Supplementary Materials for:

The spread of low-credibility content by social bots

Chengcheng Shao,[1,2] Giovanni Luca Ciampaglia,[3] Onur Varol,[1]
Kaicheng Yang,[1] Alessandro Flammini,[1,3] Filippo Menczer,[1,3*]

[1]School of Informatics, Computing, and Engineering, Indiana University, Bloomington, USA
[2]College of Computer, National University of Defense Technology, Changsha, Hunan, China
[3]Indiana University Network Science Institute, USA

[*]To whom correspondence should be addressed; E-mail: fil@iu.edu.

This PDF file includes:

- Supplementary Background
- Supplementary Methods
- Supplementary Text
- Tables S1–S3
- Figures S1–S13

Supplementary Background

Tracking abuse of social media has been a topic of intense research in recent years. The analysis in the main text leverages Hoaxy, a system focused on tracking the spread of links to articles from low-credibility and fact-checking sources (*38*). Here we give a brief overview of other systems designed to monitor the spread of misinformation on social media. This is related to

the problems of mining and detecting misinformation and fake news, which are the subjects of recent surveys (*39, 40*).

Beginning with the detection of simple instances of political abuse like *astroturfing* (*41*), researchers noted the need for automated tools for monitoring social media streams and detecting manipulation or misinformation. Several such systems have been proposed in recent years, each with a particular focus or a different approach. The Truthy system (*41*) relied on network analysis techniques to classify memes, such as hashtags and links. TraceMiner (*42*) also models the propagation of messages, but by inferring embeddings of social media users with social network structures. The TweetCred system (*43, 44*) focuses on content-based features and other kind of metadata, and distills a measure of overall information credibility. The Hierarchical Credibility Network (*45*) considers credibility as propagating through a three-layer network consisting of event, sub-events, and messages classified based on their features.

Specific systems have been proposed to detect rumors (*46*). These include RumorLens (*47*), TwitterTrails (*48*), FactWatcher (*49*), and News Tracer (*50*). The news verification capabilities of these systems range from completely automatic (TweetCred), to semi-automatic (TwitterTrails, RumorLens, News Tracer). In addition, some of them let the user explore the propagation of a rumor with an interactive dashboard (TwitterTrails, RumorLens). These systems vary in their capability to monitor the social media stream automatically, but in all cases the user is required to enter a seed rumor or keyword to operate them.

Our analysis is based on the spread of content from low-credibility sources rather than focusing on individual stories that are labeled as misinformation. Due to the impossibility to fact-check millions of articles, this approach of using sources as proxies for misinformation labels is increasingly adopted in the literature cited in the main text, and more (*51–55*).

Since misinformation can be propagated by coordinated online campaigns, it is important to detect whether a meme is being artificially promoted. Machine learning has been applied

successfully to the task of early discriminating between trending memes that are either organic or promoted by means of advertisement (*56*).

Finally, there is a growing body of research on social bot detection. The level of sophistication of bot-based manipulation can vary greatly (*57*). As discussed in the main text, there is a large gray area between human and completely automated accounts. So-called cyborgs are accounts used to amplify content generated by humans (*58*). It is possible that a significant portion of the manipulation discussed in this paper, aimed to amplify the spread of low-credibility content, is carried out by this kind of bot. The Botometer system used in this paper has been publicly available for a few years (*59*). Its earliest version was trained on simple spam bots, detected through a social honeypot system (*60, 61*). The version used here was trained on public datasets that also included more sophisticated bots.

Supplementary Methods

List of Sources

Our list of low-credibility sources was obtained by merging several lists compiled by third-party news and fact-checking organizations or experts. It should be noted that these lists were compiled independently of each other, and as a result they have uneven coverage. However, there is overlap between them. The full list of sources is shown in Table S1. Some lists annotate sources in different categories. In the case of OpenSources (`www.opensources.co`), we only considered sources tagged with any of the following labels: *fake, satire, bias, conspiracy, rumor, state, junksci, clickbait, hate.* In the case of Starbird's list of alternative domains (*52*), we considered those with primary orientation coded as one of *conspiracy theorists, political agenda, tabloid—clickbait news.*

For robustness analysis (below), we also consider a "consensus" subset of sites that are each listed among low-credibility sources by at least three organizations or experts. This subset

includes 65 sources, also shown in Table S1. We track 10,663,818 tweets (79% of the total) with links to 327,840 articles (86% of the total) from consensus low-credibility source, generated by 1,135,167 accounts (84% of the total).

We additionally tracked the websites of seven independent fact-checking organizations: `badsatiretoday.com`, `factcheck.org`, `hoax-slayer.com`,[1] `opensecrets.org`, `politifact.com`, `snopes.com`, and `truthorfiction.com`. In April 2017 we added `climatefeedback.org`, which does not affect the present analysis.

Table S1: Low-credibility sources. For each source, we indicate which lists include it. The lists are: Fake News Watch (FNW), OpenSources (OS), Daily Dot (DD), US News & World Report (US), New Republic (NR), CBS, Urban Legends (UL), NPR, Snopes Field Guide (Sn), Starbird Alternative Domains (KS), BuzzFeed News (BF), and PolitiFact (PF). Headers link to the original lists. The date indicates when Hoaxy started following a source: June 29 or December 20, 2016. Consensus sources (in three or more lists) are shown in italics.

Source	FNW	OS	DD	US	NR	CBS	UL	NPR	Sn	KS	BF	PF	Date
21stcenturywire.com	✓	✓	✓							✓			Jun
70news.wordpress.com		✓	✓			✓							Dec
abcnews.com.co		✓	✓			✓					✓	✓	Dec
activistpost.com	✓	✓	✓	✓						✓			Jun
addictinginfo.org		✓	✓							✓			Dec
americannews.com	✓	✓	✓	✓								✓	Jun
americannewsx.com		✓											Dec
amplifyingglass.com	✓												Jun
anonews.co			✓										Dec
beforeitsnews.com	✓	✓		✓						✓		✓	Jun
bigamericannews.com	✓	✓											Jun
bipartisanreport.com		✓	✓										Dec
bluenationreview.com		✓	✓										Dec
breitbart.com		✓	✓							✓			Dec
burrardstreetjournal.com		✓				✓					✓		Dec
callthecops.net		✓	✓				✓						Dec
christiantimes.com						✓							Dec
christwire.org	✓	✓	✓										Jun
chronicle.su	✓	✓											Jun
civictribune.com	✓	✓	✓			✓					✓	✓	Jun
clickhole.com	✓	✓	✓	✓									Jun
coasttocoastam.com	✓		✓										Jun
collective-evolution.com			✓										Dec
consciouslifenews.com	✓	✓	✓							✓			Jun

Continued on next page

[1] `hoax-slayer.com` includes its older version `hoax-slayer.net`.

Table S1 – continued from previous page

Source	FNW	OS	DD	US	NR	CBS	UL	NPR	Sn	KS	BF	PF	Date
conservativeoutfitters.com	✓	✓	✓										Dec
countdowntozerotime.com	✓	✓	✓										Jun
counterpsyops.com	✓	✓											Jun
creambmp.com	✓	✓	✓										Jun
dailybuzzlive.com	✓	✓		✓								✓	Jun
dailycurrant.com	✓	✓					✓				✓		Jun
dailynewsbin.com		✓											Dec
dcclothesline.com	✓	✓								✓			Jun
demyx.com					✓								Dec
denverguardian.com		✓						✓			✓		Dec
derfmagazine.com	✓	✓											Jun
disclose.tv	✓	✓		✓								✓	Jun
duffelblog.com	✓		✓	✓								✓	Jun
duhprogressive.com	✓	✓											Jun
empireherald.com		✓				✓					✓	✓	Dec
empirenews.net	✓	✓	✓			✓	✓		✓		✓	✓	Jun
empiresports.co	✓	✓			✓		✓		✓		✓	✓	Jun
en.mediamass.net	✓		✓		✓		✓						Jun
endingthefed.com		✓											Dec
enduringvision.com	✓	✓	✓										Jun
flyheight.com		✓											Dec
fprnradio.com	✓	✓											Jun
freewoodpost.com		✓					✓				✓	✓	Dec
geoengineeringwatch.org	✓	✓											Jun
globalassociatednews.com					✓		✓				✓		Dec
globalresearch.ca	✓	✓								✓			Jun
gomerblog.com	✓												Jun
govtslaves.info	✓	✓								✓			Jun
gulagbound.com	✓	✓											Jun
hangthebankers.com	✓	✓											Jun
humansarefree.com	✓	✓											Jun
huzlers.com	✓	✓			✓	✓	✓		✓		✓	✓	Jun
ifyouonlynews.com		✓				✓							Dec
infowars.com	✓	✓	✓	✓		✓				✓			Jun
intellihub.com	✓	✓								✓			Jun
itaglive.com	✓												Jun
jonesreport.com	✓	✓											Jun
lewrockwell.com	✓	✓								✓			Jun
liberalamerica.org		✓											Dec
libertymovementradio.com	✓	✓											Jun
libertytalk.fm	✓	✓											Jun
libertyvideos.org	✓	✓											Jun
lightlybraisedturnip.com					✓								Dec
nationalreport.net	✓	✓	✓		✓	✓	✓	✓	✓		✓	✓	Jun
naturalnews.com	✓	✓	✓	✓									Jun
ncscooper.com		✓							✓		✓		Dec
newsbiscuit.com	✓	✓	✓								✓		Jun

Continued on next page

Table S1 – continued from previous page

Source	FNW	OS	DD	US	NR	CBS	UL	NPR	Sn	KS	BF	PF	Date
newslo.com[a]	✓	✓	✓	✓							✓	✓	Jun
newsmutiny.com	✓	✓	✓										Jun
newswire-24.com	✓	✓											Jun
nodisinfo.com	✓	✓								✓			Jun
now8news.com		✓				✓			✓		✓	✓	Dec
nowtheendbegins.com	✓	✓											Jun
occupydemocrats.com		✓	✓							✓			Dec
other98.com		✓	✓										Dec
pakalertpress.com	✓	✓											Jun
politicalblindspot.com	✓	✓											Jun
politicalears.com	✓	✓											Jun
politicops.com[a]	✓	✓				✓					✓	✓	Jun
politicususa.com		✓											Dec
prisonplanet.com	✓	✓											Jun
react365.com		✓				✓			✓		✓	✓	Dec
realfarmacy.com	✓	✓											Jun
realnewsrightnow.com	✓	✓	✓			✓					✓	✓	Jun
redflagnews.com	✓	✓		✓						✓			Jun
redstate.com		✓	✓										Dec
rilenews.com	✓	✓	✓			✓					✓	✓	Jun
rockcitytimes.com	✓												Jun
satiratribune.com		✓							✓		✓	✓	Dec
stuppid.com		✓							✓		✓		Dec
theblaze.com		✓											Dec
thebostontribune.com		✓				✓					✓		Dec
thedailysheeple.com	✓	✓								✓			Jun
thedcgazette.com[b]	✓		✓	✓		✓							Jun
thefreethoughtproject.com		✓	✓							✓			Dec
thelapine.ca	✓						✓						Jun
thenewsnerd.com	✓	✓			✓						✓		Jun
theonion.com	✓	✓	✓	✓	✓	✓	✓						Jun
theracketreport.com		✓					✓				✓	✓	Dec
therundownlive.com	✓	✓											Jun
thespoof.com	✓	✓					✓						Jun
theuspatriot.com	✓	✓											Jun
truthfrequencyradio.com	✓	✓											Jun
twitchy.com			✓										Dec
unconfirmedsources.com	✓	✓											Jun
USAToday.com.co								✓	✓				Dec
usuncut.com		✓	✓										Dec
veteranstoday.com	✓	✓								✓			Jun
wakingupwisconsin.com	✓	✓											Jun
weeklyworldnews.com	✓	✓		✓			✓						Jun
wideawakeamerica.com	✓												Jun
winningdemocrats.com		✓											Dec
witscience.org	✓	✓									✓		Jun
wnd.com		✓											Dec

Continued on next page

Table S1 – continued from previous page

Source	FNW	OS	DD	US	NR	CBS	UL	NPR	Sn	KS	BF	PF	Date
worldnewsdailyreport.com	✓	✓	✓				✓		✓		✓	✓	Jun
worldtruth.tv	✓	✓		✓						✓		✓	Jun
yournewswire.com		✓				✓				✓	✓	✓	Dec

[a] `newslo.com` and `politicops.com` are mirrors of `politicot.com`.
[b] `thedcgazette.com` is a mirror of `dcgazette.com`.

Hoaxy Data

The back-end component of Hoaxy collects public tweets that link to a predefined list of websites. The use of the "POST statuses/filter" endpoint of the Twitter streaming API, together with the total volume of tweets collected, which is well below 1% of all public tweets, guarantee that we obtain all tweets linking to the sites in our list, and not just a sample of the tweets with these links. In addition, Hoaxy crawls all tracked websites and indexes all their articles, supporting a full-text search engine that allows users to find articles matching a given query. Furthermore, users can select subsets of these articles to visualize their spread on Twitter. To this end, Hoaxy matches the indexed articles with the tweets in our database and constructs networks based on retweets, mentions, replies, and quoted tweets. The front-end visualizes these networks interactively, allowing users to explore the accounts (nodes) and the tweets (edges) that make up these networks. The system makes all the data accessible to the public through a website (`hoaxy.iuni.iu.edu`) and an API.

Our analysis focuses on the period from mid-May 2016 to the end of March 2017. During this time, we collected 15,053 and 389,569 articles from fact-checking and low-credibility sources, respectively. The Hoaxy system collected 1,133,674 public posts that included links to fact checks and 13,617,425 public posts linking to low-credibility articles. As shown in Fig. S1, low-credibility websites each produced approximately 100 articles per week, on average. Toward the end of the study period, this content was shared by approximately 30 tweets per article per week, on average. However, as discussed in the main text, success is extremely heteroge-

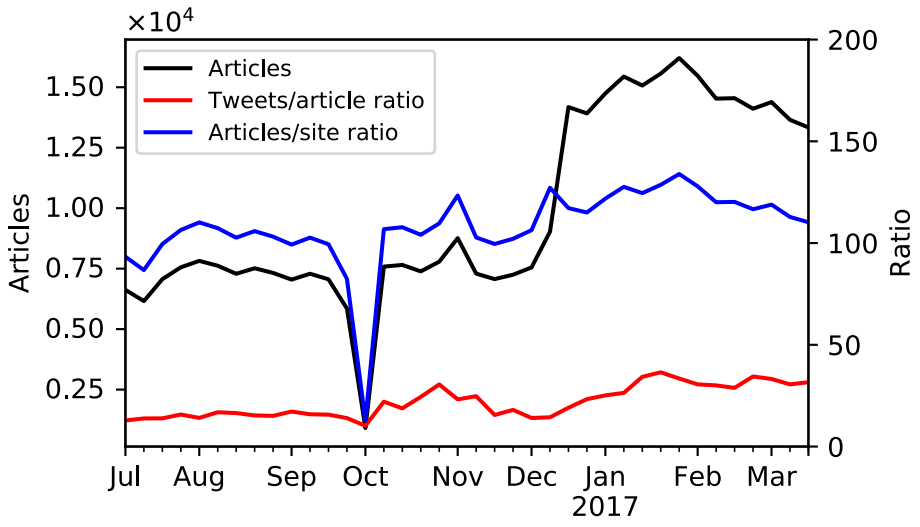

Figure S1: Weekly tweeted low-credibility articles, tweets/article ratio and articles/site ratio. The collection was briefly interrupted in October 2016. In December 2016 we expanded the set of low-credibility sources, from 70 to 120 websites.

neous across articles. This is the case irrespective of whether we measure success through the number of tweets (Fig. S2(a)) or accounts (Fig. S2(b)) sharing an article. For both popularity measures, the distributions are very broad and basically indistinguishable across articles from low-credibility vs. fact-checking sources.

Content Analysis

Our analysis considers content published by a set of websites flagged as sources of misinformation by third-party journalistic and fact-checking organizations (Table S1). This source-based approach relies on the assumption that most of the articles published by our compilation of sources are some type of misinformation, as we cannot fact-check each individual article. We validated this assumption by estimating the rate of false positives, i.e, verified articles, in the corpus. We manually evaluated a random sample of articles ($N = 50$) drawn from our corpus, stratified by source. We considered only those sources whose articles were tweeted at least once

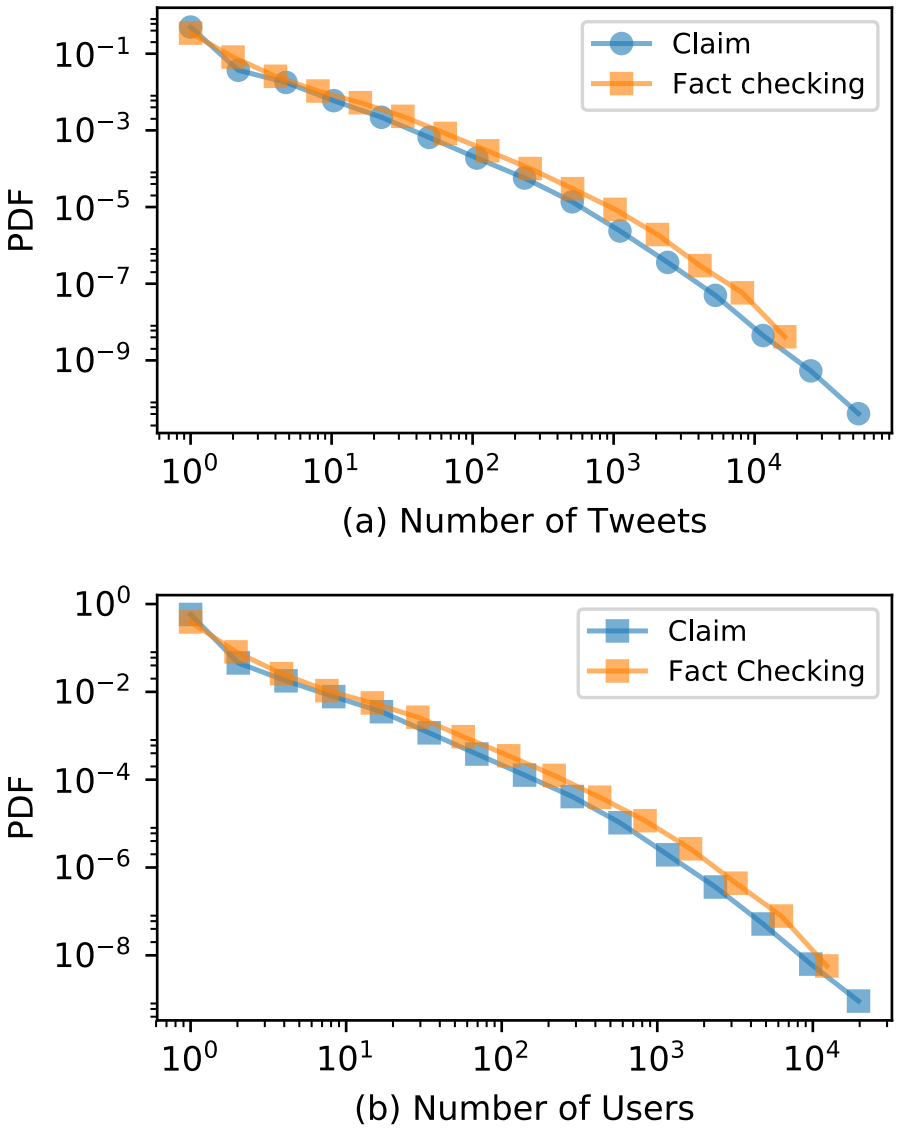

Figure S2: Probability distributions of popularity of articles from low-credibility and fact-checking sources, measured by (a) the number of tweets and (b) the number of accounts sharing links to an article.

in the period of interest. To draw an article, we first selected a source at random with replacement, and then chose one of the articles it published, again at random but without replacement. We repeated our analysis on an additional sample ($N = 50$) in which the chances of drawing an article are proportional to the number of times it was tweeted. This 'sample by tweet' is thus biased toward more popular sources.

It is important to note that articles with unverified claims are sometimes updated after being debunked. This happens usually late, after the article has spread, and could lead to overestimating the rate of false positives. To mitigate this phenomenon, the earliest snapshot of each article was retrieved from the Wayback Machine at the Internet Archive (`archive.org`). If no snapshot was available, we retrieved the version of the page current at verification time. If the page was missing from the website or the website was down, we reviewed the title and body of the article crawled by Hoaxy. We gave priority to the current version over the possibly more accurate crawled version because, in deciding whether a piece of content is misinformation, we want to consider any form of visual evidence included with it, such as images or videos.

After retrieving all articles in the two samples, each article was evaluated independently by two reviewers (two of the authors), using a rubric summarized in Fig. S3. Each article was then labeled with the majority label, with ties broken by a third reviewer (another author). Fig S4 shows the results of the analysis. We report the fractions of articles that were verified and that could not be verified (inconclusive), out of the total number of articles that contain any factual claim. The rate of false positives is below 15% in both samples.

Concentration

In the main text we use the Gini coefficient to calculate the concentration of posting activity for an article, based on the accounts that post links to the article. For each article, the Lorenz curve plots the cumulative share of tweets versus the cumulative share of accounts generating

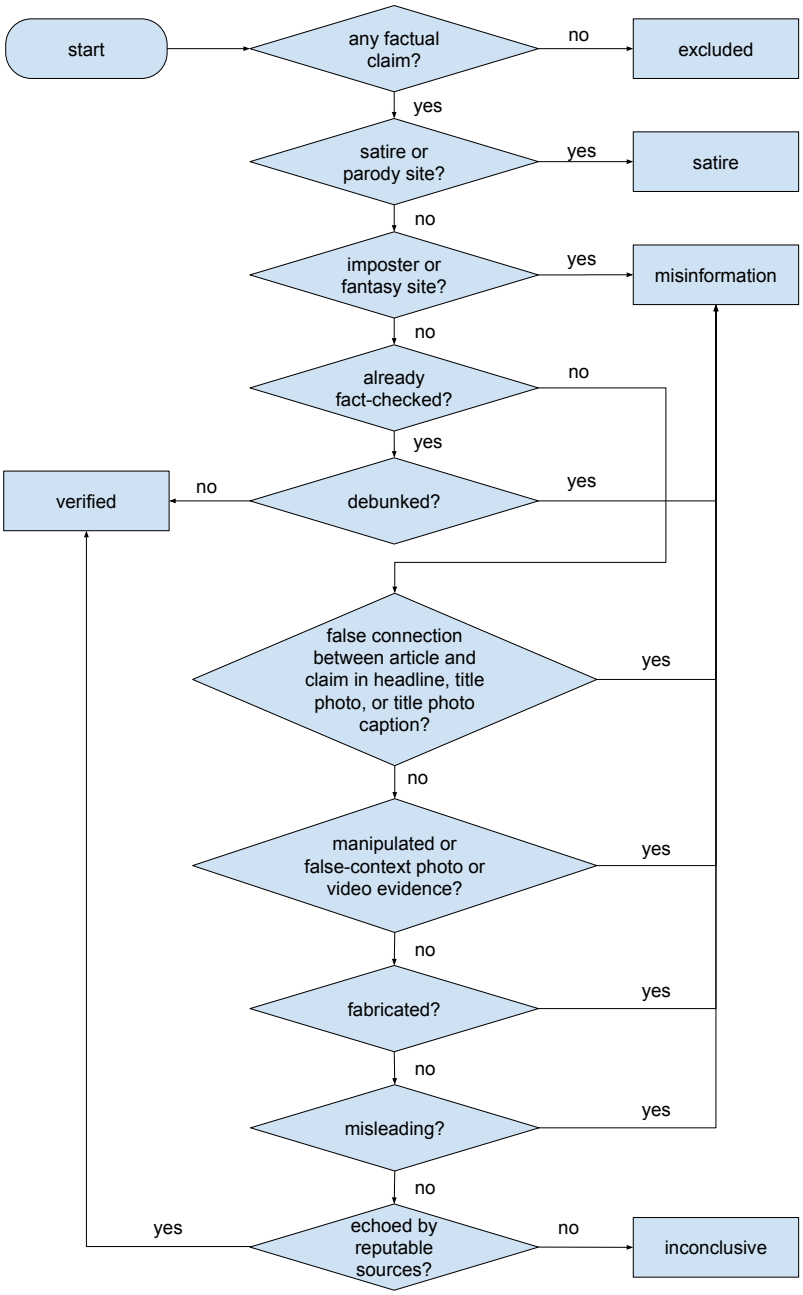

Figure S3: Flowchart summarizing the annotation rubric employed in the content analysis.

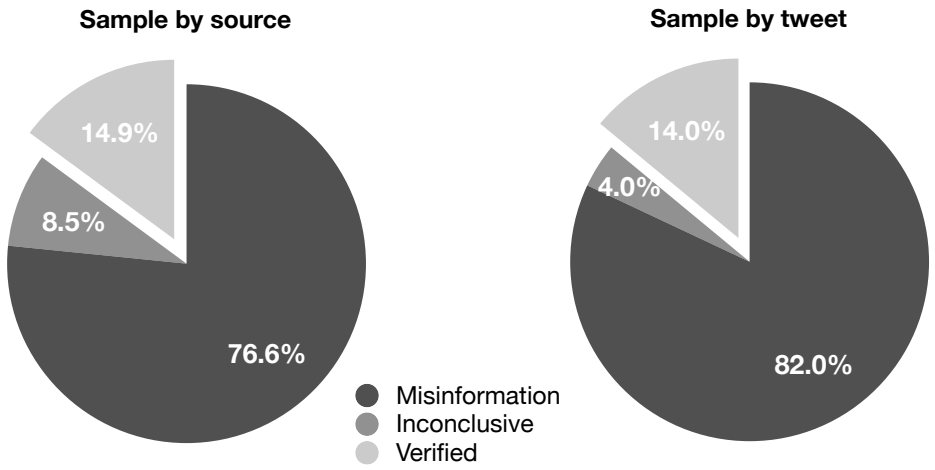

Figure S4: Content analysis based on two samples of articles. Sampling by source gives each source equal representation, while sampling by tweets biases the analysis toward more popular sources. We excluded from the sample by source three articles that did not contain any factual claims. Satire articles are grouped with misinformation, as explained in the main text.

these tweets. The Gini coefficient is the ratio of the area that lies between the line of equality (diagonal) and the Lorenz curve, over the total area under the line of equality. A high coefficient indicates that a small subset of accounts was responsible for a large portion of the posts.

Bot Classification

To show that a few social bots are disproportionately responsible for the spread of low-credibility content, we considered a random sample of accounts that shared at least one article from a low-credibility source, and evaluated these accounts using the bot classification system Botometer. Out of 1,000 sampled accounts, 85 could not be inspected because they had been either suspended, deleted, or turned private. For each of the remaining 915, Botometer returned a *bot score* estimating the level of automation of the account. To quantify how many account are likely bots, we transform bot scores into binary assessments using a threshold of 0.5. This is a conservative choice to minimize false negatives and especially false positives, as shown in prior work (cit. in main text). Table S2 shows the fraction of accounts with scores above the

Table S2: Analysis of likely bots and their content spreading activity based on a random sample of Twitter accounts sharing at least one article from a low-credibility source.

	Total	Likely bots	Percentage
Accounts	915	77	8%
Tweets with low-credibility articles	11,656	3,857	33%
Unique low-credibility articles	7,726	2,819	36%
Tweets with fact-checks	598	27	5%
Unique fact-checks	395	25	6%

threshold. To give a sense of their overall impact in the spreading of low-credibility content, Table S2 also shows the fraction of tweets with articles from low-credibility sources posted by accounts that are likely bots, and the number of unique articles included in those tweets overall. As a comparison, we also tally the fact-checks shared by these accounts, showing that accounts that are likely bots tended to focus on sharing low-credibility content.

In the main text we show the distributions of bot scores for this sample of accounts, as well as for a sample of accounts that have been most active in spreading low-credibility content (*super-spreaders*). To select the super-spreaders, we ranked all accounts by how many tweets they posted with links to low-credibility sources, and considered the top 1,000 accounts. We then performed the same classification steps discussed above. For the same reasons mentioned above, we could not obtain scores for 39 of these accounts, leaving us with a sample of 961 scored accounts.

Supplementary Text

Super-Spreaders of Low-Credibility Content

In the main text we show that the more popular a low-credibility article, the more its posting activity is concentrated around a relative small number of active accounts. We also find that the most active spreaders of content from low-credibility articles are more likely to be social bots. To further illustrate the anomalous activity patterns of these "super-spreaders", Fig. S5 plots the

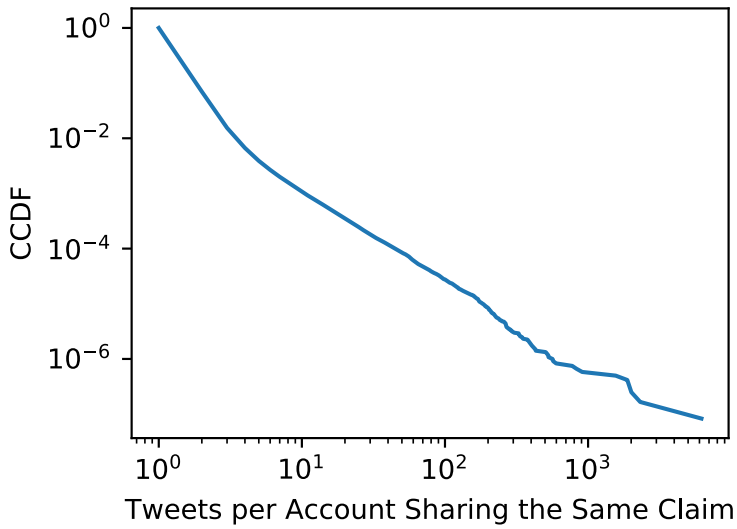

Figure S5: Cumulative distribution of repetitions, i.e., the number of times a single account tweets the same link to an article from a low-credibility source.

distribution of repeated tweets by individual accounts sharing the same low-credibility article. While it is normal behavior for a person to share an article once, the long tail of the distribution highlights inorganic, automated support. A single account posting the same article over and over — hundreds or thousands of times in some cases — is likely controlled by software.

Bots Targeting Influentials

The main text discusses a strategy used by bots, by which influential users are mentioned in tweets that link to low-credibility content. Bots seem to employ this targeting strategy repetitively. Fig. S6 offers an illustration: in this example, a (now suspended) single account produced 19 tweets linking to the article shown in the figure and mentioning `@realDonaldTrump`.

Amplification by Bots

The analysis in the main text focuses on the role of bots in the spread of articles from low-credibility sources, assuming that bots do not equally support the spread of articles from fact-checking sources. In fact, we show in the main text that articles from low-credibility and fact-

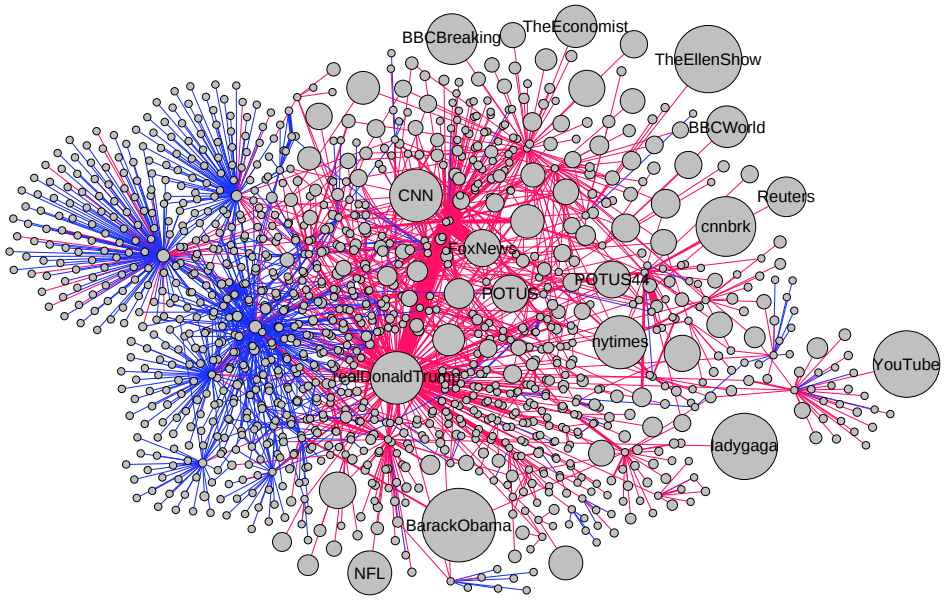

Figure S6: Example of targeting for the article *Report: three million votes in presidential election cast by illegal aliens*, published by `Infowars.com` on November 14, 2016 and shared over 18 thousand times on Twitter. Only a portion of the diffusion network is shown. Nodes stand for Twitter accounts, with size representing number of followers. Links illustrate how the article spreads: by retweets and quoted tweets (blue), or by replies and mentions (red).

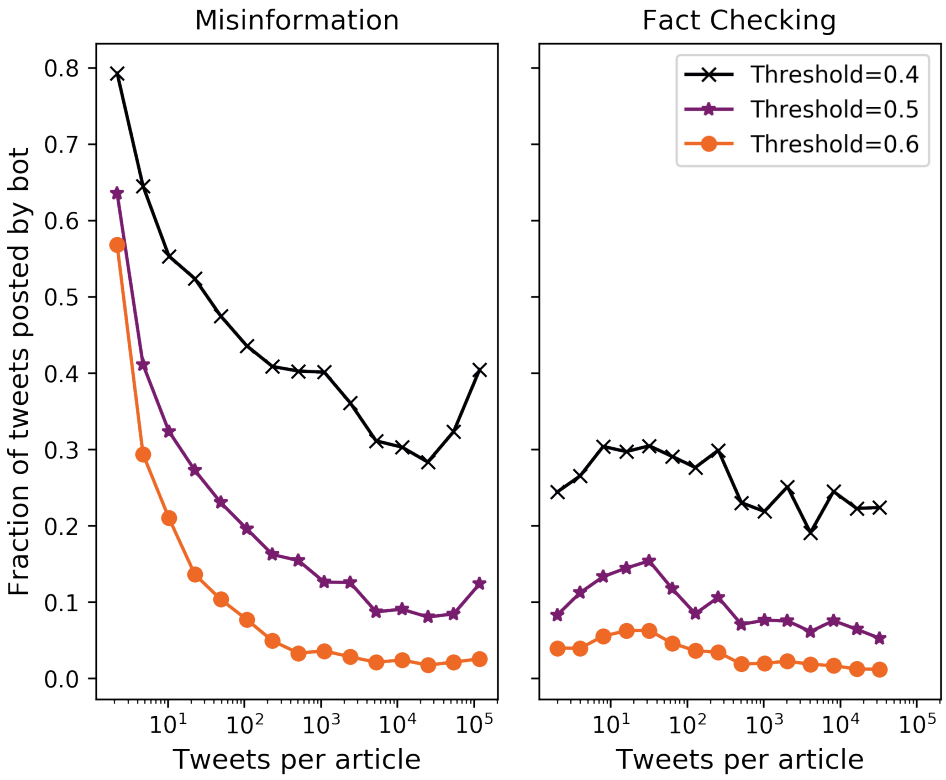

Figure S7: Fraction of tweets linking to news articles that are posted by accounts with bot score above a threshold, as a function of the popularity of the linked articles. We see different bot activity for articles from low-credibility (left) versus fact-checking (right) sources.

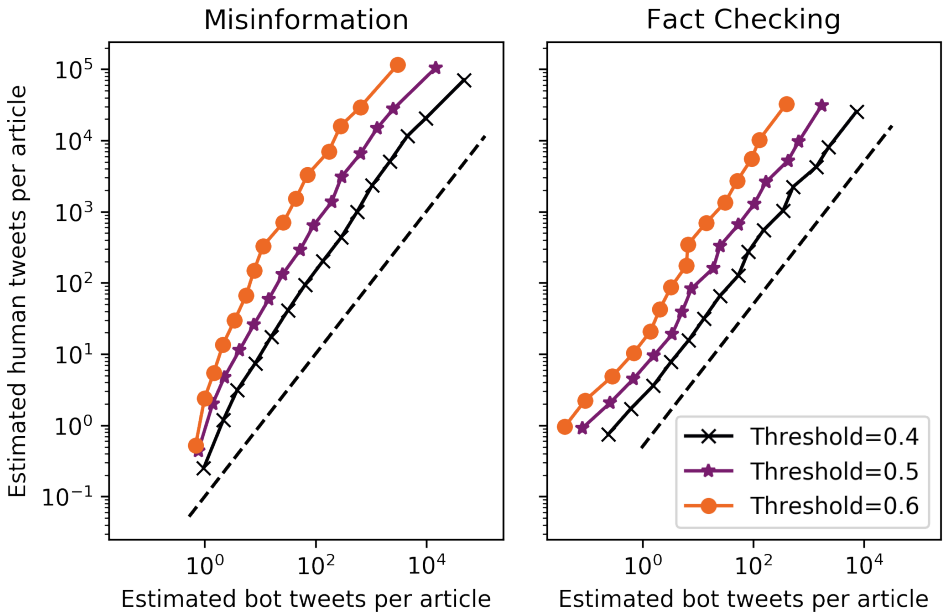

Figure S8: For links to articles from low-credibility (left) and fact-checking (right) sources, the number of tweets by accounts with bot score above a threshold is plotted versus the number of tweets by accounts with bot score below the threshold. A super-linear relationship is a signature of amplification by bots.

checking sources spread through different mixes of original tweets, retweets, and replies. And we also find that low-credibility sources have greater support from bots than fact-checking and satire sources. To further confirm the assumption that bots do not play an equal role in the spread of fact-checking articles, we observe in Fig. S7 that the fraction of tweets posted by likely bots is much higher for articles from low-credibility sources. Further, the fraction depends on popularity in the case of articles from low-credibility sources (it gets diluted for more viral ones), whereas it is flatter for articles from fact-checking sources. Here, bots and humans are separated based on a threshold in the bot score. These findings are robust to the choice of threshold, and point to selective amplification of articles from low-credibility sources by bots.

To focus on amplification more directly, let us consider how exposure to humans varies

Table S3: Spearman rank correlation ρ between U.S. state tipping probability and relative bot activity. High tipping probabilities indicate 2016 swing states (data from FiveThirtyEight at `projects.fivethirtyeight.com/2016-election-forecast/`). Bot activity is measured by numbers of tweets posting links to low-credibility articles by accounts with bot score above threshold that reported a U.S. state location in their profile, in the period between August and October 2016. The counts by states are normalized by those expected from a sample of 1,393,592,062 tweets in the same period, obtained from a 10% random sample of public posts from the Twitter streaming API. High p-values indicate that the correlation is not significant.

Bot score threshold	Tweets by likely bots	ρ	p-value
0.4	2,166,625	0.09	0.5
0.5	939,647	0.12	0.4
0.6	304,434	0.23	0.1

with activity by bots. Fig. S8 estimates the numbers of tweets by likely humans/bots, using a threshold on bot scores to separate them. Results are robust with respect to the choice of threshold. For articles from low-credibility sources, the estimated number of human tweets per article grows faster than the estimated number of bot tweets for article. For fact-checking articles, instead, we find a linear relationship. In other words, bots seem to amplify the reach of articles from low-credibility sources, but not the reach of articles from fact-checking sources.

Geographic Targeting

We examined whether bots (or rather their programmers) tended to target voters in certain states by creating the appearance of users posting from those locations. To this end, we considered accounts with bot scores above a threshold that shared articles from low-credibility sources in the three months before the election, and focused on those accounts with a state location in their profile. The location is self-reported and thus trivial to fake. We compared the distribution of bot account locations across states with a baseline obtained from a large sample of tweets in the same period. A χ^2 test indicates that the location patterns produced by bots are inconsistent with the geographic distribution of conversations on Twitter ($p < 10^{-4}$). This suggests that as part of their disguise, social bots are more likely to report certain locations than others. However,

we did not find evidence that bots used this strategy for targeting swing states (Table S3).

Robustness Analyses

The results in the main text are robust with respect to various choices and assumptions, presented next.

Criteria for selection of sources

We repeated the analyses in the main text using the more restrictive criterion for selecting low-credibility sources, based on a consensus among three or more news and fact-checking organizations. The 65 consensus sources are listed in Table S1. To carry out these analyses, we inspected 33,115 accounts and could obtain bot scores for 32,250 of these; the rest had been suspended or gone private. The results are qualitatively similar to those in the main text and support the robustness of the findings, namely: super-spreaders of articles from low-credibility sources are likely bots (Fig. S9), bots amplify the spread of information from low-credibility sources in the early phases (Fig. S10), bots target influential users (Fig. S11), and humans retweet low-credibility content posted by bots (Fig. S12).

Absence of correlation between activity and bot score

Our notion of super-spreader is based upon ranking accounts by activity and taking those above a threshold. The analysis about super-spreaders of low-credibility content being likely bots assumes that this finding is not explained by a correlation between activity and bot score. In fact, although the bot classification model does consider volume of tweets as one among over a thousand features, it is not trained in such a way that there is an obvious monotonic relation between activity and bot score. A simple monotonic relation between overall volume and bot score would lead to many false positives, because many bots produce very few tweets or appear to produce none (they delete their tweets); these accounts still get high bot scores. Figure S13

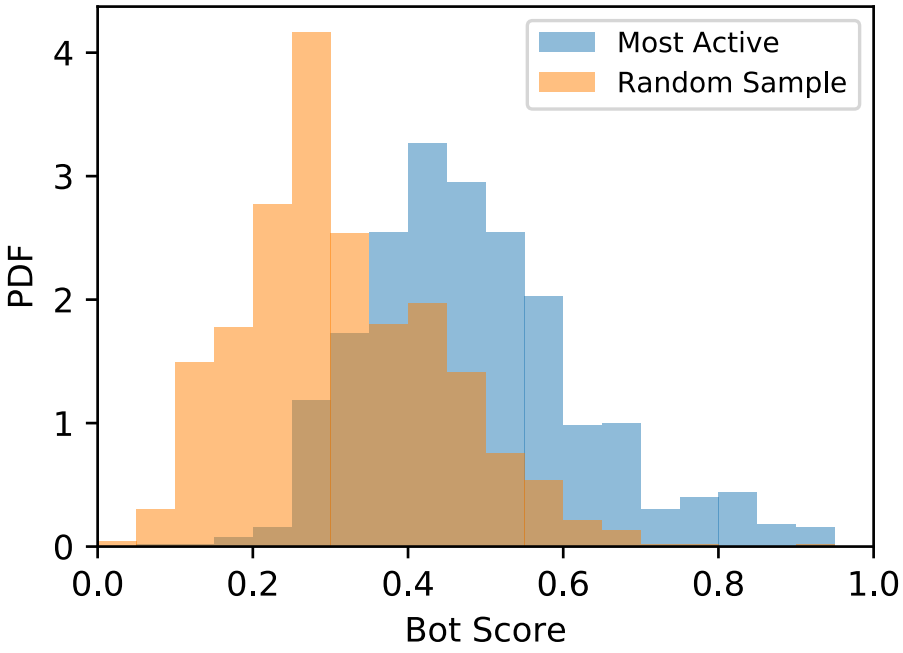

Figure S9: Bot score distributions for super-spreaders vs. randomly selected sharers of links to low-credibility sources selected by the consensus criterion. The random sample includes 992 accounts who posted at least one link to an article from a low-credibility source. Their bot scores are compared to 997 accounts that most actively share such links. The two groups have significantly different scores ($p < 10^{-4}$ according to a Mann-Whitney U test). 7% of accounts in the random sample and 37% of accounts in the most active group have bot score above 0.5.

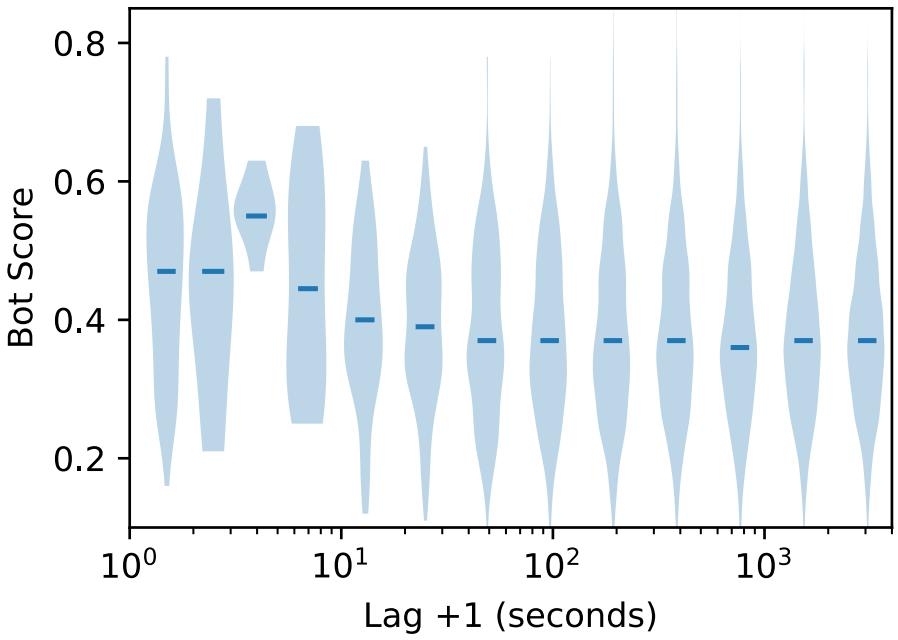

Figure S10: Temporal evolution of bot support after the first share of a viral story from a consensus low-credibility source. We consider a random sample of 20,000 accounts out of the 163,563 accounts that participate in the spread of the 1,000 most viral articles. After articles from from *The Onion* are excluded, we are left with 42,202 tweets from 13,926 accounts. We align the times when each link first appears. We focus on a one-hour early spreading phase following each of these events, and divide it into logarithmic lag intervals. The plot shows the bot score distribution for accounts sharing the links during each of these lag intervals.

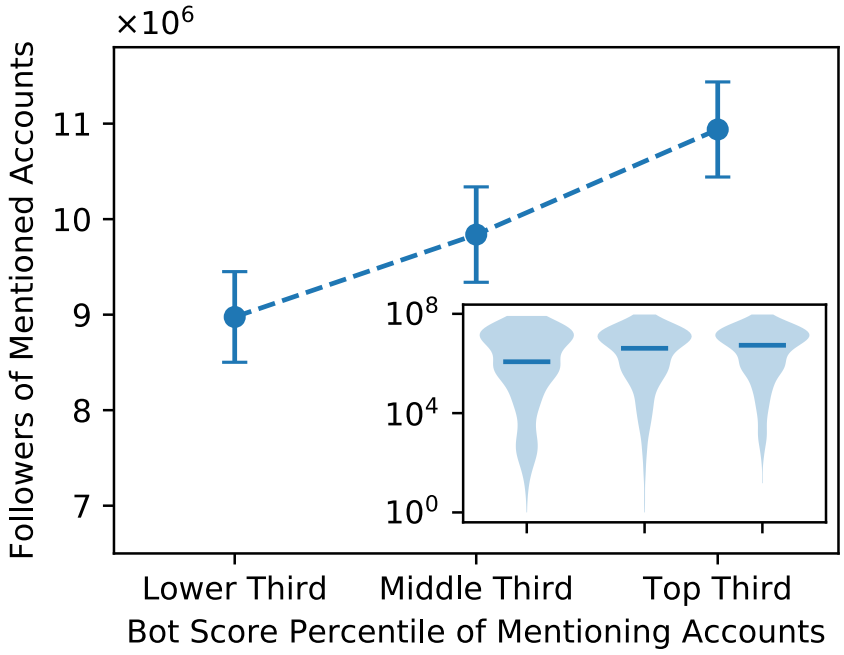

Figure S11: Average number of followers for Twitter users who are mentioned (or replied to) by a sample of 20,000 accounts that link to the 1,000 most viral articles from consensus low-credibility sources. We obtained bot scores for 4,006 unique mentioning accounts and 4,965 unique mentioned accounts, participating in 33,112 mention/reply pairs. We excluded 13,817 of these pairs using the "via @screen_name" mentioning pattern. The mentioning accounts are aggregated into three groups by bot score percentile. Error bars indicate standard errors. Inset: Distributions of follower counts for users mentioned by accounts in each percentile group.

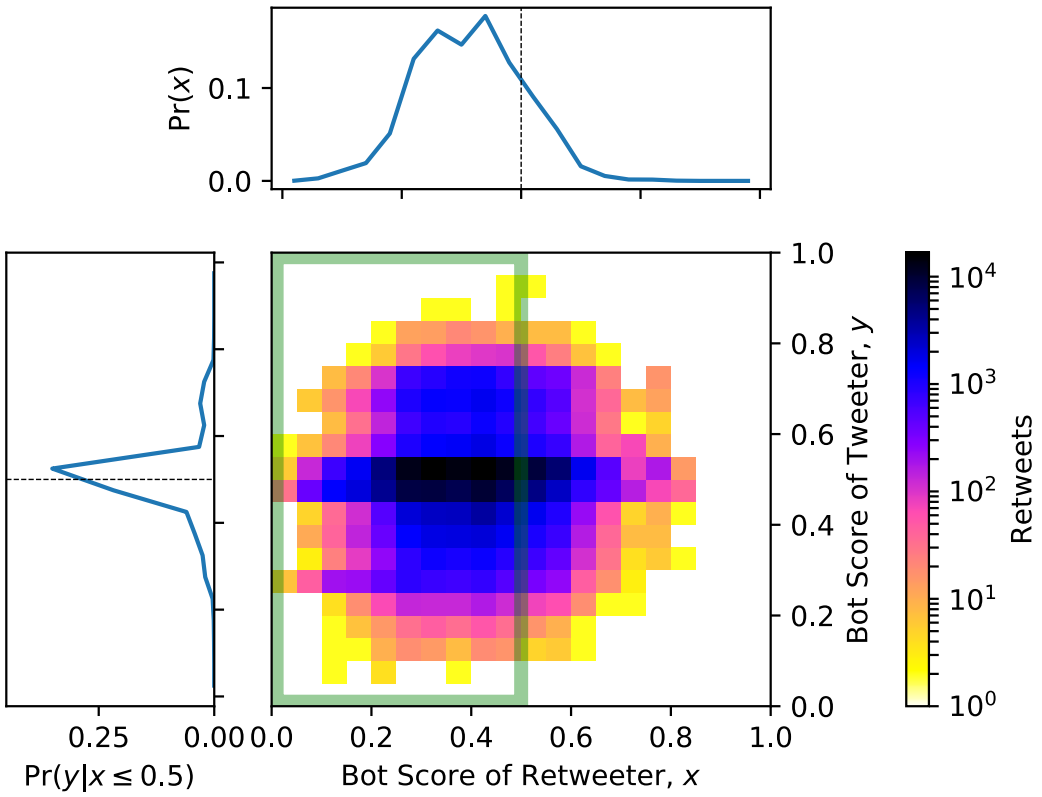

Figure S12: Joint distribution of the bot scores of accounts that retweeted links to articles from consensus low-credibility sources and accounts that had originally posted the links. We considered retweets by a sample of 20,000 accounts that posted the 1,000 most viral articles. We obtained bot scores for 12,792 tweeting accounts and 17,664 retweeting accounts, participating in 229,725 retweet pairs. Color represents the number of retweeted messages in each bin, on a log scale. Projections show the distributions of bot scores for retweeters (top) and for accounts retweeted by likely humans (left).

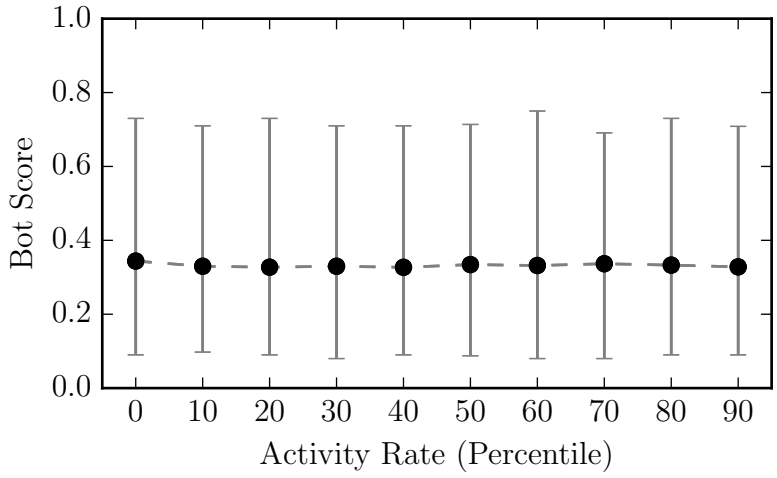

Figure S13: Distributions of bot scores versus account activity. For this analysis we randomly selected 48,517 distinct Twitter accounts evaluated by Botometer. Of these, 11,190 were available for crawling their profiles and measuring their activity (number of tweets). Bins correspond to deciles in the activity rate. We show the average and 95% confidence interval for the bot score distribution of the accounts in each activity bin. There is no correlation between activity and bot score (Pearson's $\rho = -0.007$).

confirms that account activity volume and bot scores are uncorrelated.

Bot-score threshold values

The results are also not affected by the use of different bot-score thresholds to separate social bots and human accounts. For example, we experimented with different thresholds and found that they do not change our conclusions that super-spreaders are more likely to be social bots. Figs. S7 and S8 and Table S3 show that other findings are also robust with respect to the bot score threshold, even though the estimated percentages of likely humans/bots, and the estimated numbers of tweets posted by them, are naturally sensitive to the threshold.

Bot-score calibration

Calibration is an applicable method in case a classifier outputs probabilistic scores. Well calibrated classifiers are probabilistic classifiers for which the estimates can be directly interpreted

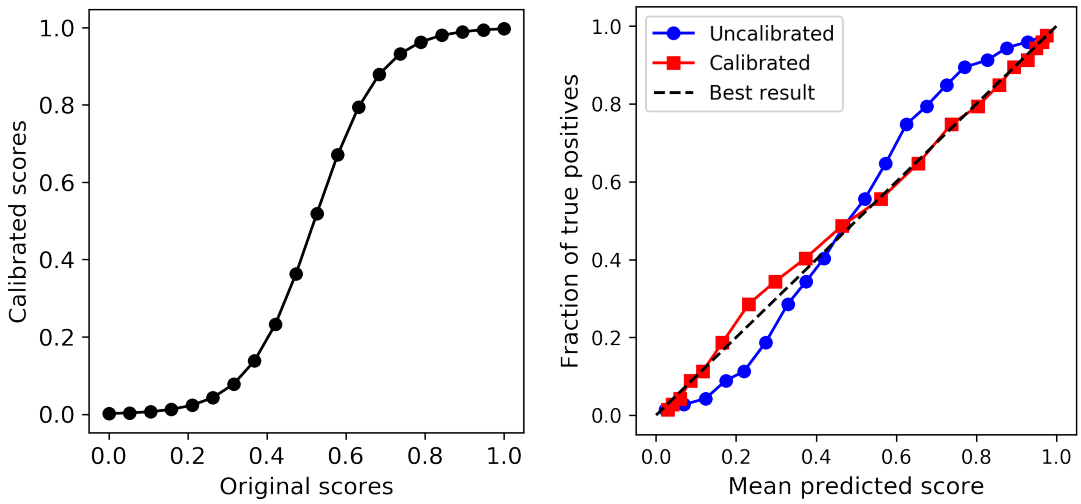

Figure S14: Bot score calibration curves. Calibration mapping functions (left) projects classifier outcomes to calibrated scores. Reliability curves (right) plots mean predicted scores against fraction of true positives.

as a confidence level. We use Platt's scaling, a logistic regression model trained on classifier outputs, to calibrate our probabilities (*62*).

In Fig. S14(a), we present mapping between classifier estimates and calibrated scores. Note that calibration only changes scores within [0,1] internal but retain ranking among the tested instances. On real world uses where the true conditional probabilities are unavailable, model calibration can be visualized with reliability diagrams (*63*). In this analysis, we split prediction space [0,1] into ten bins. Each instance in the training data set assigned to a bin based on estimated scores. For each bin, the mean predicted score is computed and compared against the true fraction of positive cases. In a well-calibrated model points will align to diagonal line as in the calibrated model (see Fig. S14(b)).

References

38. C. Shao, G. L. Ciampaglia, A. Flammini, F. Menczer, *Proceedings of the 25th International Conference Companion on World Wide Web* (2016), pp. 745–750.

39. L. Wu, F. Morstatter, X. Hu, H. Liu, *Big Data in Complex and Social Networks* (CRC Press, 2016), pp. 123–152.

40. K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, *ACM SIGKDD Explorations Newsletter* **19**, 22 (2017).

41. J. Ratkiewicz, *et al.*, *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11 (2011), pp. 249–252.

42. L. Wu, H. Liu, *Proc. 11th ACM International Conference on Web Search and Data Mining (WSDM)* (2018).

43. C. Castillo, M. Mendoza, B. Poblete, *Proceedings of the 20th International Conference on World Wide Web* (2011), p. 675.

44. A. Gupta, P. Kumaraguru, C. Castillo, P. Meier, *International Conference on Social Informatics* (2014), pp. 228–243.

45. Z. Jin, J. Cao, Y.-G. Jiang, Y. Zhang, *Proc. IEEE International Conference on Data Mining (ICDM)* (2014), pp. 230–239.

46. A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, R. Procter, *ACM Computing Surveys* **50** (2018). Forthcoming.

47. P. Resnick, S. Carton, S. Park, Y. Shen, N. Zeffer, *Proc. Computational Journalism Conference* (2014).

48. P. T. Metaxas, S. Finn, E. Mustafaraj, *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing*, CSCW'15 Companion (2015), pp. 69–72.

49. N. Hassan, *et al.*, *Proc. VLDB Endow.* **7**, 1557 (2014).

50. X. Liu, A. Nourbakhsh, Q. Li, R. Fang, S. Shah, *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15 (ACM, New York, NY, USA, 2015), pp. 1867–1870.

51. C. J. Vargo, L. Guo, M. A. Amazeen, *New Media & Society* **20**, 2028 (2018).

52. K. Starbird, *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM)* (2017), pp. 230–239.

53. S. Zannettou, *et al.*, *Proceedings of the 2017 Internet Measurement Conference*, IMC '17 (2017), pp. 405–417.

54. A. Bessi, *et al.*, *PLoS ONE* **10**, 1 (2015).

55. A. Guess, B. Nyhan, J. Reifler, Selective Exposure to Misinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign, Unpublished manuscript (2018).

56. O. Varol, E. Ferrara, F. Menczer, A. Flammini, *EPJ Data Science* **6**, 13 (2017).

57. Y. Boshmaf, I. Muslukhov, K. Beznosov, M. Ripeanu, *Proceedings of the 27th annual ACM computer security applications conference* (2011), pp. 93–102.

58. Z. Chu, S. Gianvecchio, H. Wang, S. Jajodia, *Proceedings of the 26th annual ACM computer security applications conference* (2010), pp. 21–30.

59. C. A. Davis, O. Varol, E. Ferrara, A. Flammini, F. Menczer, *Proceedings of the 25th International Conference Companion on World Wide Web* (2016), pp. 273–274.

60. S. Webb, J. Caverlee, C. Pu, *Proc. CEAS* (2008).

61. K. Lee, J. Caverlee, S. Webb, *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (2010), pp. 435–442.

62. A. Niculescu-Mizil, R. Caruana, *Proceedings of the 22nd international conference on Machine learning* (ACM, 2005), pp. 625–632.

63. M. H. DeGroot, S. E. Fienberg, *The statistician* pp. 12–22 (1983).