and shares by the panel of URLs from fake news sources in the weeks leading up to the election.

S.8 Evaluation of Panel Matches

Validation of panel matching

The panel matching process was designed to be high precision—that is, our focus was on ensuring that the resulting voter-Twitter matches had a high probability of linking the same individual. In order to evaluate the precision of matches, we compared the Twitter profiles of a sample of panel members to their linked voter data. While the tests below cannot definitively determine

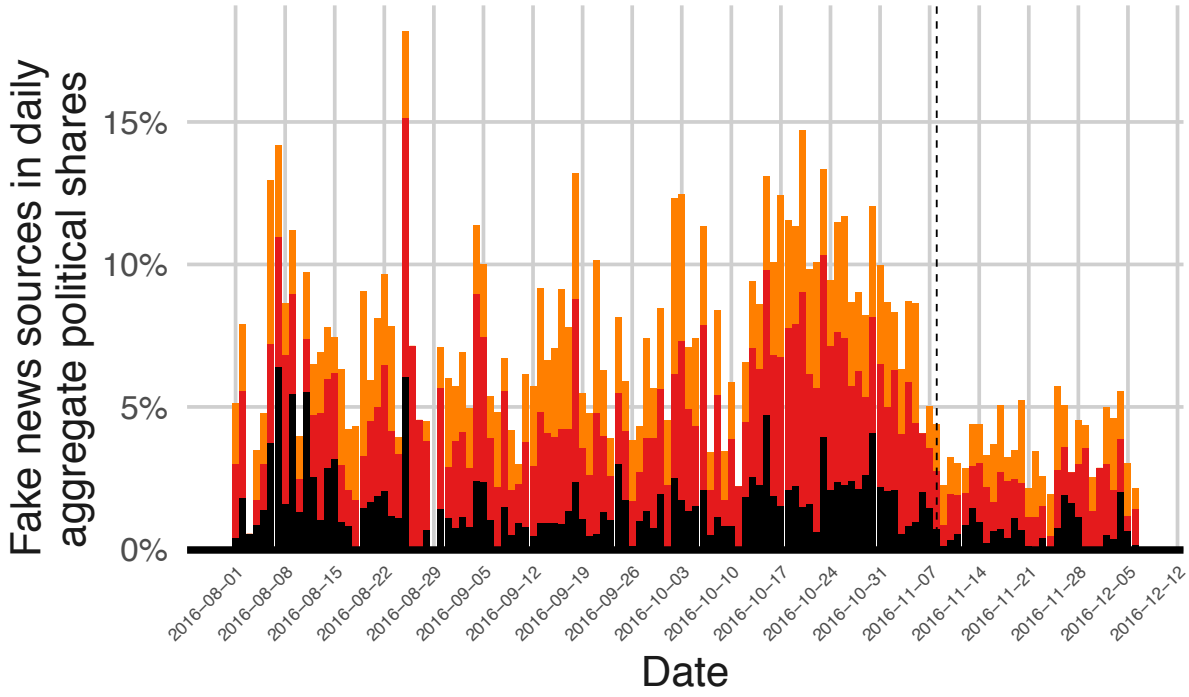

Figure S6: Among the panel's shares of political URLs, percentage each day from black, red and orange fake news sources.

whether matches are correct, they show that in general, there is a high correlation between the attributes of the Twitter profiles and their linked voter records. Specifically, we find that the perceived characteristics of the account holders agreed well with the matched voter records on gender (97%), race (94%) and current age (78% within 10 years, 54% within 5 years).

We used the Twitter API to obtain profile information and Twitter profile photographs for 500 accounts as of September 2018. As a first check, we asked whether the profiles' names and locations still matched the voter records. Among the 489 accounts whose data was publicly available, 95% displayed the voter's first and/or last name in the name field, and 84% contained the voter's city or state in the location field. Although these fields were originally used to construct the matches, they may have been updated in the time since the profile data used for matching was first recorded, several years earlier (2014 and 2015). In addition, the initial

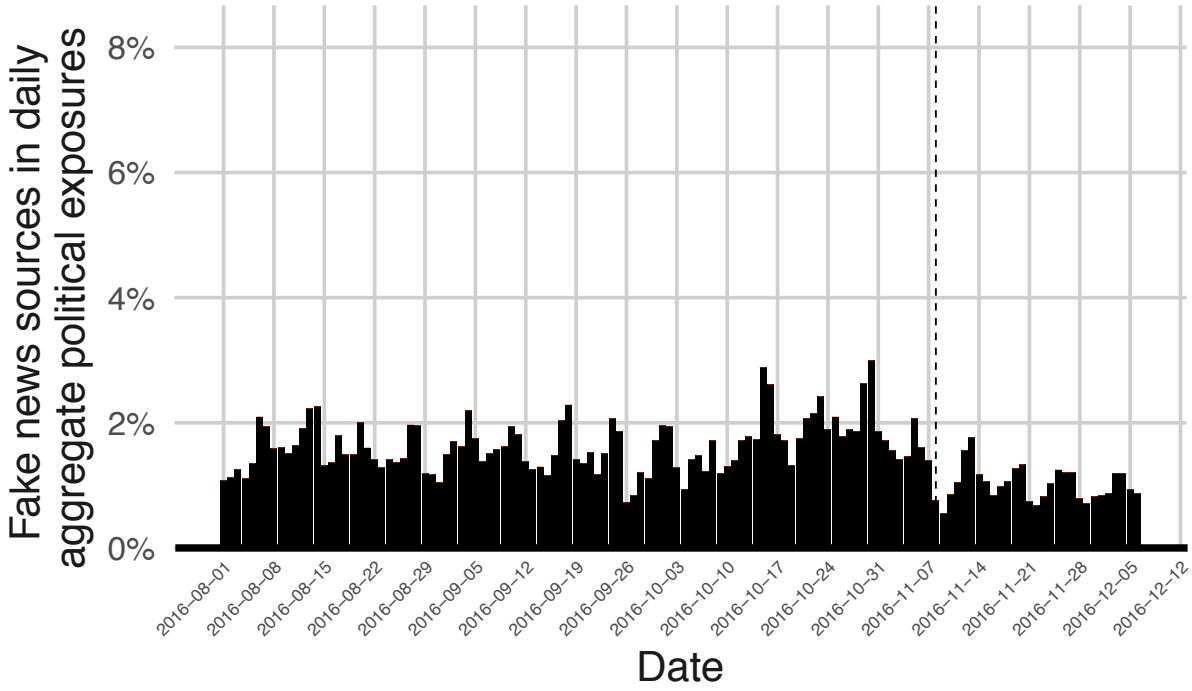

Figure S7: The same results as in Figure 1A in the main document, except with the list constructed by (*9*).

matching used more complex text processing (see Section S.1), whereas these checks used simple string matching.

We then manually labeled the profile pictures of 200 accounts. Two annotators independently examined each picture. First, annotators determined whether a single individual appeared in the photograph and whether the photograph seemed recent. Out of the 200 accounts sampled, 131 accounts passed this step. Those that did not often had either multiple individuals in the photograph or a picture of a non-human (e.g., a pet). For these 131 accounts, annotators then estimated the person's gender, age, and race. We assessed inter-rater reliability using Krippendorff's *alpha*, which produces a score of 1 for perfect agreement and 0 for agreement no better than chance.

Annotators were unanimous in their estimates of gender (*alpha* = 1), and they had 97%

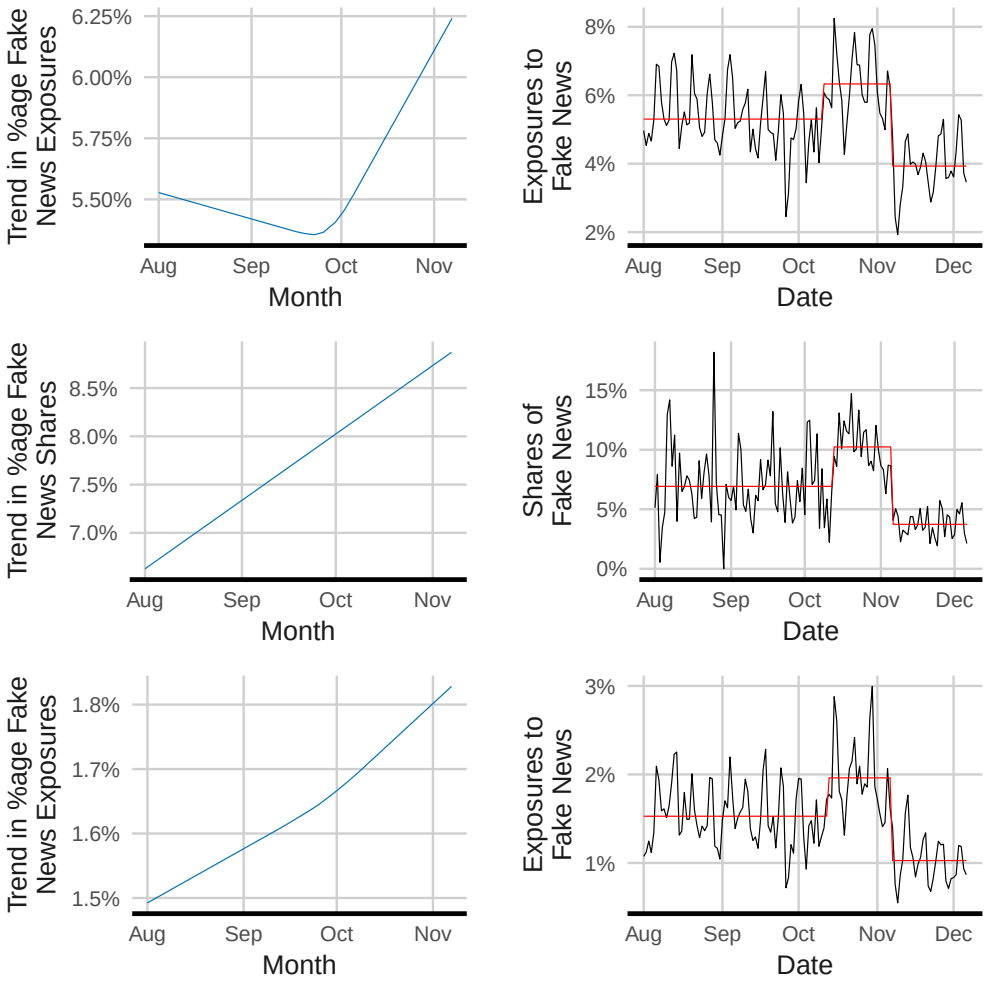

Figure S8: Left column: Estimated trend in percentage of exposures (top), shares (middle), and exposures using the Guess et al. list (bottom) of content from fake news sources overall by the panel from July 31st through election day. Right column: Estimated change points in percentage of exposures (top), shares (middle), and exposures using the Guess et al. list (bottom) of content from fake news sources overall by the panel from July 31st through December 6th, 2016. In each plot in the right column, the black line represents the data; the red line follows inferred means. Where the mean value changes, a changepoint was inferred.

agreement with the voter records on gender. Closer examination of the full profiles for the (four) discrepancies revealed that annotators were mistaken in two cases, the voter record appeared to be incorrect in one case, and one account that appeared to be a spammer.

For race, we annotated using the categories white, black, Asian and Hispanic. However,

we merged Hispanic into the white category upon noting that annotators' judgments of that distinction were effectively random. Using the three categories, annotators agreed for 126 of the 131 cases (*alpha* = 0.82). For the cases where they agreed, they matched the voter record 94% of the time.

For age, annotators had an *alpha* of 0.74. For each profile, we took the average of the two age estimates to produce a single "annotator estimate." Figure S9 displays the estimated vs. true ages. The annotator estimates were within 10 years of the current age 78% of the time, and within 5 years 54% of the time. This accuracy was higher for younger people: for those aged 50 or less, who constituted 80% of the sample, 88% of estimates were within 10 years of the current age. For those older than 50, only 38% of estimates were within 10 years, and 65% were within 15 years. Informal manual inspection of eight age outliers did uncover three mismatches, in which the Twitter account holder was a student much younger than the voter, and one more possibly spammer account.

Validation of outlier accounts

Upon examining the extreme statistics associated with the supersharer and superconsumer accounts, we were concerned that these accounts might not be run by the voters we had associated them with. Imperfect heuristics during panel construction might have allowed a voter to be matched to someone else's account, or worse, to a bot or deceptive profile. Bot and troll accounts—controlled by algorithms or malicious impersonators, respectively—play well-documented roles in manipulating political discourse (*21, 28, 29*). Accounts that look human are especially valuable and can be obtained by hijacking (taking over) normal accounts (*40*) or duplicating their profile data (*57*).

As a first safeguard to the panel's validity, we excluded the 141 ($< 1\%$ of) accounts having high BotOrNot scores (0.7 or higher, among accounts with at least 50 tweets) (*35*). However,

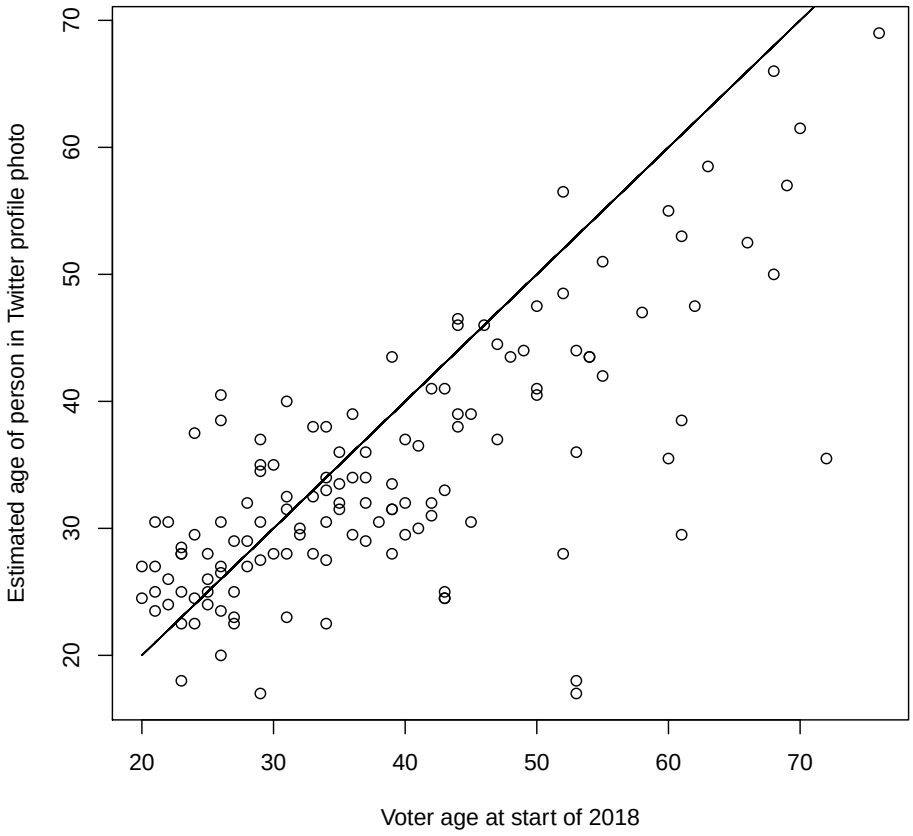

Figure S9: Voter age vs. perceived age of the person in their Twitter profile photo, as estimated by annotators. Krippendorff's *alpha* is 0.74.

when we informally spot-checked these accounts, most seemed to be controlled by the humans portrayed in the profiles, who tended to have extensive professional presences. More specifically, almost all of these high-exposure profiles contained links to professional websites. If that or another of the person's professional pages contained a link back to the Twitter profile, and if the Twitter content seemed consistent with their professional persona, we considered the account to be human-managed.

In order not to miss any human-run accounts in the analyses of supersharers and supercon-

sumers, we manually checked those accounts marked as bots who otherwise met the thresholds for supersharers or superconsumers—either overall (i.e., 1% of the population) or for content from fake news sources (responsible for 80% of content from fake news sources). All 15 of these "bot-like" accounts were judged human and restored to the panel (and counted among the 16,442).

This difference of judgment we had with BotOrNot we attribute partly to the difficulty of accurate bot detection, but also to a fundamental question of definitions. Most work on bot detection considers any automated behavior to stem from bots. However, many of the accounts we examined appear to be cyborgs, or partially automated accounts controlled by humans (*15*).

Second, we validated 23 outlier accounts by comparing the Twitter profile to the rest of the voter's online footprint. These accounts included the 16 fake news supersharers and the top 10 overall supersharers and superconsumers (many accounts overlapped among these sets). The check revealed two apparently hijacked accounts, which we excluded from the entire study. Each used stock profile photos, had modified the original owner's name, and posted content that was inconsistent with the voter's other social media presences. The profile information of the remaining accounts seemed consistent with the voter records and human-generated.

S.9 Account Statistics of Supersharers and other Panel Members

In addition to the supersharers of fake news sources (SS-F), the set responsible for 80% of shares from fake news sources, we defined *overall* supersharers as the top 1% of the panelists among those who shared one or more political URLs ($n = 38$, 0.2% of the entire panel). Similarly, we defined *overall* superconsumers as the 1% of panelists exposed to the most political URLs ($n = 164$). Like the SS-F accounts, many of the overall supersharers and superconsumers appeared to use automation. Some accounts posted mainly about politics, while others promoted their professional or business presences through high-volume posting or following of other accounts.

	Rest of panel	SC-R	SS-R	SS-F	
Size of group	16,260	142	24	16	
Portion of all political shares	42.6%	7.8%	24.6%	25.0%	
Portion of all shares from fake news sources	11.0%	5.5%	3.7%	79.8%	
Portion of all exposures to fake news sources	25.2%	36.1%	6.9%	31.8%	
Median by user:					
Total tweets / days in study	0.1	6.1	64.4	71.0	†
Shares of content from fake news sources	0	0	3	213	*†
Other shares (non-fake political)	0	9	687	759	†
Exposures to content from fake news sources	10	34,875	5360	130,000	*†
Other exposures (non-fake political)	6020	635,335	327,825	1,134,850	
Number of followers	104	1,479	1,633	1,583	
Number of followees	199	2,365	1,378	2,097	†
Political shares / total tweets	0	0.02	0.09	0.13	†
Political affinity score	-0.205	-0.252	-0.454	0.532	*†
Retweets / political shares	0.50	0.47	0.48	0.83	
"Via @"/ political shares	0	0.05	0.05	0.10	
Fraction of followees' accounts still active	1	0.98	0.99	0.95	*†
Fraction of accounts using "Via @" (outside of RTs or quotes)	0.143	0.469	0.833	0.938	†
Fraction of accounts verified	0.005	0.042	0.125	0.000	
Fraction of accounts male	0.447	0.577	0.542	0.250	†

Table S3: Activity measures for supersharers of content from fake news sources (SS-F, responsible for 80% of content from fake news sources), regular supersharers (SS-R: among top 1% of sharers but not in SS-F), regular superconsumers (SC-R: among top 1% of exposures, not in SS-F or SS-R), and the rest of the panel. Asterisks indicate where SS-F is significantly different than SS-R, using bootstrapped 95% confidence intervals; likewise, daggers indicate where SS-F is significantly different than SC-R. Notes: (1) "Followees' accounts still active": among a panel member's followees who had political URLs seen in the exposure data, the fraction of followees whose accounts were still open in February 2018. (2) The retweet and "via" statistics are calculated using only people that shared any political URLs.

However, as seen in Figure 2 in the main text, supersharers of fake news sources dominated the top of overall sharing and exposure distributions, sharing as many political URLs as the rest of the supersharers and consuming nearly a quarter of the superconsumers' political exposures.

Table S3 shows the cumulative percentages of political content attributable to the outlier

groups and to the rest of the panel. It also displays medians and averages of various activities for these groups across the (128-day) study period. Beyond the data in this table, the way in which SS-F shared content differed in qualitatively interesting ways from the other panel members. Specifically, when these accounts shared URLs, the accompanying text often simply repeated the article's headline without any modification or commentary; for some accounts, the text matched templates that also included hashtags, another URL, a comment, and/or an attribution to a source application or Twitter handle. Such an attribution, of the form "via @[source]", was seen at least once in 29% of the tweets from SS-F accounts (ignoring retweets and quotes) but only 12% from the rest of the panel.

We observed that supersharers of fake news sources were from across the country and were disproportionately aged 50 or above, Republican, and female; some listed professional experience in sales, communications, politics, religion, or the military. For the superconsumers of fake news sources, the distinguishing feature was the number of accounts they followed; beyond that, they were harder to characterize, as they posted infrequently.

Regarding the concentration of content from fake news sources, 61% of panel members had at least one exposure to a URL from a fake news source. Broken out by category, 1% of the panel was responsible for 86.6%, 83.7% and 74.8% of black, red and orange fake news source exposures, respectively. For sharing, 0.1% of the panel was responsible for 84.3%, 85.5% and 74.1% of all shares to black, red and orange sites, respectively.

S.10 Estimating Voters' Political Affinity

We devise a continuous political affinity score for panel members and evaluate its accuracy in three different ways. The score estimates the similarity of an individual's exposures to those of registered Democrats and Republicans using a logistic regression model, given individuals' news diet on Twitter and precinct-level vote share in the 2012 general election. With respect

to vote share, we use as a feature, for each individual panel member, the percentage of the vote received by Obama in the precinct in which their voter registration address is located[4]. As a measure of people's news diet, we average the political alignment of news sources the individual is exposed to on Twitter as described below.

We infer the political alignment of news sources using a method similar to the one used by Bakshy et al. (*58*), with a key distinction – we use exposure information rather that sharing of the source by partisans. This distinction lets us base our score on a much larger set of people who consume politics, but rarely tweet about it. As such, we compute a news source's alignment as the proportion of registered Republicans and Democrats who were exposed to the source, and reweight to correct for the imbalance of the two parties in our sample. In order to reduce the impact of cases where exposure to a news source is unlikely to reflect one's political affinity, we only consider individuals with a minimum of 100 observed exposures to politics, and sites that occupying 1% or more of all political URLs in a person's timeline. In addition, we only compute alignment scores for news sources with at least 30 registered voters. Fake news sources were excluded from the scores computation since a major part of our analysis pertains to the consumption of fake news as a dependent variable.

Evaluation: We evaluate our methodology by assessing its ability to predict party registration of held-out individuals, and by examining the congruence of our site-level alignment scores with those documented in the literature. Our logistic regression model, trained on data from 80% of registered votes, was able to predict party registration of the remaining 20% held-out individuals with high accuracy (AUC = 0.82). A fair amount of this accuracy stems from the precinct-level data alone (AUC = 0.73), which is further improved by including information about the alignment of sites in people's news feeds.

In terms of site alignment score, 109 sites overlapped with the list provided by Bakshy

[4]As compiled by the Huffington Post and retrieved from `https://github.com/huffpostdata/election-2012-results`

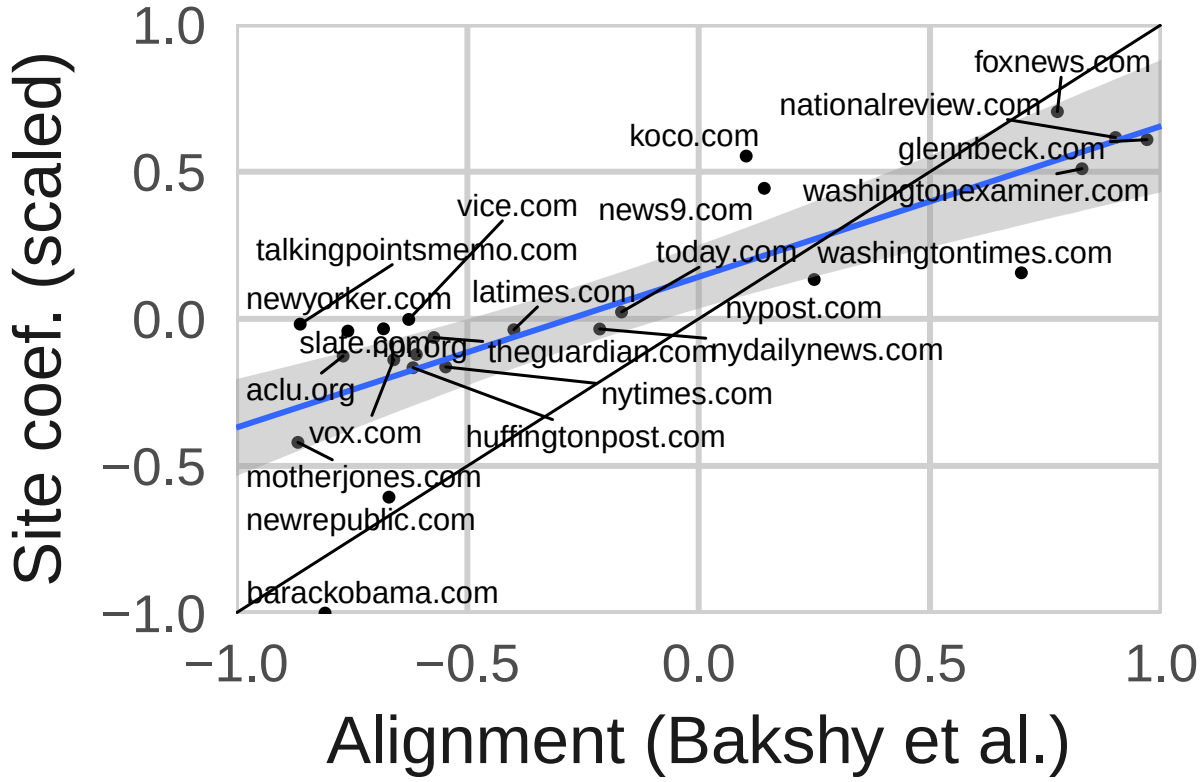

Figure S10: Correlation between our site-level coefficients and political alignment scores from Bakshy et al. (*58*).

et al., and a strong correlation exists between the two sets of scores (Pearson's $\rho = 0.91$, 95% CI=$[0.87, 0.94]$). Figure S10 shows the alignment scores of a sample of news sites that overlapped with the work of Bakshy et al. Similarly, we find a strong correlation with news sources slant, as computed by Budak et al. (*59*) using a mix of crowd-labels and a machine learning model. We matched 14 out of the 15 sites in their work with a Pearson correlation of $0.89\,[0.64, 0.96]$. The large confidence interval in this case stems from a three noticeable cases where the alignment scores diverge. We find the New York Times and the Chicago Tribune more left-leaning, and Fox News more right-leaning than Budak's work. Otherwise, the two sets of scores are consistent with other up to a scaling factor. The full set of site alignment scores computed in this work is publicly available at DOI *10.5281/zenodo.2483311*.

Finally, using the continuous political affinity score we assign individuals into distinct political affinity groups in the following manner. We divide the score range into seven equal parts. Individuals in the most extreme parts were assigned to "extreme left" or "extreme right". People in the central part of the score range were assign to the group "center". Individuals in the two remaining parts on left were assigned to "left", and similarly we assigned people on the remain two parts on the right to the "right" group. Supersharers and superconsumers were assigned to separate groups as described in the main body of the article, as well as "apolitical" individuals with little exposure to politics.

Table S4 provides descriptive statistics for the different subgroups. As mentioned in the main text, people who had 5% or more of their political exposures from fake news sources constituted 2.5% of individuals on the left (L and L*), and 16.3% of the right (R and R*). Ten thousand bootstrap samples from these two populations confirm that the differences are indeed statistically significant with none of the sample resulting in higher percentage on the left ($p < 10^{-4}$). Similarly, significant differences emerge in the fraction of people who shared of content from fake news sources across different subgroups. Among those who shared any political content in each subgroup, less than 5% of individuals on the center or left of it ever shared content from fake news sources, whereas more than 11% of individuals on right or the extreme right did so, respectively (bootstrapped samples $p < 10^{-4}$). The table shows other interesting patterns that are not discussed in the main results - for example, we see that in the last month of the campaign there was an average of 121 potential exposures for those on the extreme left, 172 for left, 246 for center, 476 for right, and 576 for extreme right. Note that there are 164 superconsumers when considering superconsumers who are also supersharers (as shown in Fig. 2), and 144 strictly superconsumers accounts as shown in Table S4.

	Extreme left	Left	Center	Right	Extreme right	A-political	Super-consumer	Super-sharer
N	1386	6011	2195	2609	570	3489	144	38
% Female	57.8	58.2	57.8	51.6	45.4	52.1	42.4	57.9
% White	80.9	82.3	85.8	90.4	94.6	83.9	93.1	84.2
% Swing state	14.4	29.6	31.8	28.3	25.6	31.2	26.4	31.6
Avg. followers to followees ratio	2.1	1.1	0.8	0.8	0.8	4.0	1.6	4.6
Avg. % of exposures to content from fake news sources	0.5	0.8	1.3	2.5	3.3	0.7	6.8	8.0
Avg. # of exposures to content from fake news sources[1]	121	172	246	476	576	2	35305	126090
% people with significant exposure to content from fake news sources[2]	1.2	2.8	6.3	15.4	21.1	3.5	45.8	47.4
% people who shared a fake news source[3]	4.9	4.6	4.8	11.6	21.3	3.7	37.8	86.8
Avg. # of weekly exposures to politics[4]	2684	2045	1495	1084	809	21	50259	91978
Total # of tweets[5]	2572	2583	2452	2602	1848	1165	10943	62512

Table S4: Summary statistics describing the different panel subgroups. Notes: [1] during the last month of the election; [2] significant level defined as having more than 5% of political exposures from fake news sources; [3] Among people who any shared political content on Twitter ; [4] excluding exposures to content from fake news sources; [5] excluding political URLs analyzed during the study period.

S.11 Regression Models for Exposure

We examined individual characteristics associated with increased exposure to content from fake news sources using a binomial regression model. Our dependent variable is the proportion of exposures to content from fake news sources relative to exposures of all political URLs. We thus model the *fraction* of political URLs that panel members were exposed to that came from fake news sites. We considered three classes of explanatory variables for each panel member: *demographic information* (age, gender, race and whether the individual resides in a swing state); *Twitter profile and activity information* (ratio of account followers to followees, total number of potential exposures to politics excluding content from fake news sources, and the total number of tweets posted, excluding political URLs shared during the study); and *political affinity* based on our assignment of panel members to political subgroups as described earlier.

We experimented with a variety of regression models and specifications; our exploration showed that fitting separate models to separate subgroups of panel members outperforms a single group model (in terms of both AIC and BIC). Due to the low number of superspreaders in the sample ($N = 38$), we excluded this subgroup from further analysis. We used a quasi-binomial model to better capture the overdispersion in our dependent variable towards no exposure to content from fake news sources and to obtain more conservative error estimates than a binomial model. In all cases, variance inflation factors (VIF) were smaller than four, suggesting a low degree of collinearity in our covariates.

Table S5 provides full results for the regression on exposure rate of content from fake news sources presented in the paper for each political subgroup described in the article. In addition to subsetting results by political subgroup, we performed a variety of robustness checks to ensure that main results – positive associations with political exposure and age, and distribution of scores in Figure 4 held across various potential artifacts in our analysis. First, in order to ensure that our results were not based on how panel members were split into political groups, we find

Table S5: Full results for regression on exposure rate to fake news sources for each political subgroup.

	Proportion of fake news sources in politics						
	extreme left	left	center	right	extreme right	apolitical	superconsumer
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Constant	-8.6^{***}	-7.5^{***}	-7.0^{***}	-7.2^{***}	-6.0^{***}	-4.8^{***}	-5.8^{***}
	$(-9.9, -7.4)$	$(-8.0, -6.9)$	$(-7.7, -6.2)$	$(-7.7, -6.7)$	$(-6.6, -5.5)$	$(-5.5, -4.1)$	$(-8.1, -3.6)$
Followers/followees	-0.2	0.1	0.3^{**}	-0.04	0.1	0.1	0.2
ratio (logged)	$(-0.5, 0.1)$	$(-0.1, 0.2)$	$(0.1, 0.6)$	$(-0.2, 0.2)$	$(-0.1, 0.3)$	$(-0.2, 0.4)$	$(-0.1, 0.5)$
Political exp.	0.4^{**}	0.2^{**}	0.5^{***}	0.8^{***}	0.9^{***}	-0.4^{**}	0.7^{***}
(weekly, logged)	(0.1, 0.8)	(0.02, 0.3)	(0.3, 0.7)	(0.7, 0.9)	(0.7, 1.1)	$(-0.7, -0.00)$	(0.3, 1.1)
Num. tweets	0.3^{**}	0.4^{***}	0.2^{***}	0.2^{***}	0.00	0.00	-0.2
(logged)	(0.05, 0.5)	(0.3, 0.5)	(0.1, 0.3)	(0.1, 0.3)	$(-0.1, 0.1)$	$(-0.1, 0.1)$	$(-0.4, 0.04)$
Age	0.02^{***}	0.02^{***}	0.01^{***}	0.02^{***}	0.01^{**}	0.02^{***}	0.01^{*}
	(0.01, 0.03)	(0.02, 0.03)	(0.01, 0.02)	(0.01, 0.02)	(0.00, 0.02)	(0.01, 0.02)	$(-0.00, 0.02)$
Male	0.1	0.3^{***}	0.4^{***}	0.3^{***}	0.1	-0.1	-0.1
	$(-0.2, 0.3)$	(0.2, 0.4)	(0.2, 0.5)	(0.2, 0.4)	$(-0.04, 0.3)$	$(-0.3, 0.1)$	$(-0.4, 0.2)$
Nonwhite	-0.2	0.1^{*}	-0.2	-0.3^{**}	-0.2	-0.3^{*}	0.2
	$(-0.5, 0.2)$	$(-0.02, 0.3)$	$(-0.5, 0.1)$	$(-0.6, -0.02)$	$(-0.5, 0.2)$	$(-0.6, 0.03)$	$(-0.2, 0.7)$
Swing state	1.0^{***}	-0.1^{**}	0.3^{***}	0.2^{***}	0.3^{***}	-0.1	0.3^{**}
	(0.8, 1.3)	$(-0.3, -0.01)$	(0.1, 0.4)	(0.1, 0.3)	(0.1, 0.4)	$(-0.3, 0.2)$	(0.04, 0.6)
Observations	1,386	6,011	2,195	2,609	570	3,489	144

Note: *p<0.1; **p<0.05; ***p<0.01

that main results are robust to small changes (5%) in the discretization of the score. Second, in order to ensure that results were not based on sites that produced low levels of fake news, we validated that our results were robust to removal of all orange websites from our fake news list. Finally, in order to ensure that results were not due to only a few websites, we validated that the results were robust to removing the top five fake news sources overall. While the percentage of exposures to content from fake news sources obviously changed, the general trends were found to be consistent in all cases.

S.12 Regression Model for Sharing

Table S6 provides the results of a logistic regression, modeling whether panel members shared any content from fake news sources during the course of the study as described in the main article text.

The covariates used to model sharing of content from fake news sources are similar to the ones in the exposure regression with some important distinctions. First, we focus on the 3,534

individuals who shared at least one political link during the course of the analysis, excluding superspreaders and superconsumers. Second, we use a logistic regression to distinguish between individuals who shared *any* URLs from fake news sources during the four months of the analysis[5]. Third, we fit a single logistic regression with dummy variables for the discretized political affinity score, as this model outperformed all other specifications in terms of AIC. Last, we omit the political exposure variable because it is moderately correlated with the number of political tweets people post (Pearson correlation of 0.35) and since we examine the relationship between exposure and sharing in greater depth in a separate analysis. We checked for collinearity by verifying that all variance inflation factors are smaller than four. We also assess the robustness of the results to the removal of the top five fake news sources, as well as removing all orange sites. We find here that patterns of both substantive effects and statistical significance are consistent across these different analyses.

S.13 Regression Models for Sharing Rates

In addition to modeling overall propensity to share content from fake news sources, we also examined the rate of fake news sharing *per exposure*. We linked each URL shared by a panel member with the most recent exposure that preceded it (i.e., the most recent tweet a followee posted with that same URL prior to the panel member tweeting or retweeting it). We excluded shares that could not be matched to a preceding exposure, and we considered unmatched exposures as cases that did not lead to sharing. This linking is insufficient to establish a causal path between exposure and sharing. However, it does preserve the temporal ordering of exposure and sharing behaviors and the logic that the influence of an exposure on a future share decays over time, which are important properties for a causal model of exposure. Based on this data, we used a logistic regression model to estimate the likelihood that a URL was shared by an

[5]Modeling the rate or count of sharing of content from fake news sources would be unstable in this case because the vast majority of people posted fewer than 10 political URLs.

	Likelihood of Sharing Content from Fake News Sources
	Center
Constant	-4.0^{***}
	$(-5.4, -2.6)$
Extreme Left	-0.4
	$(-1.1, 0.2)$
Left	-0.4
	$(-0.9, 0.2)$
Right	0.9^{***}
	$(0.4, 1.5)$
Extreme Right	1.7^{***}
	$(1.0, 2.3)$
Followers/followees	-0.9^{***}
ratio (logged)	$(-1.3, -0.4)$
Political tweets	2.4^{***}
(weekly, logged)	$(2.0, 2.7)$
Other tweets	0.00
(logged)	$(-0.3, 0.3)$
Fake news exp.	0.6^{***}
(weekly, logged)	$(0.5, 0.8)$
Age	0.02^{***}
	$(0.01, 0.03)$
Male	0.2
	$(-0.2, 0.5)$
Nonwhite	0.4^{*}
	$(-0.01, 0.9)$
Swing state	-0.1
	$(-0.5, 0.3)$
Observations	3,534
Note:	*p<0.1; **p<0.05; ***p<0.01

Table S6: Coefficients from a logistic regression modeling whether an individual who shared one political URL during the four months of the analysis will also share a fake news URL.

individual who was previously exposed to it.

The logistic regression model estimates the likelihood of sharing an exposed URL depending on a combination of three binary factors: source veracity, individual partisanship, and source partisanship. We control for individual differences by including the same demographic and Twitter profile information as previously described. Since sharing of URLs is a relatively rare event, happening roughly once every 10,000 exposures, we used a Bayesian logistic regression with a Cauchy(0, 2.5) prior to slightly encourage zero coefficients while allowing for large values when the data supports it (*60*).

We assigned partisanship to fake news sources based on a weighted average of political affinity scores of panel members exposed to a site. In order to reduce noise, we only consider individuals with a minimum of 100 observed exposures to politics overall, and at least 3 exposures to a site in question. We only compute the weighted political affinity scores for news sources with at least 10 individuals with the minimum exposure described above. Weighting was used to correct for the imbalance of partisans in the panel. We examined the distribution of weighted political affinity scores for the 123 fake news sites that satisfied the above criteria. Two distinct modes appeared in the distribution separated around an affinity score of -0.2. Manual examination of sites around this threshold confirmed that sites to the left (right) of the threshold were left-leaning (right-leaning). All sites examined appeared to have a clear political leaning. In addition, the partisanship assigned to fake news sites was highly consistent with the partisanship of source in the list by Bakshy et al. (*58*). The partisanship of hard news sources was simply assigned based on the polarity alignment scores, excluding the central seventh part of the score range. We found that 45 out of 46 (97.8%) of fake news sources appearing on the list by Bakshy et al. have the same partisanship assigned to them in both datasets. The single mismatch was for the site yournewswire.com, which upon manual inspection confirmed to be right-leaning as correctly labelled by our process. The full regression results are in Table S7.

	Sharing per exposure
Constant	-11.2^{***}
	$(-12.7, -9.7)$
Age	0.01^{*}
	(0.00, 0.01)
Male	-0.00
	$(-0.1, 0.1)$
Nonwhite	0.1
	$(-0.1, 0.3)$
Swing state	-0.5^{***}
	$(-0.7, -0.3)$
Followers/friends ratio (logged)	0.5^{***}
	(0.4, 0.7)
Num. nonpolitical tweets (logged)	0.8^{***}
	(0.6, 0.9)
Conservative exposed to congruent fake news	0.3
	$(-1.1, 1.7)$
Conservative exposed to congruent hard news	0.6
	$(-0.8, 1.9)$
Liberal exposed to congruent fake news	0.3
	$(-1.1, 1.7)$
Liberal exposed to congruent hard news	0.5
	$(-0.9, 1.8)$
Conservative exposed to incongruent fake news	-0.8
	$(-4.1, 2.4)$
Conservative exposed to incongruent hard news	0.4
	$(-1.0, 1.8)$
Liberal exposed to incongruent fake news	-1.2
	$(-2.7, 0.3)$
Liberal exposed to incongruent hard news	-0.4
	$(-1.8, 1.0)$
Observations	2,721,265
Log Likelihood	-7,826.4
Akaike Inf. Crit.	15,682.8
Note:	*p<0.05; **p<0.01; ***p<0.001

Table S7: Regression coefficients for the rate of sharing political URLs per exposure. The model controls for demographic and Twitter profile information, and includes the interaction of three binary variables: source veracity (fake or hard news), individual partisanship (conservative or liberal), and source partisanship (congruent or incongruent with an individual's partisanship).

S.14 Constructing and Analyzing the Co-exposure Network

In the main text we describe a co-exposure network intended to show patterns in consumption of different political news websites. Here we describe how this network is constructed and provide additional details about the clustering using an ensemble of clustering algorithms. In addition, Figure S11 provides a version of the co-exposure network where site names appear for each node.

To begin, we subset our analyses to websites that were labeled as fact-checking websites or popular sources of political news (or blog) websites (using the manual labeling process discussed in Section S.8 above), or websites we labeled as yellow, orange, red or black sources in our hand-coding task. We included fact-checking websites given previous work suggesting interesting and important potential links between exposure to fake news and fact-checking on Twitter (*61*). These 785 websites accounted for 67% percent of all potential exposures to panel members. We then construct a matrix M of these websites, where an entry within the matrix, M_{ij} determines the number of users exposed to at least one URL from website i that were also exposed to at least one URL from website j. In doing so, we consider only the subset of users who we did not classify as supersharers or superconsumers of non-fake or fake news, or as bots. This matrix can be considered as network, where the cell M_{ij} gives the link weight of the edge between i and j, and such a link exists when $M_{ij} > 0$.

As is, this network is extremely dense. Further, M is biased by the popularity of sites; two sites that are very popular will have a high value of the corresponding cell in M simply because of their popularity, and not necessarily because they share an important relationship with respect to co-exposure. How these issues impact network analyses has been well documented in the social network analysis literature, see (*62*) for recent work and a review. Essentially, network patterns dictated by high density and differences in popularity tend to obscure the multi-scale nature of network structure (*63*). To address these issues, we adopt the methodology proposed

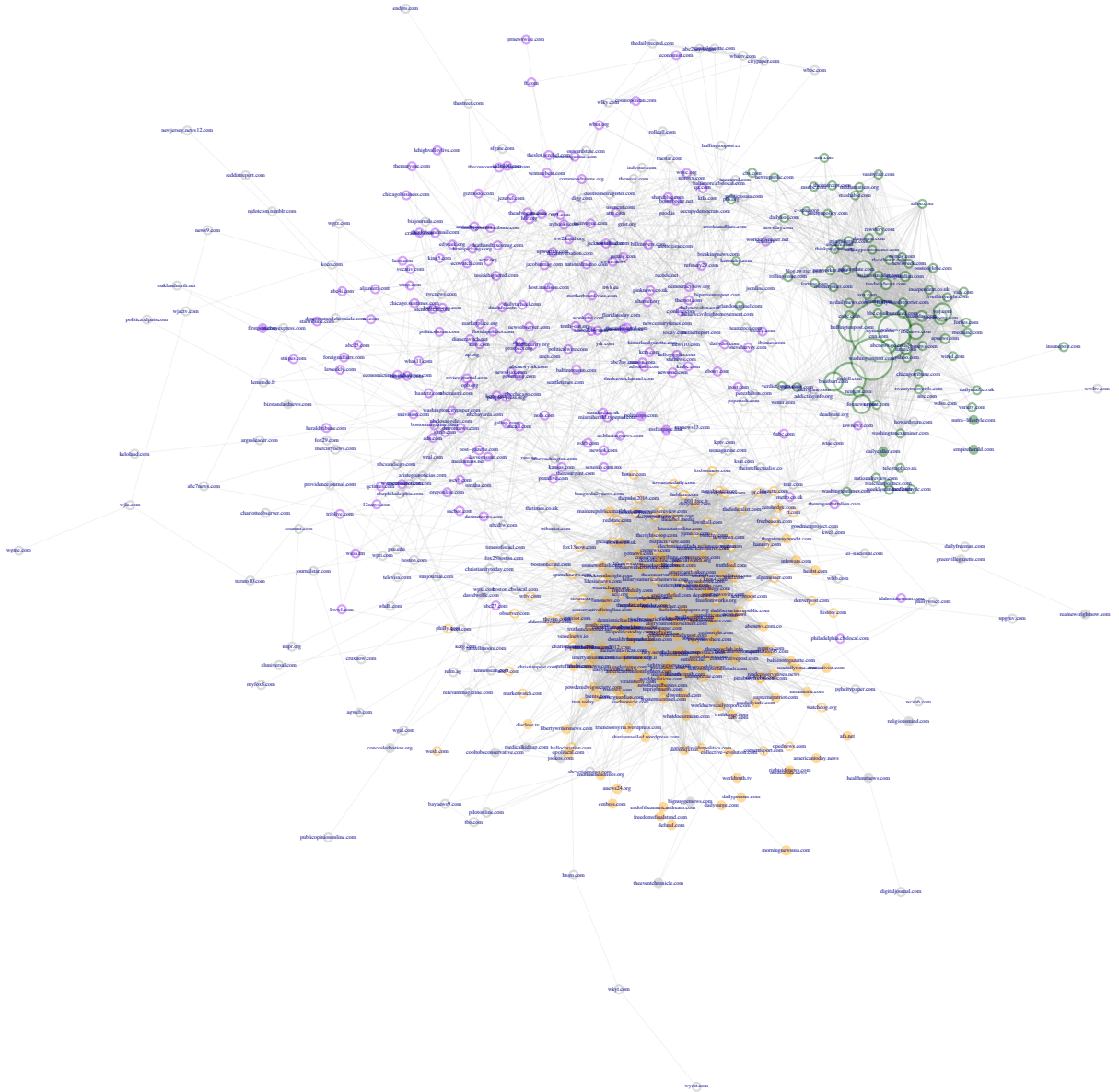

Figure S11: A labeled version of the co-exposure network shown in the main text.

by (*20*), which allows us to extract a co-exposure network that characterizes the backbone of the co-exposure network structure at multiple scales. This technique prevents the network structure from being dominated by site popularity. Roughly, this method retains an edge, M_{ij}, in M if an unusual number of users exposed to i are also exposed to j, relative to a null model based on

the expected number of users shared between two sites of similar levels of popularity, and sets $M_{ij} = 0$ otherwise.

As recommended by (*20*), we set a high threshold for deviation from the null model such that only the top 2%, most meaningful relationships between sites are retained. We set the weight of these edges to 1, and set all others to 0, resulting in a binary network. We further restrict our analysis to the largest weakly connected component of the network, which contains 99% of the sites in the binary network. This results in a network with 606 websites and 10,239 edges.

In the main text, we described an ensemble of network clustering techniques to identify three large groups of nodes, as well as a fourth group that showed inconsistent clustering patterns across the three algorithms. To identify these groups, we first cluster the network using three different, widely-used algorithms – Louvain (or multi-level) method (*64*), label propagation (*65*), and walktrap (*66*). The results of these three algorithms are shown in Figure S12. We then examined the frequency of all combinations of the three cluster labels. We found that 430 out 606 of websites belong to just three label combinations (three different groups of size 197, 150 and 83 nodes had the same values across all three algorithms, suggesting high consensus in their clustering) and that no other combination had more than 30 sites labeled.

By examining the labels from the three unsupervised clustering algorithms, we identify four groups - three large groups that are consistently identified across all clustering algorithms, and a fourth group of all other nodes in the network for comparison. Figure S11 presents a labeled version of Figure 5 in the main text so that websites within each group can be better understood.

In the main text we also describe levels of partisanship for each of the groups. To do so, we use an estimate of the partisan leaning of a website's audience from a prior work (*58*). Although computed in a similar fashion to our data, these estimates draw from vastly different data, and thus represent a relatively independent measure of site partisanship from those computed in the

present work. Across Groups 1, 2, 3 and 4, we identify 65, 79, 33, and 43 websites, respectively, for which partisanship scores were calculated by Bakshy et al. We then use these websites to estimate partisan leaning of each cluster. Statistical details of the claims made in the text are

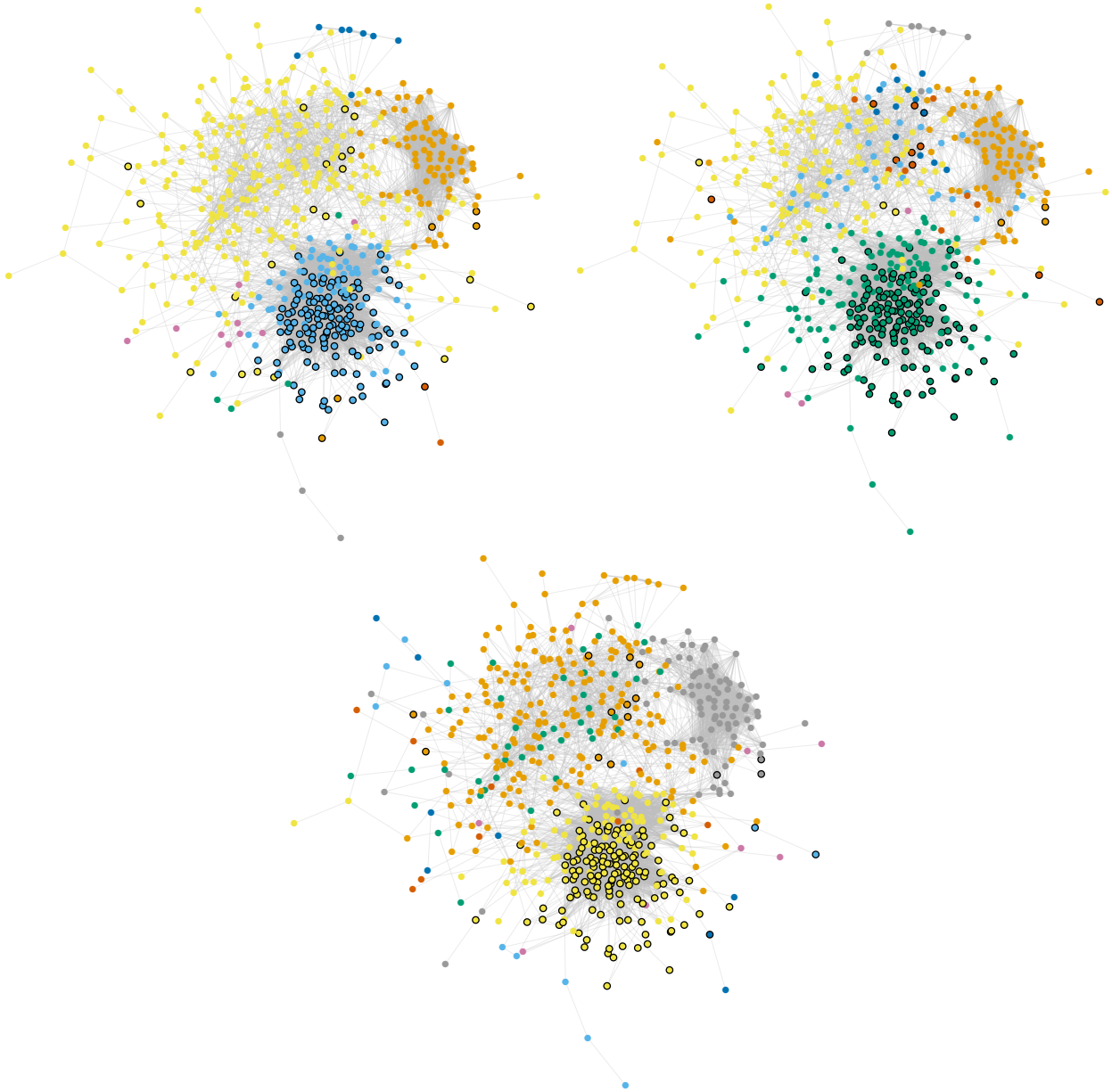

Figure S12: Three clustering algorithms applied to the co-exposure network: Louvain (top left), Label Propagation (top right), and WalkTrap (bottom).

given here - websites in Group 1 were significantly less conservative than those in Group 2 (one sided t-test; t = 11.33, df = 122.7, p-value $< .01$) and significantly less liberal than nodes in Group 3 (one sided t-test; t=2.24; df=117.44; $p < .01$). There was no significant difference in average political alignment of sites between Group 1 and Group 4 (two sided t-test; t = -0.29, df = 104.79, $p > .05$).

We also give details of statistical tests for differences in the number of fake news sites in each cluster - sites in Group 1 were significantly less likely to be fake news sources than in Group 2 (χ^2 = 98.08, df = 1, $p < .01$), but not Group 3 (χ^2 = 0.09, df = 1, p-value $> .05$) or Group 4 (χ^2 = 3.37, df = 1, p-value $> .05$). We also note that, as mentioned in the text,attention to Group 1 varied across partisanship: Group 1 made up 86%, 82%, 77%, 73% and 72% of exposures for those on the extreme left, left, center, right and extreme right, respectively. Websites in Group 1 were mostly centric in their political leaning with only a slight leaning to the left.

Finally, we provide additional details about Figure 5. We size each website by the total number of exposures, but set a minimum size in order to ensure that all nodes are visible. We use the Kamada and Kawai layout algorithm (*67*) to plot the network with the `igraph` package (*68*). The same layout is used for Figure S11 and the three network diagrams in Figure S12, thus one can compare labels directly to Figure 5 via position in the former and color directly to Figure 5 in the latter.

S.15 Concentration of Fake News

Concentration compared to other categories of content

In this section, we compare the concentration of exposure to and sharing of content from fake news sources to other kinds of content. The comparison consists of randomly sampling sites both in politics and outside of politics, sampling of sites similar to the fake news sites, and sites associated with a certain topic. In each case, we calculate the Gini coefficient to assess

concentration of URLs shared or potentially seen by panel members. For example, if each panel member had exactly the same number of potential exposures to a given set of sites, the Gini coefficient would be zero. In contrast, the Gini coefficient would approach one as more exposure volume gets concentrated in a small number of panel members. For the set of fake news sites appearing on our black, red or orange lists, the Gini coefficient is 0.96 for exposure and 0.88 for sharing. The analysis below shows that these concentrations levels of fake news sources are extremely high not only in absolute terms, but also relative to other categories of content shared or seen on Twitter.

The first set of comparisons are based on the sharing and exposure data only of political URLs, i.e. those used throughout this work. Within this set of comparisons to political URL content, we make four comparisons. The first consists of **political non-fake** sites, that is, any site in the data of political URLs that is not on our list of fake news sites. Second, we compare with a set of sites not on fake news lists that are most similar to fake news sites. Each fake news site is matched, without replacement, to five similar non-fake sites. The category **political matched vol.** refers to matching based on the total volume of the site's exposure or sharing, and the category **political matched ppl** refers to matching based on the total number of people exposed to or sharing content from a site. After matching, the set of fake and non-fake sites are statistically indistinguishable based on volume of exposure (sharing) or number of people exposed to (sharing) these sites. The fourth category for comparison is based on sites that appear with **political hashtags**. Since hashtags tend to represent a particular topic or issue of interest, this category enables a comparison with a set of sites that appear in a certain topical context. We identify hashtags posted along with a similar number of sites as the list of all fake news sites. For exposure we included hashtags appearing with $292 \pm 10\%$ sites and for sharing $173 \pm 10\%$ sites.

The second set of comparisons extends beyond politics. We analyze additional data from the

historical 10% sample of all tweets in our possession (the Twitter "Decahose"), and additional content from panel members obtained through the Twitter API. The extended dataset covers the same study period as the main dataset (Aug. 1–Dec. 6, 2016), but without the political classifier applied to it. Therefore, the data is approximately ten times larger in size. At this scale it is impractical to expand all URLs and follow redirects, and even expanding a meaningful sample of URLs is extremely time-consuming. Therefore, we take the approach of subsampling 10% of URLs obtained from followees of panel members, excluding URLs from known URL shorteners, and avoiding URL expansion altogether in order to keep the computation tractable. This approach yields a dataset with nearly six million URLs posted by followees of panel members and close to a million URLs posted by panel members, all pointing to web pages outside of Twitter. We analyze the concentration of content from sites in this dataset overall (**random** category), and by identifying hashtags that appear the same number of sites as described before (**hashtag** category).

We obtain a distribution of Gini coefficients for each category in the following manner. For categories that are not based on hashtags we draw 10,000 samples of domains from the category and compute the Gini coefficient. Each of the 10,000 samples subsets the entire exposure (sharing) data to a set of sites with equal probability from the category. In order to match the size of the fake news list, each sample consists of 292 sites for exposure or 173 sites for sharing, drawn from the list of sites in the category with equal probability. The distribution for categories based on hashtags are obtained by computing the Gini coefficient of exposure or sharing for every hashtag separately, subsetting the data to sites that appear with that hashtag.

Figure S13 shows the distribution of concentration levels (Gini coefficients) for the various categories along with the concentration levels of content from fake news sources (vertical, red, dashed line). The left panel shows concentration in potential exposures of panel members and the right panel shows concentration in sharing by panel members. Each curve shows the distri-

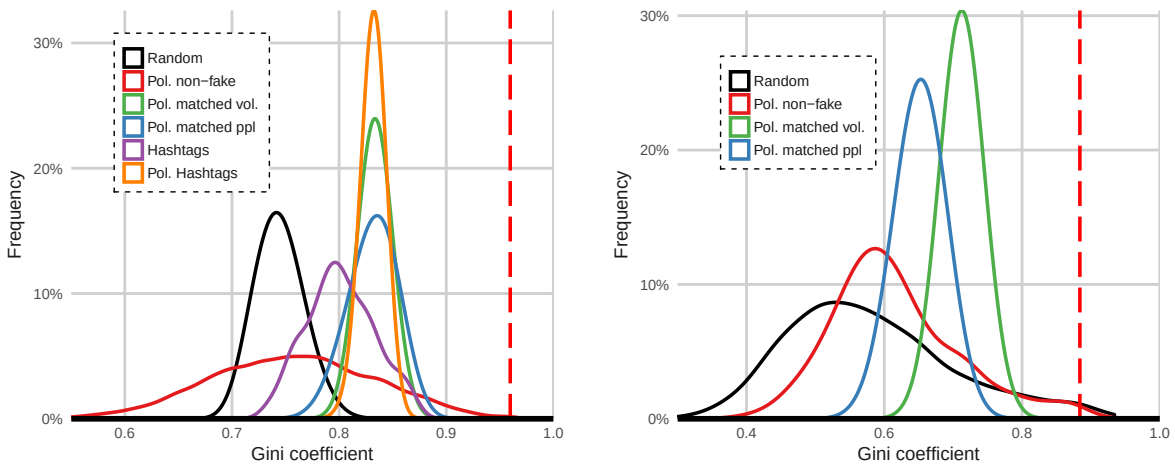

Figure S13: Concentration in exposure (left) and sharing (right) of content from fake news sources compared to other categories of content as measured by the Gini coefficient over panel members. The categories include: (1) Random sample of sites from all of Twitter (Random), (2) Random sample of sites with political content that are not on the fake news list (Pol. non-fake), (3) Political sites not on the fake news list that match the fake news sites in terms of total volume shared / exposed (Pol. matched vol.), (4) Political sites not on the fake news list that match the fake news sites in terms of total number of people sharing / exposed (Pol. matched ppl), (5) Sites that appear with different hashtags in all of Twitter, and (6) Sites that appear with different hashtags in political content. The distributions in categories 1–4 are based on 10,000 random samples of sites from the category, each of which subsetting the exposure or sharing data to the sampled sites. The distributions for categories 5–6 are based on the Gini coefficient calculated for each hashtag separately, when subsetting the data to sites that appear with the hashtag. Categories 5–6 are omitted from the right panel due to data sparsity.

bution of Gini coefficient values obtained in either the 10,000 samples from the category or the distribution among different hashtags. For example, the black curve on the left panel shows that a random sample of domains on Twitter is most likely to have a Gini coefficient of about 0.73. Exposure and sharing of content from political sites not on our fake news list are slightly more concentrated than a random sample of all sites, especially for sites with similar characteristics to the fake news sites and hashtags. The distributions for hashtag concentration in sharing were omitted because there were fewer than 10 hashtags appearing with $173 \pm 10\%$ sites. Across all categories of comparison – random, matched, topically-consistent, political or not – content

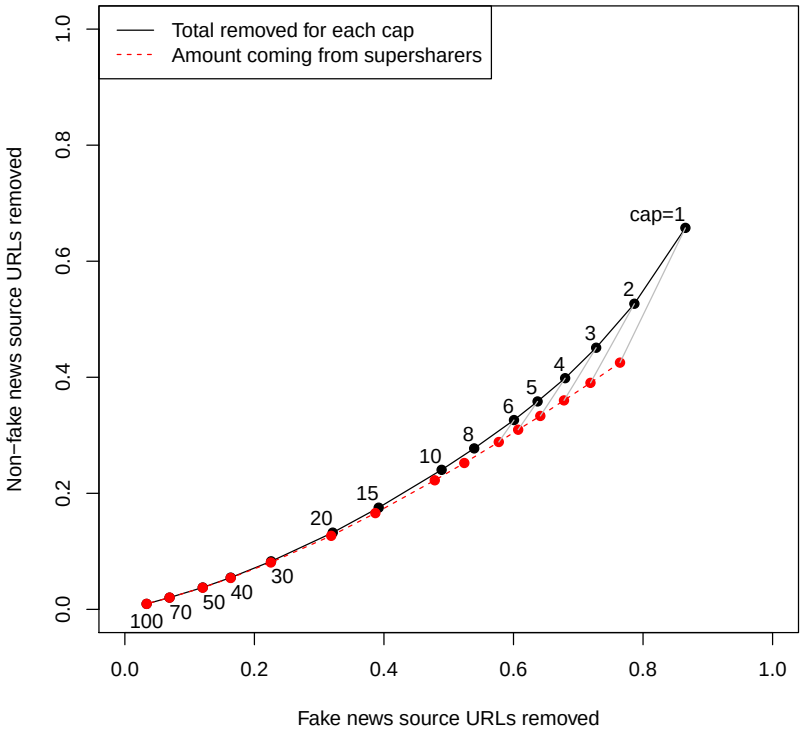

Figure S14: Fractions of the total political URLs from fake and non-fake news sources removed when simulating a daily cap on the shares of political URLs per user. Selected values from 1 to 100 are used as caps. Bootstrapped confidence intervals (not shown) are less than 0.001. Red dotted line shows how much of the amount removed comes from just the supersharer accounts (top 1%). Grey lines connect the same values of the cap and represent the remaining amounts, which come from non-supersharers.

from fake news sites was more concentrated than 99.45% of samples of sites shared and 99.97% of samples of sites potentially seen.

Concentration in people: capping simulation

Since most of the voter panel's shares of fake news sources come from a tiny population of supersharers, we explored whether it would be feasible to reduce the volume of fake news on Twitter by simply limiting the number of URLs each user could post per day. For this experiment, we re-analyzed the shares of political URLs by panel members but simulated a daily limit, or cap.

To simulate a given cap, for each person and day, we randomly sampled from their actual shares until the cap was reached. We analyzed the resulting volume of shares, subdivided into fake and non-fake sources, in comparison to the original (non-capped) volume. Figure S14 shows the fractions of fake and non-fake source URLs that would be removed with different caps. As we expected, more fake news than non-fake news is removed for any value of the cap. Reasonably high fractions of fake news (e.g., up to 40% of political URLs from fake news sources) could be prevented using caps that affect non-fake URLs at rates half as large.

These caps mainly affect content from supersharers, the top 1% of panel members in terms of political URLs shared, as the red line demonstrates. For instance, at a cap of 20 political URLs per day, 32.1% of fake news is reduced, of which 31.9% comes from supersharers and only 0.2% from the remaining 99% of the population (non-supersharers). Non-fake news is reduced by 13.2%, of which 12.7% comes from supersharers and only 0.6% from non-supersharers (differences due to rounding). The amounts removed from non-supersharers constitute 1.3% and 1.0% of all the fake and non-fake content, respectively, posted by non-supersharers.

References

1. E. McKernon, Fake news and the public, *Harper's Magazine* **151** (1925).

2. J. H. Kuklinski, P. J. Quirk, J. Jerit, D. Schwieder, R. F. Rich, Misinformation and the currency of democratic citizenship, *The Journal of Politics* **62**, 790 (2000).

3. C. Silverman, This analysis shows how viral fake election news stories outperformed real news on Facebook (2016). `https://www.buzzfeed.com/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook`.

4. D. Ruths, J. Pfeffer, Social media for large studies of behavior, *Science* **346**, 1063 (2014).

5. S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, *Science* **359**, 1146 (2018).

6. M. Del Vicario, *et al.*, The spreading of misinformation online, *Proceedings of the National Academy of Sciences* **113**, 554 (2016).

7. C. Shao, *et al.*, The spread of low-credibility content by social bots, *Nature Communications* **9**, 4787 (2018).

8. H. Allcott, M. Gentzkow, Social media and fake news in the 2016 election, *Journal of Economic Perspectives* **31**, 211 (2017).

9. A. Guess, B. Nyhan, J. Reifler, Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 U.S. presidential campaign (2018). `http://www.dartmouth.edu/~nyhan/fake-news-2016.pdf`.

10. Y. Benkler, R. Faris, H. Roberts, *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics* (Oxford Univ. Press, 2018).

11. K. Starbird, Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on Twitter, *in Proceedings of the 11th International AAAI Conference on Web and Social Media* (AAAI, 2017), pp. 230–239.

12. A. L. Schmidt, *et al.*, Anatomy of news consumption on Facebook, *Proceedings of the National Academy of Sciences* **114**, 3035 (2017).

13. D. M. J. Lazer, *et al.*, The science of fake news, *Science* **359**, 1094 (2018).

14. S. Greenwood, A. Perrin, M. Duggan, Social media update 2016, *Pew Research Center* (2016).

15. Z. Chu, S. Gianvecchio, H. Wang, S. Jajodia, Detecting automation of Twitter accounts: Are you a human, bot, or cyborg?, *IEEE Transactions on Dependable and Secure Computing* **9**, 811 (2012).

16. L. X. Wang, A. Ramachandran, A. Chaintreau, Measuring click and share dynamics on social media: A reproducible and validated approach, *in Workshops of the 10th International AAAI Conference on Web and Social Media* (AAAI, 2016), pp. 108–113.

17. P. N. Howard, B. Kollanyi, S. Bradshaw, L.-M. Neudert, Social media, news and political information during the US election: Was polarizing content concentrated in swing states?, *Data Memo 2017.8*, Oxford Project on Computational Propaganda (2017).

18. S. Bhatt, S. Joglekar, S. Bano, N. Sastry, Illuminating an ecosystem of partisan websites, *in Proceedings of the 27th International World Wide Web Conference* (ACM, 2018), pp. 545–554.

19. Z. Kunda, The case for motivated reasoning., *Psychological bulletin* **108**, 480 (1990).

20. N. Dianati, Unwinding the hairball graph: Pruning algorithms for weighted complex networks, *Physical Review E* **93**, 012304 (2016).

21. E. Mustafaraj, S. Finn, C. Whitlock, P. T. Metaxas, Vocal minority versus silent majority: Discovering the opinions of the long tail, *in Proceedings of the 3rd International Conference on Social Computing* (IEEE, 2011), pp. 103–110.

22. C. R. Sunstein, *#Republic: Divided democracy in the age of social media* (Princeton University Press, 2018).

23. B. Swire, U. K. Ecker, S. Lewandowsky, The role of familiarity in correcting inaccurate information., *Journal of Experimental Psychology: Learning, Memory, and Cognition* **43**, 1948 (2017).

24. Z. Tufekci, Big questions for social media big data: Representativeness, validity and other methodological pitfalls, *in Proceedings of the 8th International AAAI Conference on Weblogs and Social Media* (AAAI, 2014), pp. 505–514.

25. M. E. Roberts, *Censored: Distraction and Diversion Inside China's Great Firewall* (Princeton University Press, 2018).

26. N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, D. Lazer, Public replication package for Fake News on Twitter, Zenodo (2019). `https://doi.org/10.5281/zenodo.2483311`.

27. N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, D. Lazer, Protected replication data for Fake News on Twitter, Zenodo (2019). `https://doi.org/10.5281/zenodo.2485428`.

28. A. Bessi, E. Ferrara, Social bots distort the 2016 U.S. presidential election online discussion, *First Monday* **21** (2016).

29. S. C. Woolley, Automating power: Social bot interference in global politics, *First Monday* **21** (2016).

30. P. Barberá, Less is more? How demographic sample weights can improve public opinion estimates based on Twitter data. (2016). Working paper.

31. U. Pavalanathan, J. Eisenstein, Confounds and consequences in geotagged Twitter data, *in Proceedings of the Conference on Empirical Methods in Natural Language Processing* (ACL, 2015), pp. 2138–2148.

32. J. Kulshrestha, F. Kooti, A. Nikravesh, K. Gummadi, Geographic dissection of the Twitter network, *in Proceedings of the 6th International AAAI Conference on Web and Social Media* (AAAI, 2012).

33. D. Jurgens, T. Finethy, J. McCorriston, Y. Xu, D. Ruths, Geolocation prediction in Twitter using social networks: A critical analysis and review of current practice, *in Proceedings of the 9th International AAAI Conference on Web and Social Media* (AAAI, 2015).

34. K. Imai, K. Khanna, Improving ecological inference by predicting individual ethnicity from voter registration records, *Political Analysis* **24**, 263 (2016).

35. O. Varol, E. Ferrara, C. Davis, F. Menczer, A. Flammini, Online human-bot interactions: Detection, estimation, and characterization, *in Proceedings of the 11th International AAAI Conference on Web and Social Media* (AAAI, 2017).

36. A. Smith, M. Anderson, Social media use in 2018: Demographics and statistics, *Report*, Pew Research Center (2018).

37. T. Lumley, Analysis of complex survey samples, *Journal of Statistical Software* **9**, 1 (2004).

38. A. Agresti, B. A. Coull, Approximate is better than "exact" for interval estimation of binomial proportions, *The American Statistician* **52**, 119 (1998).

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

40. K. Thomas, F. Li, C. Grier, V. Paxson, Consequences of connectivity: Characterizing account hijacking on Twitter, *in Proceedings of the 21st Conference on Computer and Communications Security* (ACM, 2014), pp. 489–500.

41. G. Pennycock, T. D. Cannon, D. Rand, Implausibility and illusory truth: Prior exposure increases perceived accuracy of fake news but has no effect on entirely implausible statements, *Journal of Experimental Psychology: General* (in press).

42. M. Marchetti-Bowick, N. Chambers, Learning for microblogs with distant supervision: Political forecasting with Twitter, *in Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12 (ACL, 2012), pp. 603–612.

43. P. Vijayaraghavan, S. Vosoughi, D. Roy, Automatic detection and categorization of election-related tweets, *in Proceedings of the 10th International AAAI Conference on Web and Social Media* (AAAI, 2016).

44. N. Beauchamp, Predicting and interpolating state-level polls using Twitter textual data, *American Journal of Political Science* **61**, 490 (2017).

45. J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, *Journal of Statistical Software* **33**, 1 (2010).

46. E. C. Tandoc Jr, Z. W. Lim, R. Ling, Defining "fake news": A typology of scholarly definitions, *Digital Journalism* **6**, 137 (2018).

47. C. Silverman, J. Singer-Vine, The true story behind the biggest fake news hit of the election, `https://www.buzzfeed.com/craigsilverman/the-strangest-fake-news-empire` (2016).

48. C. Silverman, Here are 50 of the biggest fake news hits on Facebook from 2016, https://www.buzzfeed.com/craigsilverman/top-fake-news-of-2016 (2016).

49. C. Silverman, How a false story about a husband and wife being twins ended up on major news websites, https://www.buzzfeed.com/craigsilverman/this-false-story-about-a-husband-and-wife-discovering (2017).

50. C. Silverman, J. Singer-Vine, L. T. Vo, In spite of the crackdown, fake news publishers are still earning money from major ad networks, https://www.buzzfeed.com/craigsilverman/fake-news-real-ads (2017).

51. C. Silverman, S. Spary, Trolls are targeting Indian restaurants with a create-your-own fake news site, https://www.buzzfeed.com/craigsilverman/create-your-own-fake-news-sites-are-booming-on-facebook-and (2017).

52. J. Gillin, Politifact's guide to fake news websites and what they peddle, http://www.politifact.com/punditfact/article/2017/apr/20/politifacts-guide-fake-news-websites-and-what-they (2017).

53. S. Schaedel, Websites that post fake and satirical stories, http://www.factcheck.org/2017/07/websites-post-fake-satirical-stories (2017).

54. M. J. Metzger, Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research, *Journal of the Association for Information Science and Technology* **58**, 2078 (2007).

55. S. J. Taylor, B. Letham, Forecasting at scale, *The American Statistician* **72**, 37 (2018).

56. R. Killick, I. Eckley, Changepoint: An R package for changepoint analysis, *Journal of Statistical Software* **58**, 1.

57. O. Goga, G. Venkatadri, K. P. Gummadi, The doppelgänger bot attack: Exploring identity impersonation in online social networks, *in Proceedings of the 2015 Internet Measurement Conference* (ACM Press, 2015), pp. 141–153.

58. E. Bakshy, S. Messing, L. A. Adamic, Exposure to ideologically diverse news and opinion on Facebook, *Science* **348**, 1130 (2015).

59. C. Budak, S. Goel, J. M. Rao, Fair and balanced? Quantifying media bias through crowdsourced content analysis, *Public Opinion Quarterly* **80**, 250 (2016).

60. A. Gelman, A. Jakulin, M. G. Pittau, Y.-S. Su, A weakly informative default prior distribution for logistic and other regression models, *The Annals of Applied Statistics* pp. 1360–1383 (2008).

61. C. Shao, G. L. Ciampaglia, A. Flammini, F. Menczer, Hoaxy: A platform for tracking online misinformation, *in Proceedings of the 25th International World Wide Web Conference companion* (2016), pp. 745–750.

62. Z. Neal, The backbone of bipartite projections: Inferring relationships from co-authorship, co-sponsorship, co-attendance and other co-behaviors, *Social Networks* **39**, 84 (2014).

63. M. A. Serrano, M. Boguná, A. Vespignani, Extracting the multiscale backbone of complex weighted networks, *Proceedings of the national academy of sciences* **106**, 6483 (2009).

64. V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, *Journal of Statistical Mechanics: Theory and Experiment* **2008**, P10008 (2008).

65. U. N. Raghavan, R. Albert, S. Kumara, Near linear time algorithm to detect community structures in large-scale networks, *Physical review E* **76**, 036106 (2007).

66. P. Pons, M. Latapy, Computing communities in large networks using random walks, *International symposium on computer and information sciences* (Springer, 2005), pp. 284–293.

67. T. Kamada, S. Kawai, An algorithm for drawing general undirected graphs, *Information processing letters* **31**, 7 (1989).

68. G. Csardi, T. Nepusz, The igraph software package for complex network research, *InterJournal, Complex Systems* **1695**, 1 (2006).