Supplementary Materials for

The spread of true and false news online

Soroush Vosoughi, Deb Roy, Sinan Aral*

*Corresponding author. Email: sinan@mit.edu

Published 9 March 2018, *Science* **359**, 1146 (2018)
DOI: 10.1126/science.aap9559

This PDF file includes:

Contents

S1 Definitions and Terminology

S1.1 True News, False News, Rumors and Rumor Cascades

Some work develops theoretical models of rumor diffusion [37, 38, 39, 40], or methods for rumor detection [41, 42, 43, 44], credibility evaluation [45] or interventions to curtail the spread of rumors [46, 47, 48]. But, almost no studies comprehensively evaluate differences in the spread of truth and falsity across topics or examine why false news may spread differently than the truth. For example, while Bessi et al [49, 50] study the spread of scientific and conspiracy-theory stories, they do not evaluate their veracity. We therefore focus our analysis on veracity and stories that have been verified as true or false.

We also purposefully adopt a broad definition of the term "news." Rather than defining what constitutes "news" based on the institutional source of the assertions in a story, we refer to any asserted claim made on Twitter as "news" regardless of the institutional source of that "news" (we defend this decision in the next section of the SI on "Reliable Sources").

We define "news" as any story or claim with an assertion in it and a "rumor" as the social phenomena of a news story or claim spreading or diffusing through the Twitter network. A rumor's diffusion process can be characterized as having one or more "cascades," which we define as instances of a rumor spreading pattern that exhibit an unbroken retweet chain with a common, singular origin. For example, an individual could start a rumor cascade by tweeting a story or claim with an assertion in it and another individual could independently start a second cascade of the same rumor (pertaining to the same story or claim) that is completely independent of the first cascade except that it pertains to the same story or claim. If they remain independent, then they represent two cascades of the same rumor. Cascades can be as small as size 1 (meaning no one retweeted the original tweet). The number of cascades that make up a rumor is equal to the number of times the story or claim was independently tweeted by a user (not retweeted).

S1.2 A Note on Reliable Sources and the News

Some colleagues have suggested that we classify or somehow sample "actual news," as opposed to errant rumors, by turning to what they have referred to as "reliable sources." However, after careful consideration, we have rejected this approach in favor of a broader definition of "news" and more objectively verifiable

definitions “truth” and “falsity.”

We believe the only way to robustly study “true” and “false” news is to study stories that have been verified as true or false by multiple independent fact checking organizations. Since our focus is on veracity and as we clearly argue in the main text why veracity it a key feature of interest in the spread of news, we are committed to analyzing true and false news that has been verified.

We also think that a reliance on “reliable sources” to distinguish “news” from other types of information is extremely problematic for at least two reasons. First counterclaims of (unverified) reliability are the subject of considerable disagreement in our polarized political landscape in the United States and around the world. We expand on this point below. Second, politicians are labeling news as “fake” as a political strategy and claiming that sources that don’t support them are “unreliable” while sources that do support them are “reliable,” in effect politicizing the meaning and classification of “reliable sources.”

A PEW research study [51] of American’s confidence in the media has found that the sources that “consistently conservative” and “mostly conservative” Americans find reliable or trustworthy are the exact sources that “consistently liberal” and “mostly liberal” Americans find unreliable and untrustworthy (see the Figures below). Although one person may find certain sources more reliable, chances are there are a significant number of people who see those sources as unreliable. There is simply no agreement about which sources are “reliable sources” and which are not.

Given this evidence, we do not see how a scientific study could remain objective and take a position on which sources are “reliable” and which are not. Instead, to get at the difference between true and false news, we feel it is imperative to focus on which stories (from any source) have been verified as true or false by multiple independent fact checking organizations.

To demonstrate the point further, we considered the “most reliable sources” listed in the PEW study (those that are the most trusted by the greatest number of Americans) and, for any source with at least one verified study, examined the fraction of their verified stories which were deemed true or false by the six independent fact checking organizations we worked with. This analysis revealed, first, that the most trusted sources are not necessarily the ones that record the greatest fraction of verified stories which are true; and second, that there is no correlation between the degree to which the American public finds a source “reliable” and the fraction of its verified stories which are true (see below). For these reasons, we cannot see a more reliable way determining what should be considered ”news” than adopting a broad and inclusive view of news.

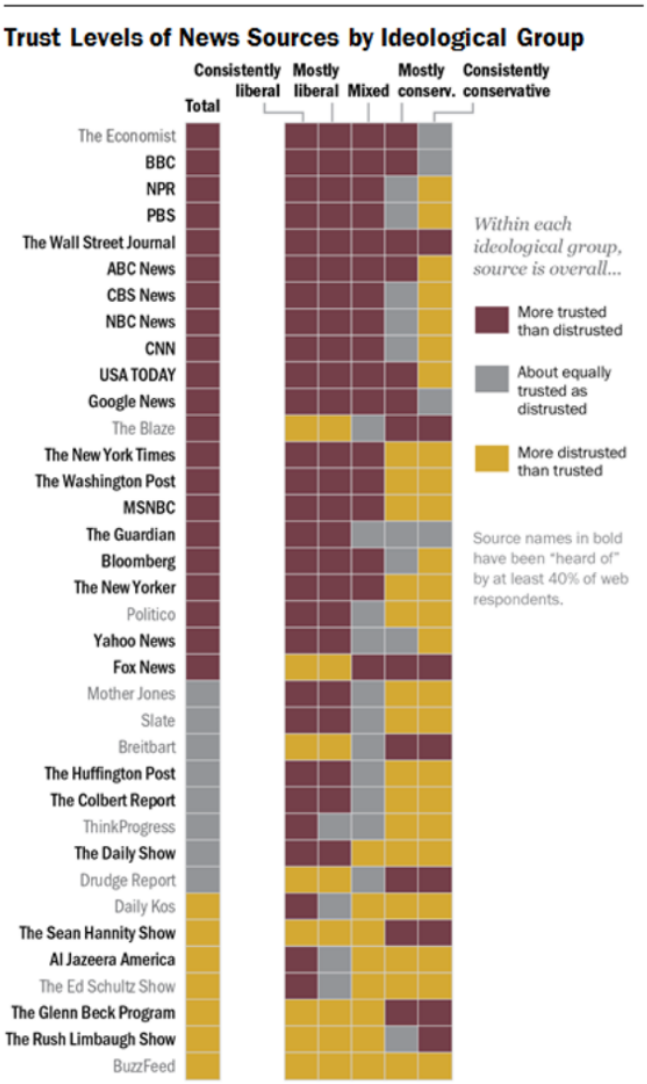

Figure S1

While one may argue that analyzing stories verified by the six independent

Trust of News Sources

% of respondents saying they trust each source

Source	Overall	Consistently liberal	Mostly liberal	Mixed	Mostly conservative	Consistently conservative
CNN	54%	56%	66%	61%	39%	14%
ABC News	50%	52%	59%	56%	40%	18%
NBC News	50%	56%	63%	54%	37%	16%
CBS News	46%	51%	55%	50%	36%	16%
Fox News	44%	6%	28%	47%	72%	88%
MSNBC	38%	52%	48%	39%	26%	7%
PBS	38%	71%	50%	31%	23%	8%
BBC	36%	69%	45%	28%	22%	12%
New York Times	34%	62%	45%	29%	17%	3%
USA TODAY	33%	29%	38%	38%	29%	14%
Wall Street Journal	31%	35%	34%	28%	32%	30%
NPR	29%	72%	36%	19%	14%	3%
Washington Post	27%	48%	33%	23%	17%	7%
Google News	25%	25%	26%	29%	20%	11%
Yahoo News	20%	17%	25%	22%	14%	10%
Huffington Post	18%	38%	21%	13%	10%	5%
Daily Show	16%	45%	21%	10%	5%	0%
Colbert Report	15%	36%	20%	10%	7%	2%
New Yorker	14%	32%	18%	11%	7%	1%
Economist	12%	30%	14%	7%	8%	4%
Sean Hannity Show	12%	0%	1%	6%	28%	62%
Rush Limbaugh Show	12%	0%	2%	6%	27%	58%
Bloomberg	11%	18%	13%	8%	8%	7%
Glenn Beck Program	10%	0%	1%	4%	24%	51%
Al Jazeera America	9%	28%	11%	3%	3%	3%
Drudge Report	8%	1%	2%	4%	15%	34%
Guardian	7%	21%	8%	4%	3%	2%
Politico	7%	21%	7%	2%	4%	5%
TheBlaze	6%	1%	0%	2%	13%	37%
Mother Jones*	6%	25%	5%	1%	0%	1%
Breitbart	4%	0%	0%	2%	8%	25%
Slate	4%	14%	4%	1%	1%	0%
Ed Schultz Show*	3%	14%	3%	0%	0%	0%
BuzzFeed	2%	6%	3%	1%	1%	0%
Daily Kos*	2%	10%	1%	0%	0%	0%
ThinkProgress*	2%	10%	2%	0%	0%	1%

Source: American Trends Panel (wave 1). Survey conducted March 19-April 29, 2014. Based on all Web respondents (representative of the 89% of Americans with internet access). (Overall N=2,901; see About the Study for sample sizes of each ideological group.) Respondents were asked which (of 36 sources for news about government and politics) they have heard of, trust, distrust and got news from in the past week. Ideological consistency based on a scale of 10 political values questions. To see audience profiles, click each source. *Note that ThinkProgress, Daily Kos, Mother Jones and The Ed Schultz Show do not have audience profiles because the sample sizes for these audiences are too small to analyze.

Figure S2

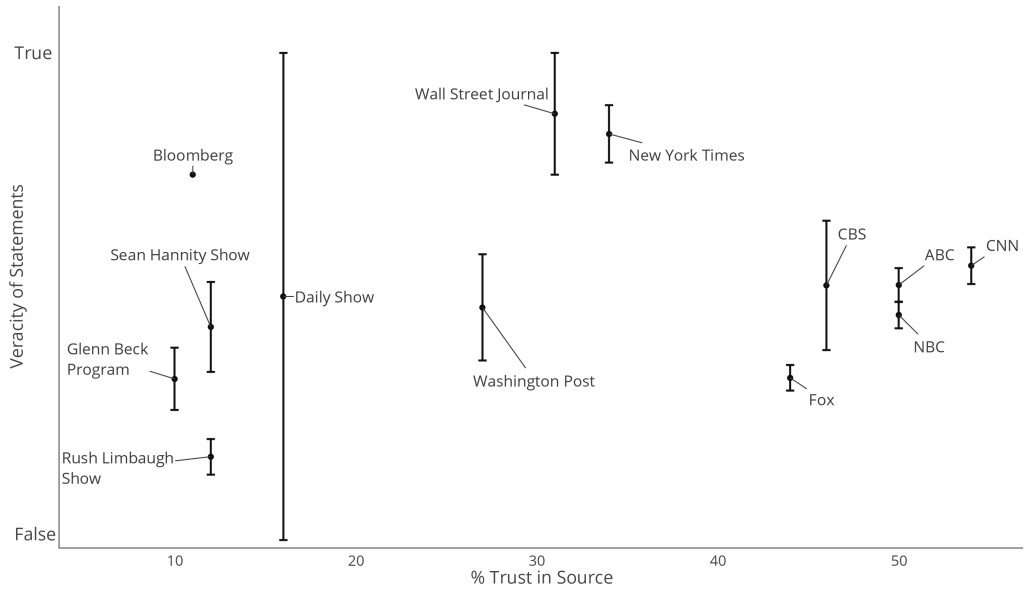

Figure S3: The percent of Americans that trust an outlet (recorded in the the PEW study for trust in media) vs the average veracity of statements investigated by the fact checking organization Politifact in our sample.

fact checking organizations may introduce its own selection bias, as we describe in the main text and expound on below, we cannot think of a more objective way to distinguish true from false content than to rely on multiple independent fact checking organizations. Furthermore, it is for this reason that we analyze a second set of news stories that were never fact checked by any of the original fact checking organizations, but that instead were fact checked by three independent fact checkers that we recruited to verify a robustness sample of approximately 13,000 rumor cascades independently (see sections S8.1 and S8.2). We feel this addresses the potential selection bias introduced by our reliance on the six independent fact checking organizations in our main analysis and makes ours the most rigorous approach to defining truth and falsehood, without wading into a debate about which institutional sources are reliable and which are not. While our approach is certainly not the only way to analyze the diffusion of true and false news, we encourage future research to also clearly define the terms used in analyses to enable comparability across disparate studies.

S2 Data

A rumor cascade on Twitter starts with a user making an assertion about a topic (this could be text, a photo or a link to an article); people then propagate the news by retweeting it. In many cases, people also reply to the original tweet. These replies sometimes contain links to fact checking organizations that either confirm or debunk the rumor in the origin tweet. We used such cascades to identify rumors that are propagating on Twitter. We explain the rumor detection, classification and collection methods in detail below.

S2.1 Rumor Dataset

We identified six fact checking organizations well-known for thoroughly investigating and debunking or confirming rumors. The websites for these organizations are as follows: snopes.com, politifact.com, factcheck.org, truthorfiction.com, hoax-slayer.com, and urbanlegends.about.com. We automatically scraped these websites, collected the archived rumors and parsed the title, body and verdict of each rumor. These organizations have various ways of issuing a verdict on a rumor, for instance snopes articles are given a verdict of "False", "Mostly False", "Mixture", "Mostly True" and "True"; while politifact articles are given a "Pants on Fire" rating for false rumors. We normalized the verdicts across the different sites by mapping them to a score of 1 to 5 (1="False", 2="Mostly False", 3="Mixture", 4="Mostly True", 5="True"). For our analysis, we grouped all rumors with a score of 1 or 2 as false, those with a score of 4 or 5 as true and the ones with score of 3 as mixed or undetermined. Mixed rumors are those that are either a mixture of false and true; all fact checking organizations we looked at have a few categories that fall under this label.

It is not uncommon for a rumor to be investigated by multiple organizations. We can use these cases to measure the agreement between various fact checking proceedures across organizations. Table S1 shows the agreement between various organizations' verdicts. Note that all cases of disagreement were between "mixture" and "mostly true" (scores 3 and 4) or "mixture" and "mostly false" (scores 3 and 2). We did not observe any disagreement between the organizations' verdicts for rumors that were "false" or "true." In cases where we saw disagreements, we assigned the veracity score based on the majority verdict.

	snopes	politifact	factcheck	truthorfiction	hoax-slayer
politifact	96%				
factcheck	98%	97%			
truthorfiction	95%	95%	96%		
hoax-slayer	96%	95%	95%	97%	
urbanlegends	95%	95%	95%	96%	97%

Table S1: Agreement between various rumor debunking websites.

S2.1.1 Rumor Topics

Most of the aforementioned rumor debunking organizations (henceforth referred to as trusted organizations) already tag rumors with a topic (e.g., politics, terrorism, science, urban legends). Using these classifications, we divided the rumors into seven overarching topics: Politics, Urban Legends, Business, Science and Technology, Terrorism and War, Entertainment, and Natural Disasters. For rumors that did not have a topic tag, or had multiple or uncertain tags, we asked three annotators (political science undergraduates at MIT and Wellesley) to label them using one of the seven topics. We showed the annotators several example rumors from each of the categories and explained the topic hierarchy for classification (for instances where a rumor might fall under more than one category). We labeled the rumor based on the majority label. The Fleiss' kappa (κ) for the annotators was 0.93 (Fleiss' kappa is a statistical measure of the reliability agreement between annotators [52]). Table S2 shows the agreement amongst the annotators. For 91% of the rumors there was agreement amongst all three annotators, the remaining 9% had agreement between two out of three annotators. There were no rumors for which there was no agreement amongst at least two of the annotators.

	Annotator 1	Annotator 2
Annotator 2	97%	
Annotator 3	92%	93%

Table S2: Agreement between annotators on rumor topics. $\kappa = 0.93$

S2.2 Twitter Data

We used our access to the full Twitter historical archives (which gives us access to all tweets ever posted, going back to the first tweet) to collect all English-language tweets that contained a link to any of the websites of the trusted fact checking organizations, from September 2006 to December 2016. There were 500K tweets containing a link to these websites and we were interested in tweets containing these links that were replies to other tweets. For each reply tweet, we extracted the original tweet that they were replying to and then extracted all the retweets of the original tweet. Each of these retweet cascades is a rumor propagating on Twitter. We also know the veracity of each cascade, through the reply that linked to one of the rumor investigating sites.

We took extreme care to make sure that the replies containing a link to any of the trusted websites were in fact addressing the original tweet. We did this through a combination of automatic and manual measures. First, we only considered replies that were directly targeting the original tweet, in other words, we did not consider replies to replies, only replies to the original tweet. Second, we compared the headline of the linked article to that of the original tweet. We also removed all original tweets that were directly linking to one of the fact-checking websites as we wanted to study how unverified and contested information spreads, and tweets linking to one of the fact-checking websites do not qualify as they are no longer unverified. Around 158K cascades passed this stage.

We then used ParagraphVec [53] and Tweet2Vec [54] algorithms to convert the headline and the original tweet respectively to vectors that capture their semantic content. We then used cosine similarity to measure the distance between the vectors (we note that some tweets had images with text on them, therefore, we used an OCR algorithm[1] to extract the text from the images.) If the similarity was lower than .5 the tweet was discarded, if it was higher than .5, but lower than .9, it was manually inspected, if it was higher than .9 it was assumed to be correct. We removed 10,331 cascades from our dataset through this process.

S2.2.1 Canonicalization

Once we had identified the rumor cascades that had been debunked/confirmed through replies, we canonicalized them by identifying images and links to external articles in the original tweets (root of the cascades). Images on Twitter also have a url, however, there could be hundreds of different links for a given photo.

[1] https://ocr.space/

Therefore, we passed each image to Google's reverse image search to identify all links that point to that image. Moreover, as mentioned earlier, we employed OCR to identify the text in the images. Next, using the Twitter historical API, which has full url and text search capabilities, we extracted all English-language original tweets containing any of these urls (photos and external articles) or text, from September 2006 to December 2016. Finally, we extracted all the retweets to these tweets.

S2.2.2 Removing bots

As a last step, we used a state-of-the-art bot detection algorithm by Varol et al. [55] to remove all accounts that were identified as bots.[2] 13.2% of the accounts were identified as bots and were removed. Our bot analysis is explained in greater detail in section S8.3 below.

S2.2.3 Approach to Tweet Deletion

As shown in previous work, tweet deletion may impact the results of rumor studies on Twitter [56]. We therefore included all tweets that were made available to us by the full Twitter historical archives. Since our data is anonymized and since we have a direct relationship with Twitter, we can continue to include in our analysis any tweet that was deleted after we received our data, which means our analysis is less prone to errors from tweet deletions than other studies of rumor cascades on Twitter.

S2.3 Dataset Summary

After all of the data processing, we were left with 126,285 rumor cascades corresponding to 2,448 rumors. Of the 126,301 cascades, 82,605 were false, 24,409 were true and 19,287 were mixed, corresponding to 1,699 false, 490 true and 259 mixed rumors. The earliest rumor cascades that we were able to identify were from early October 2008 and the latest cascades were from late December 2016. Figure 1b in the main text shows the complimentary cumulative distribution function (CCDF) (see section S10 for an explanation of the CCDF) of the number of cascades for false, true and mixed rumors (Figure 1d shows this for political rumors). Figure 1c in the main text shows the number of false, true and mixed

[2]The bot detection API can be found here: https://truthy.indiana.edu/botornot

rumor cascades over time, from mid 2008 to the end of 2016 (Figure 1e shows this for political rumors). Figure 1f in the main text shows the number of cascades by topic.

We are aware that there may be a selection bias in the collection of our dataset as we only consider rumors that were eventually investigated by a fact-checking organization. To address this issue, we include a robustness check by looking at human-identified stories (described later in this document). It may also be that there is a bias towards stories that are of greater diffusion volume, even in the robustness dataset. However, we argue this implies we are studying rumors/stories that have at least a visible footprint on Twitter (i.e., they have been picked up/shared by enough people to have an impact). So, while our robustness dataset may under-sample stories that never diffused, our main sample is representative of verified stories and our robustness sample is representative of stories with a visible footprint on Twitter.

S3 Quantifying and Comparing Rumor Cascades

S3.1 Time-Inferred Diffusion of Rumor Cascades

Each of the retweet cascades described in section S2.2, corresponds to a rumor cascade. The root of the cascade is the original tweet containing a rumor. All other nodes in the cascade correspond to retweets of the original tweet. Since each tweet and retweet is labeled with a timestamp, one can track the temporal diffusion of messages on Twitter. However, the Twitter API does not provide the true retweet path of a tweet. Figure S4a shows the retweet tree that the Twitter API provides. As you can see, all retweets point to the original tweet. This does not capture the true retweet tree since in many cases a user retweets another user's retweet, and not the original tweet. But as you can see in Figure S4a, all credit is given to the user that tweeted the original tweet, no matter who retweeted whom.

Fortunately, we can infer the true retweet path of a tweet by using Twitter's follower graph. Figure S5 shows how this is achieved. The left panel in the figure shows the retweet path provided by Twitter's API. The middle panel shows that the bottom user is a follower of the middle user but not of the top user (the user who tweeted the original tweet). Finally, the right panel shows that using this information, and the fact that the bottom user retweeted after the middle user, it can be inferred that the bottom user retweeted the middle user and not the top

user. If the bottom user was a follower of the top user, then the original diffusion pattern shown in the left panel would stand (i.e., it would have been inferred that both the middle and bottom users were retweeting the top user). This method of reconstructing the true retweet graph, called time-inferred diffusion, is based on work by Goel et al. [57] and is used to establish true retweet cascades in a broad range of academic studies of Twitter. Using this method, we convert our example retweet cascade shown in Figure S1a to a more accurate representation of the retweet cascade, shown in Figure S4b. The cascade shown in Figure S4b, is what we use to analyze the rumor cascades.

The follower-followee information is inferred at the time of the retweet. Since the Twitter API returns follower-followee information in reverse chronological order, combined with knowledge of when the users joined Twitter, one can probabilistically infer the followership network of a user in the past. For instance, if user U_0 is followed by users U_1, U_2, and U_3 (in this order through time) and users U_1, U_2, and U_3 joined Twitter on dates D_1, D_2, and D_3, then we can know for certain that U_2 was not following U_0 before D_1 and U_2 was not following U_0 before $min(D_1, D_2)$.

Note that we do not include quotes or replies in our propagation dynamics. This is because, generally speaking, retweets (not quotes) do not contain additional information and represent people agreeing with what is being shared. We do not include replies in our propagation analysis as we don't know if the replies are agreeing or disagreeing with the rumor. But we do analyze replies in other ways (as shown in Figure 4d and 4e in the main text).

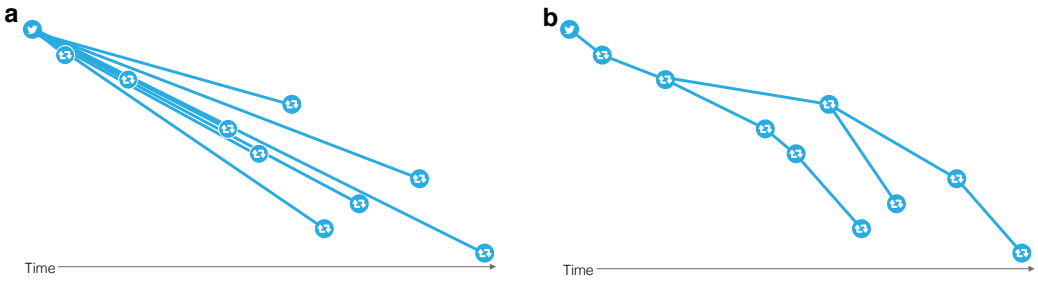

Figure S4: A sample rumor cascade. Each node represents a user and the x-axis is time. The Twitter symbol on the top left represents an original tweet and the arrows represent retweets. The tree on the left shows the retweet cascade from the Twitter API, the tree on the right shows the true cascade created using time-inferred diffusion.

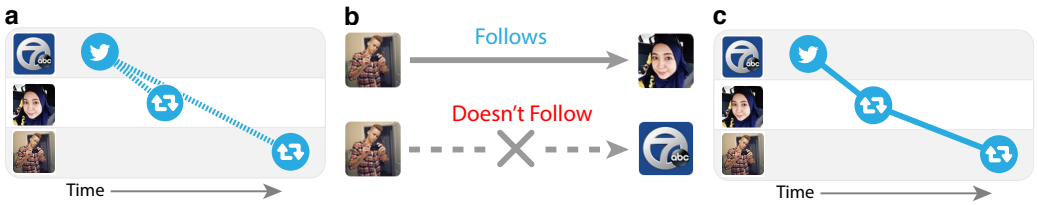

Figure S5: Using Twitter's follower graph to infer the correct retweet path of a tweet. Panel (a) shows the retweet path provided by the Twitter API. Panel (b) shows that the bottom user is a follower of the middle user but not that of the top user (the user who tweeted the original tweet). Panel (c) shows that using this information, and the fact that the bottom user retweeted after the middle user, we can infer that the bottom person retweeted the middle person and not the top person.

S3.2 Characteristics of Rumor Cascades

S3.2.1 Static Measures

We measured and compared four static characteristics of false, true and mixed rumor cascades: depth, max-breadth, structural virality [23], and size (since on Twitter a person can only retweet a tweet once, the size of a cascade corresponds to the number of unique users involved in that cascade). Here we define each of these measures.

Take an example rumor cascade shown in Figure S6a. The static measures are not dependent on time, therefore, we can reorganize the cascade based on depth, as seen in Figure S6b. Using this example, the definition of each of the four static measures is described below.

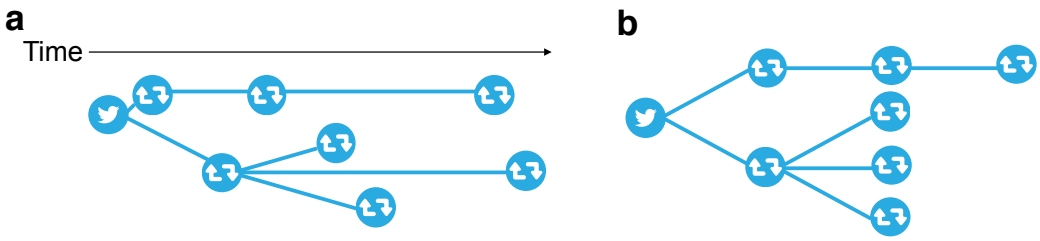

Figure S6: An example rumor cascade.

- Depth: The depth of a node is the number of edges from the node to the

root node (in this case, the original tweet). The depth of a cascade is the maximum depth of the nodes in the cascade. In other words, the depth of a cascade, D, with n nodes is defined as:

$$D = max(d_i), 0 <= i <= n \tag{S1}$$

Where d_i denotes the depth of node i. Figure S7a shows the depth measurement for our example cascade. In this case the depth of the cascade is 3.

- Size/Unique Users: The size of a cascade corresponds to the number of users in that cascade. As explained earlier, the size of a cascade corresponds to the number of unique users involved in that cascade since users can only retweet something once on Twitter. Figure S7b shows the size our example cascade at different depths; the size of the full cascade is 8.

- Structural Virality: The structural virality of a cascade, as defined by Goel et al. (2015), is the average distance between all pairs of nodes in a cascade. For a cascade with n>1 nodes, the virality v is defined as:

$$v = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} \tag{S2}$$

Where d_{ij} denotes the length of the shortest path between nodes i and j. Figure S7c shows the structural virality of our example cascade at different depths. The structural virality of the full cascade is 2.54.

- Max-Breadth: The breadth of a cascade is a function of its depth. At each depth, the breadth of a cascade is the number of nodes at that depth. As the name suggests, the max-breadth of a cascade is its maximum breadth. For a cascade with depth d, the max-breadth, B, is defined as:

$$B = max(b_i), 0 <= i <= d \tag{S3}$$

Where b_i denotes the breadth of a cascade at depth i. Figure S7d shows the breadth of our example cascade at each depth. The max-breadth of this cascade is thus 4.

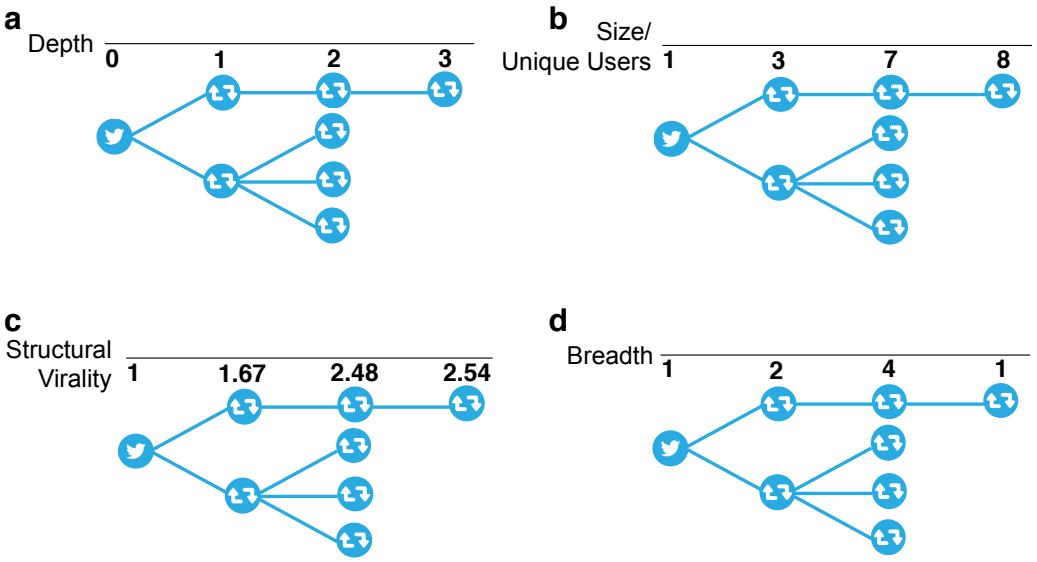

Figure S7: Depth, size, structural virality and breadth calculated for a sample cascade.

These four static measures were calculated for all cascades in our dataset. Figures 2a, 2b, 2c and 2d in the main text show the CCDF (see section S10 for an explanation of the CCDF) of these measurements for false and true cascades. Below we show the breakdown of these statistics for each veracity. Note that all standard error (SE) values correspond to cluster-robust standard errors [58, 59], clustered on rumors (i.e., cascades belonging to the same rumor are clustered together). (An explanation of cluster-robust standard errors is provided in section S9.) Since most of these measures have a heavy-tailed distribution, we logged our measurements. Tables S3, S4, S5 and S6 below show the mean (log), the cluster-robust standard errors (log) and the min and max of the depth, max-breadth, structural virality, and size for false, true and mixed cascades (the tables show the mean and SE of the logged data). For each of these measurements, we also ran two-sample Kolmogorov-Smirnov (KS) tests to compare the distributions of these measures between false and true cascades. The results of the tests are reported in the table legends.

	N	Mean (log)	Robust-SE (log)	Min	Max
False	82,605	0.156	0.0135	0	24
True	24,409	0.099	0.0183	0	12
Mixed	19,287	0.083	0.0171	0	19

Table S3: Statistics on the depth of cascades. KS-test for false and true cascades: $D = 0.134$, $p \sim 0.0$

	N	Mean (log)	Robust-SE (log)	Min	Max
False	82,605	0.289	0.0286	1	29,527
True	24,409	0.172	0.0401	1	1,559
Mixed	19,287	0.148	0.0331	1	11,783

Table S4: Statistics on the max-breadth of cascades. KS-test for false and true cascades: $D = 0.134$, $p \sim 0.0$

	N	Mean (log)	Robust-SE (log)	Min	Max
False	31,858	0.188	0.0074	1.0	10.25
True	6,149	0.164	0.0196	1.0	5.72
Mixed	4,074	0.164	0.0167	1.0	10.07

Table S5: Statistics on the structural virality of cascades. KS-test for false and true cascades: $D = 0.107$, $p \sim 0.0$

	N	Mean (log)	Robust-SE (log)	Min	Max
False	82,605	0.313	0.0305	1	46,895
True	24,409	0.186	0.0431	1	1,649
Mixed	19,287	0.160	0.0364	1	23,228

Table S6: Statistics on the size of cascades. KS-test for false and true cascades: $D = 0.134$, $p \sim 0.0$

S3.2.2 Dynamic Measures

We also measured four dynamic characteristics of the cascades: time vs depth, time vs unique users (or size), max-breadth vs depth, and unique users vs depth. As with the static measurements, here we also took the log of the data (for similar reasons) and calculated all standard errors using a cluster-robust method (again, clustering on the rumors). Below is a description of how each dynamic character-

istic was calculated. Two of the dynamic measures are a function of time (depth over time and unique users over time), while the other two are a function of depth (breadth vs depth and unique users vs depth). Note that the figures depicting the dynamic measures in the main text (Figures 2e-f and 3e-f) depict the geometric mean; this is because we log the values to transform the data to a non-heavy-tailed distribution to avoid infinite variance. To make the visualization more human-readable, we then take the exponent of the mean and the SEM of the data, which is the same as the geometric mean of the actual data.

- Depth over Time: For each cascade (e.g., the one shown in Figure S6a), we measured the time it took to reach each depth in minutes. Figure S8a shows depth over time being calculated for our sample cascade. For each depth, we averaged these times across false, true and mixed cascades, producing an average time (and standard error) for cascades to reach different depths, as shown in Figure 2e in the main text.

- Unique Users over Time: For each cascade, we measured the time it took to reach a certain number of unique users (which corresponds to cascade size) in minutes. Figure S8b shows unique users over time being calculated for our sample cascade. For each value, we averaged these times across all false, true and mixed cascades, producing an average time (and standard error) for cascades to reach different numbers of unique users, as shown in Figure 2f in the main text.

- Breadth vs Depth: For each cascade, we measured the breadth at every depth. Figure S9a shows breadth vs depth being calculated for our sample cascade. For each depth, we averaged these values across all false, true and mixed cascades, producing an average breadth (and standard error) for each depth, as shown in Figure 2h in the main article.

- Unique Users vs Depth: For each cascade, we measured the number of unique users at every depth. Figure S9b shows unique users vs depth being calculated for our sample cascade. For each depth, we averaged these values across false, true and mixed cascades, producing an average number of unique users (and standard error) for each depth, as shown in Figure 2g in the main article.

Overall, we found statistically significant differences between false and true cascades across all of our measures. In brief, false cascades tend to be more viral

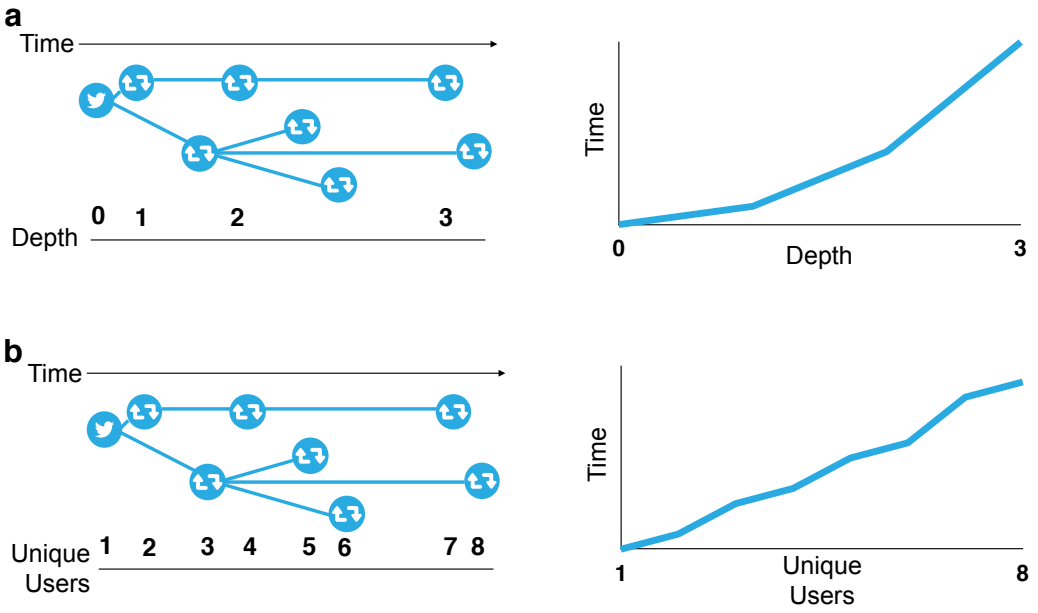

Figure S8: Time based dynamic characteristics: time vs depth and time vs unique users, calculated for a sample cascade.

and spread farther, faster, deeper and more broadly than the truth for all categories of information.

S4 Rumor Topics

As explained in section S2.1.1, each rumor is tagged with one of seven topics. As much of the discourse on false news has been focused on politics, we were interested in the differences between political and nonpolitical rumors in particular as well as the differences in diffusion dynamics across all categories more broadly. We began by combining all nonpolitical rumors into one category in order to compare political rumors to all other types of rumors. Of the 126,285 rumor cascades, 44,095 were political (27,600 false, 9,520 true, and 6,975 mixed) and 82,206 were nonpolitical (55,005 false, 14,889 true, and 12,312 mixed).

We ran the same analysis explained in section S3.2 on the political and nonpolitical rumor cascades. Figure S10 shows the results. This is the same as Figure 3 in the main article, with the addition of the results for true rumor cascades. Tables S7, S8, S9 and S10 below show the mean (log), the cluster-robust standard errors

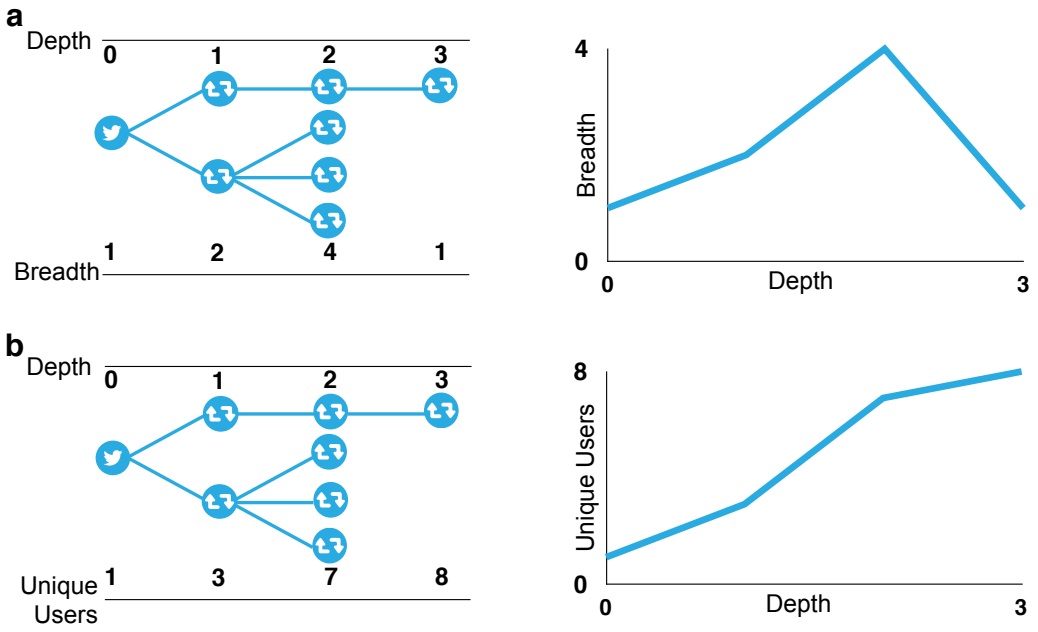

Figure S9: Depth based dynamic characteristics: breadth vs depth and unique users vs depth, calculated for a sample cascade.

(log), min and max of the depth, max-breadth, structural virality, and size for false, true and mixed political and nonpolitical cascades (the tables show the mean and SE of the logged data). We also ran two-sample Kolmogorov-Smirnov (KS) tests to compare the distributions between political and non-political cascades. The results of these tests are reported in the table legends.

Overall, we found statistically significant differences between false and true cascades within and between each rumor type across all of our measures. Political cascades (both true and false) tend to be more viral, spread faster and deeper and reach more unique users than nonpolitical rumors.

We also disaggregated the false rumor cascades and compared all 7 categories. Alhough the data are noisier when we disaggregate the categories (as is expected since the number of data points in each category is smaller), it is still clear that "Politics" spreads farther, faster, deeper and more broadly on most measures. Tables S11, S12, S13, and S14 below show the means (log), cluster-robust standard errors (log), min and max of the depth, max-breadth, structural virality, and size of false cascades across the different categories (the tables show the mean and SE of the logged data). We also ran two-sample Kolmogorov-Smirnov (KS) tests to

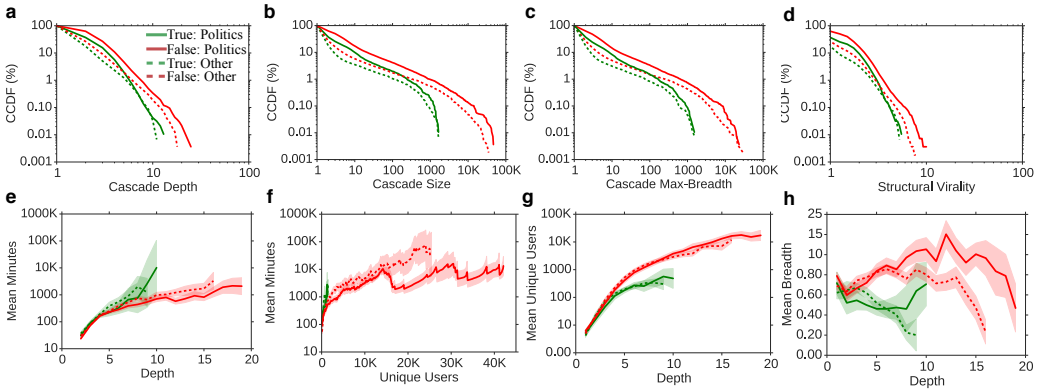

Figure S10: Difference between political and non-political rumor cascades.

	N	Mean (log)	Robust-SE (log)	Min	Max
False-political	27,600	0.258	0.0166	0	24
True-political	9,520	0.149	0.0525	0	12
Mixed-political	6,975	0.093	0.0219	0	9
False-nonpolitical	55,005	0.104	0.0094	0	17
True-nonpolitical	14,889	0.067	0.0118	0	10
Mixed-nonpolitical	12,312	0.078	0.0242	0	19

Table S7: Statistics of the depth of political and nonpolitical cascades. KS for false cascades: $D = 0.362, p \sim 0.0$. KS for true cascades: $D = 0.194, p \sim 0.0$

	N	Mean (log)	Robust-SE (log)	Min	Max
False-political	27,600	0.498	0.0380	1	23,243
True-political	9,520	0.269	0.1214	1	1,521
Mixed-political	6,975	0.175	0.0642	1	4,971
False-nonpolitical	55,005	0.185	0.0200	1	29,527
True-nonpolitical	14,889	0.111	0.0247	1	1,559
Mixed-nonpolitical	12,312	0.132	0.0406	1	11,783

Table S8: Statistics of the max-breadth of political and nonpolitical cascades. KS for false cascades: $D = 0.362, p \sim 0.0$. KS for true cascades: $D = 0.194, p \sim 0.0$

compare the distributions between political and all other topics. The results of these tests are reported in the captions of the tables. As can be seen in the tables, false political rumors significantly "outperform" false rumors of other categories

	N	Mean (log)	Robust-SE (log)	Min	Max
False-political	17,295	0.212	0.0072	1.0	10.25
True-political	3,524	0.188	0.0340	1.0	5.72
Mixed-political	1,651	0.167	0.0362	1.0	4.93
False-nonpolitical	14,563	0.159	0.0114	1.0	7.86
True-nonpolitical	2,625	0.132	0.0187	1.0	5.40
Mixed-nonpolitical	2,423	0.163	0.0147	1.0	10.07

Table S9: Statistics of the virality of political and nonpolitical cascades. KS for false cascades: $D = 0.194, p \sim 0.0$. KS for true cascades: $D = 0.218, p \sim 0.0$

	N	Mean (log)	Robust-SE (log)	Min	Max
False-political	27,584	0.540	0.0392	1	46,895
True-political	9,520	0.292	0.1304	1	1,649
Mixed-political	6,975	0.188	0.0687	1	5,075
False-nonpolitical	55,005	0.199	0.0216	1	35,016
True-nonpolitical	14,889	0.119	0.0269	1	1,647
Mixed-nonpolitical	12,312	0.144	0.0451	1	23,228

Table S10: Statistics of the size of political and nonpolitical cascades. KS for false cascades: $D = 0.362, p \sim 0.0$. KS for true cascades: $D = 0.194, p \sim 0.0$

in terms of the speed, breadth, and depth of their diffusion. Figure S11 shows the diffusion measurements for each topical category seperately. The results are quite interesting. For example, news about politics, urban legends and entertainment spread to the most people, while rumors about politics and urban legends spread the fastest and are the most viral (in terms of their structural virality).

S5 Characteristics of Users

Next, we analyzed the characteristics of users involved in spreading rumors to see if there were differences across the characteristics of users involved in spreading true and false rumors that could explain differences in the spread of these rumors. For each user, we looked at five factors that could be extracted from the Twitter API:

- Followers: Number of people who follow the user on Twitter.

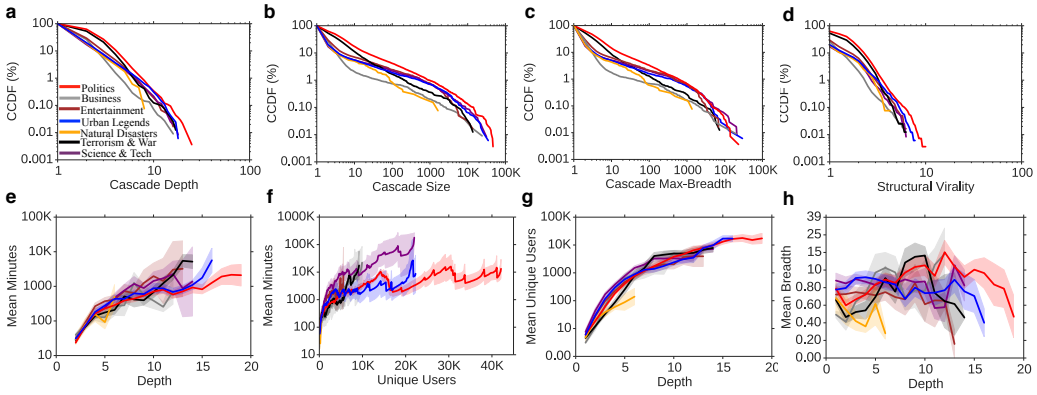

Figure S11: Difference between false rumor cascades of different topics.

	N	Mean (log)	Robust-SE (log)	Min	Max
Politics	27,600	0.258	0.0166	0	24
Urban Legends	16,458	0.083	0.0091	0	17
Science & Tech	12,043	0.071	0.0111	0	17
Business	11,086	0.092	0.0087	0	15
Terrorism & War	8,054	0.211	0.0282	0	16
Entertainment	6,046	0.116	0.0153	0	15
Natural Disasters	1,318	0.074	0.0276	0	7

Table S11: Statistics of the depth of false cascades across different categories. Politics vs Terrorism & War: $D = 0.094, p \sim 0.0$; Politics vs Science & Tech: $D = 0.456, p \sim 0.0$; Politics vs Urban Legends: $D = 0.427, p \sim 0.0$; Politics vs Entertainment: $D = 0.333, p \sim 0.0$; Politics vs Business: $D = 0.363, p \sim 0.0$; Politics vs Natural Disasters: $D = 0.439, p \sim 0.0$

- Followees: Number of people who the user follows on Twitter.
- Verified: Whether the user's account has been officially verified by Twitter.[3]
- Account age: The age of the user's account, measured in days.
- Engagement: This measures how active a user has been on Twitter since

[3]https://support.twitter.com/articles/119135

	N	Mean (log)	Robust-SE (log)	Min	Max
Politics	27,600	0.498	0.0380	1	23,243
Urban Legends	16,458	0.156	0.0237	1	29,527
Science & Tech	12,043	0.151	0.0311	1	20,998
Business	11,086	0.130	0.0146	1	20,147
Terrorism & War	8,054	0.355	0.0653	1	7,296
Entertainment	6,046	0.214	0.0411	1	6,829
Natural Disasters	1,318	0.129	0.0671	1	1,363

Table S12: Statistics of the max-breadth of false cascades across different categories. Politics vs Terrorism & War: $D = 0.120, p \sim 0.0$; Politics vs Science & Tech: $D = 0.456, p \sim 0.0$; Politics vs Urban Legends: $D = 0.427, p \sim 0.0$; Politics vs Entertainment: $D = 0.333, p \sim 0.0$; Politics vs Business: $D = 0.375, p \sim 0.0$; Politics vs Natural Disasters: $D = 0.439, p \sim 0.0$

	N	Mean (log)	Robust-SE (log)	Min	Max
Politics	17,295	0.212	0.0071	1.0	10.25
Urban Legends	3,281	0.175	0.0148	1.0	7.86
Science & Tech	2,056	0.178	0.0181	1.0	6.28
Business	2,919	0.094	0.0093	1.0	5.95
Terrorism & War	8,054	0.185	0.0131	1.0	6.28
Entertainment	1,774	0.153	0.0167	1.0	6.47
Natural Disasters	247	0.155	0.0538	1.0	4.44

Table S13: Statistics of the virality of false cascades across different categories. Politics vs Terrorism & War: $D = 0.082, p \sim 0.0$; Politics vs Science & Tech: $D = 0.171, p \sim 0.0$; Politics vs Urban Legends: $D = 0.181, p \sim 0.0$; Politics vs Entertainment: $D = 0.229, p \sim 0.0$; Politics vs Business: $D = 0.381, p \sim 0.0$; Politics vs Natural Disasters: $D = 0.240, p \sim 0.0$

joining. The Engagement, E, is calculated by the following equation:

$$E = \frac{T + R + P + F}{A} \tag{S4}$$

Where T, R, P, and F denote the number of tweets, retweets, replies and favorites by the user, respectively and A denotes the user account's age in days.

	N	Mean (log)	Robust-SE (log)	Min	Max
Politics	27,600	0.540	0.0392	1	46,895
Urban Legends	16,458	0.169	0.0257	1	34,919
Science & Tech	12,043	0.160	0.0330	1	31,342
Business	11,086	0.139	0.0161	1	23,719
Terrorism & War	8,054	0.387	0.0681	1	13,727
Entertainment	6,046	0.229	0.0438	1	13,574
Natural Disasters	1,318	0.137	0.0729	1	1,631

Table S14: Statistics of the size of false cascades across different categories. Politics vs Terrorism & War: $D = 0.119, p \sim 0.0$; Politics vs Science & Tech: $D = 0.454, p \sim 0.0$; Politics vs Urban Legends: $D = 0.427, p \sim 0.0$; Politics vs Entertainment: $D = 0.333, p \sim 0.0$; Politics vs Business: $D = 0.384, p \sim 0.0$; Politics vs Natural Disasters: $D = 0.439, p \sim 0.0$

We analyzed the difference between these factors for users involved in false and true cascades. Overall, there were 3,525,344 users (2,725,269 unique) involved in the false cascades, 202,348 users (170,918 unique) involved in the true cascades and 307,043 users (214,797 unique) involved in the mixed cascades (total of 3,092,984 unique users across all veracities). Figure 4a in the main article shows the breakdown of the user characteristics for false and true rumors. Even though there are slight differences between the characteristics of users involved in spreading false and true rumors, the differences should in fact favor the spread of true rumors (by favor here we mean these differences should drive greater virality, speed, depth and reach for true rumors), as the users spreading those have in general more followers, are more likely to be verified and are more active on Twitter. Therefore, these factors cannot be driving the differences we observe between true and false rumor cascades (we expand on this in the next section).

S5.1 Analysis of Rumor-Starters

We also passed the user accounts of the rumor starters to a Twitter demographic classifier [60], to infer the gender of the users who start false and true rumors. We found no difference between the gender distribution of false and true rumor sharers (ks-test: D=0.0182, p= 0.320), with false rumor starters being 56% male, 25% female and 19% other (organizations, websites, etc). True rumor starters were 55% male, 27% female and 18% other.

We also looked at the unique number of users responsible for starting false and true rumors. Overall, the 82,605 false cascades were started by 63,293 unique users and the 24,409 true cascades were started by 20,266 unique users. Figure S12 below shows the CCDF of the number of false and true rumors started by unique users. Note that the distribution is very heavy-tailed. As can be seen from the figure, a great majority of users (94%) start only a single rumor (both false and true), and less than 1% of users start more than 10 rumors. The most false cascades started by a user was 4,717 and the most number of true cascades started by a user was 928.

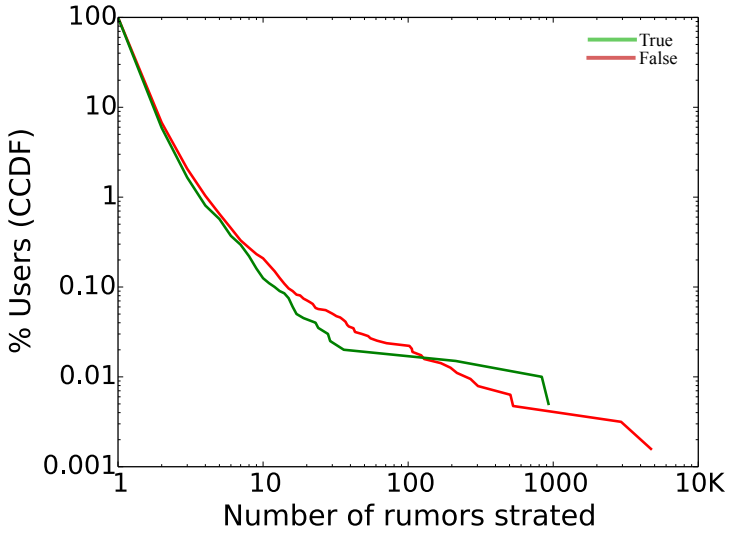

Figure S12: CCDF of number of false and true rumors started by unique users.

Another way to visualize these patterns is shown in Figure S13 below. As can be seen, in the figure, the top false rumor starter is responsible for 6% of all false rumor cascades and the top true rumor starter is responsible for 4% of all true rumor cascades. The top 10 false rumor starters are responsible for 12% of rumors started and the top 10 true rumor starters are responsible for 9% of rumors started. The figure shows a "knee" for both false and true rumors at around y=20. That point shows that 20% of all false rumor cascades are started by 1.7% (1,057) of the users and 20% of all true rumor cascades are started by 3.9% (940) of the users. Though not directly comparable with Gupta et al. [61], since they only looked at the retweets of fake images during hurricane Sandy, our results do show that a relatively small percentage of users are responsible for starting a large

number of rumor cascades and that this is slightly more true of false rumors than true rumors, however we hesitate to draw any dramatic conclusions from these observations.

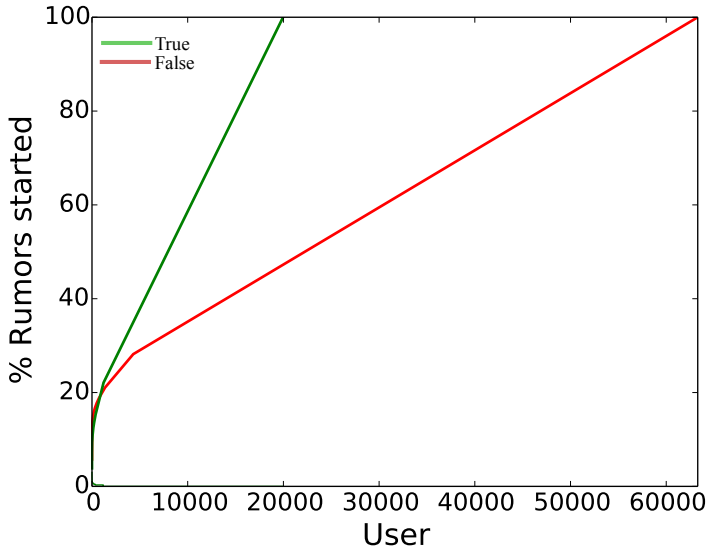

Figure S13: Cumulative plot of percent of false and true rumors started by different users.

S6 The Effect of Veracity on the Probability of Retweeting

As we showed in section S5, the difference between the spread of true and false rumors cannot be driven by the characteristics of users. Thus, we hypothesized that the veracity of a rumor has an effect on the probability of it being spread. To test this hypothesis we estimated a user-level logistic regression model of retweeting behavior as a function of falsehood and all of the user characteristics described in section S4, as follows:

$$logit(p_{retweet}) = \beta_0 + \beta_1 F + \beta_2 \mu_0 + \beta_3 \mu_1 + \beta_4 \mu_2 + \beta_5 \mu_3 + \beta_6 \mu_4 \qquad \text{(S5)}$$

Where:

- $p_{retweet}$ is the probability of a retweet of a rumor.
- β_0 is the intercept.
- $\beta_1, \beta_2 ... \beta_6$ are the coefficients (effects) of each of the parameters.
- $\mu_0, \mu 1 ... \mu_4$ are the user parameters (followers, followees, age, engagement, and verified).
- F is the falsehood of the rumor (1 if false, otherwise 0).

The model was estimated on 3,724,197 observations (impressions of a rumor comprising instances in our dataset where there could have been a retweet of a rumor). The results of the logistic regression are shown in Figure 4b in the main article. As shown in that figure, all the parameters had a statistically significant effect on the probability of a retweet, though, except for two factors, the effect size of all the parameters were small. The two factors with the largest effect sizes were whether the user was verified, followed by the falsehood of the rumor. Both effects were positive, meaning that they both increased the probability of a retweet. They had coefficients of 1.4261 and 0.5350 respectively, corresponding to odds ratios of 4.1625 and 1.7075, implying that, all else equal, a false tweet is 70% more likely to be retweeted than a true or mixed tweet.

S7 Measuring Emotional Responses and Rumor Novelty

Having shown that the probability of a rumor being retweeted is higher for false rumors, even after controlling for user characteristics, we hypothesized that there might be psychological reasons behind the differences between the spread of true and false rumors. Specifically, we hypothesized that false rumors tend to be more novel and surprising compared to true rumors. In order to test this hypothesis, we ran two analyses: 1) We measured the emotional content of the replies to the tweets containing rumors. 2) We compared the information contained in the rumor tweets to other information the users were exposed to on Twitter. We explain both analyses in detail below.

S7.1 Measuring Emotions in Responses to Rumors

As mentioned in section S2.2, a sizeable number of tweets containing rumors have replies associated with them (these are people replying to the tweet). To be more specific, 18,645 false rumors and 3,499 true rumors were replied to. We collected all the replies using our full access to the Twitter historical API. We categorized the emotions conveyed in these replies using the emotional lexicon collected and curated by the National Research Council of Canada (NRC). There are two manually curated datasets provided by the NRC, the first contains a comprehensive list of 141,820 English words and their associations with eight emotions: anger, fear, anticipation, trust, surprise, sadness, joy, and disgust [32][62]. These eight emotions are based on Plutchik's work on basic emotions [31]. The second dataset contains a list of 32,389 Twitter hashtags and their weighted associations with the same emotions [33] [63]. Both datasets have been manually curated and evaluated by the authors.

The reply tweets to a tweet are cleaned (cleaning entailed removing urls, usernames, stopwords, and correcting misspellings) and tokenized using Python's Natural Language Toolkit (NLTK) [64]. The cleaned tokens were then compared against the word and hashtag emotional dictionaries. It is possible (and likely) that there are multiple "hits" to different emotions, in these cases the scores are distributed between the emotions based on the number of hits. For example, it is possible for replies to a tweet to be 20% sadness, 50% surprise and 30% fear. If there was no match between the tokens and the dictionaries, the emotion is classified as miscellaneous. The emotional scores for false and true rumors are then aggregated and averaged to produce a mean and a standard error for each emotion. Figures 4d and 4f in the main article show the emotion scores for false and true rumors. Note that here, as with previous sections, we used cluster-robust standard errors (clustered on rumors). We found false rumors inspired replies expressing greater surprise (k-s test = .205, $p \sim 0.0$), corroborating the novelty hypothesis, and greater disgust (k-s test = .102, $p \sim 0.0$), while the truth inspired replies that expressed greater sadness (k-s test = .037, $p \sim 0.0$), anticipation (k-s test = .038, $p \sim 0.0$), joy (k-s test = .061, $p \sim 0.0$) and trust (k-s test = .060, $p \sim 0.0$) (Fig 4d and f in the main text).

We also used the linguistic inquiry and word count framework (LIWC) [65] to score replies based on assent and dissent words (as a proxy for agreement and disagreement). The LIWC scores (which correspond to the percentage of total words that match each of the categories) for assent and dissent in replies to false and true rumors are shown in Figure S14 below. You can see that that there is more dissent

than assent in the replies. There is not a statistically significant difference between assent words in replies to true and false rumors. However, there is a significant difference for dissent words, with false rumors having 13% more dissent words (ks-test D=0.07, p 0.0). We do not make any claims or generalizations about this, however, as the fraction of assent and dissent words in replies is so low (only making up between .5% and 2.8% of the total words in replies). Though this is not directly comparable with the work of Mendoza et al. [66] (since they study rumors during breaking news events), our finding, that false rumors tend to generate more disagreement, agrees with their findings (though the signal is much weaker in ours). In a related work, Zeng et al. [67] also looked at the diffusion of speed of rumor-denying and rumor-affirming posts during a hostage crisis event and found that rumor-denying posts spread faster than rumor-affirming posts. Though this work does not analyze rumor affirmation and denial (other than the analysis shown in Figure S14, this is a rich area of research that should be looked at in future research.

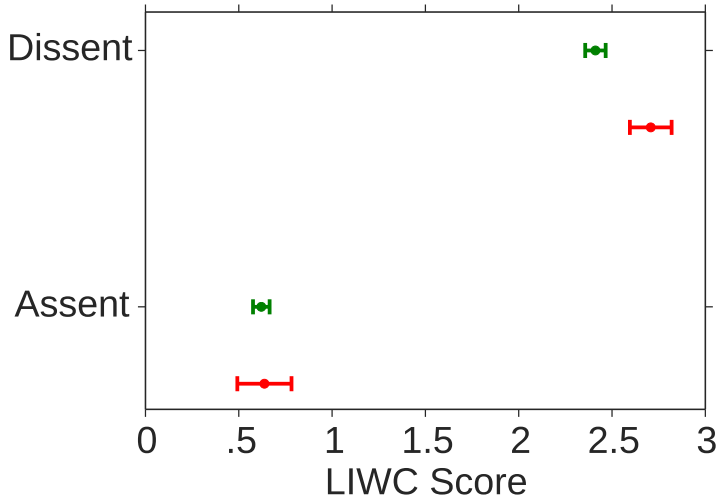

Figure S14: The fraction of assent and dissent words in replies to false and true rumors.

S7.2 Measuring the Novelty of Rumors

In order to measure the novelty of the information contained in the rumors, we randomly selected a subset of users involved in the propagation of true and false

rumors to match the prevalence of users in each type of cascade in our dataset (3,870 unique users involved in spreading false rumors and 899 unique users involved in spreading true rumors). We then randomly sampled 24,577 tweets that these 4,769 unique users were exposed to in the 60 days prior to them propagating (retweeting) one of the rumors (these are tweets posted by people the users followed at the time). Since we sampled the tweets that the users were exposed to, we are approximating what the user may have been exposed to.

Next, we calculated the information distance between the background tweets and the rumor tweets. To do so, we used a Latent Dirichlet Allocation (LDA) topic model [27] and trained on 10M English-language tweets and specifying 200 topics. Since we are dealing with tweets, we used a variant of LDA designed specifically to deal with tweets [68]. We then ran the trained LDA model on the 4,769 rumor retweets and the 24,577 background tweets (as mentioned earlier, OCR was used to extract text from images when applicable). This generated a probability distribution over the 200 topics for each of the tweets in our dataset.

Next, for each user, we compared the topic distribution of their rumor retweet and the topic distribution of the background tweets to which that particular user was exposed in the 60 days before being exposed to the rumor tweet. We used three metrics for this comparison: information uniqueness [69], KL-divergence [70] and the Bhattacharyya distance [30]. Below we explain how each of these metrics is calculated. Figure S15 shows a simplified illustration of the three steps involved in measuring novelty.

- Information Uniqueness (IU): IU measures the distance of topic distributions between two documents using cosine similarity. The formula for calculating IU is:

$$IU(\Gamma_r, \Gamma_b) = 1 - cos(\Gamma_r, \Gamma_b) \tag{S6}$$

  Where Γ_r and Γ_b correspond to the topic distribution of a retweeted rumor tweet and the background tweets respectively, and cos refers to the cosine similarity function. The higher the IU, the more unique or novel is Γ_r compared to Γ_b.

- KL-divergence (KL): KL is a measure of how one probability distribution diverges from a second expected probability distribution [71]. In our case, the two probability distributions correspond to the topic distributions of a retweeted rumor and the background tweets. The formula for calculating

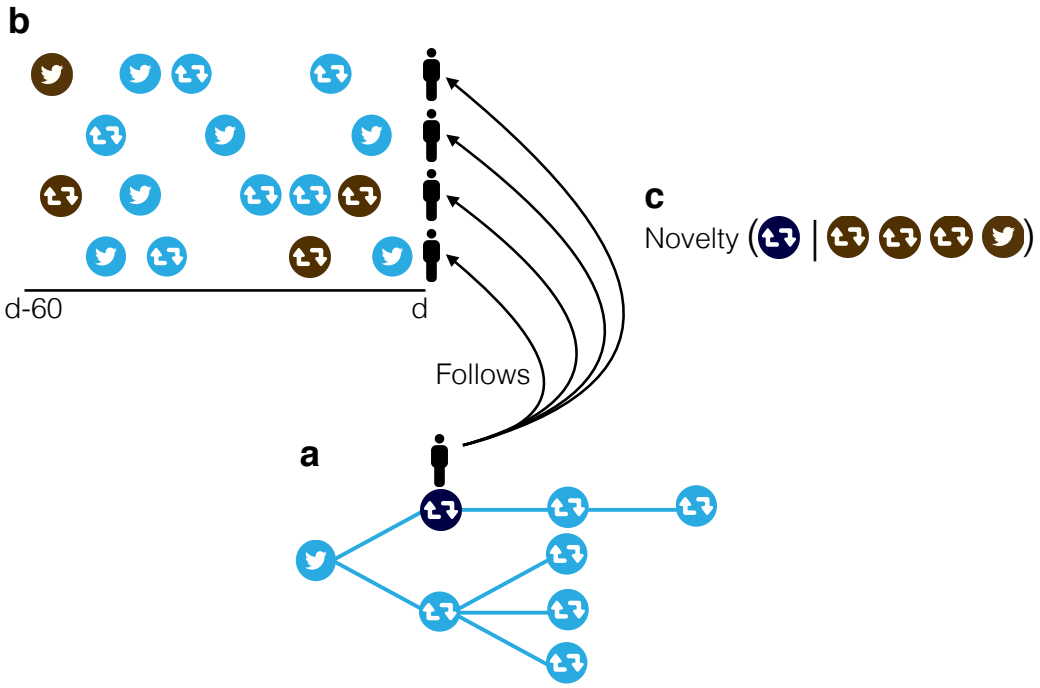

Figure S15: An illustration of the process through which the novelty of rumors is assessed: (a) a random selection of users involved in rumor propagation, (b) a random selection of tweets from the people the selected user followed in the 60 days prior to their retweet of the rumor tweet, (c) novelty is measured by comparing the tweet containing the rumor and the selected background tweets from the last 60 days.

KL-divergence from discrete probability distributions Q to P is:

$$D_{KL}(P \parallel Q) = \sum_i P(i) log(\frac{P(i)}{Q(i)}) \tag{S7}$$

KL-divergence is not symmetric, however we can create a symmetric metric for measuring the divergence using the following formula:

$$KL(\Gamma_r, \Gamma_b) = \frac{D_{KL}(\Gamma_r \parallel \Gamma_b) + D_{KL}(\Gamma_b \parallel \Gamma_r)}{2} \tag{S8}$$

Where Γ_r and Γ_b correspond to the topic distributions of a retweeted rumor

tweet and the background tweets respectively; $P(i)$ and $Q(i)$ in Equation S7 correspond to individual topics in Γ_r and Γ_b. As with all our other calculations, we again used cluster-robust standard errors, clustered on rumors. Lower KL indicates lower divergence between the two documents, therefore, the higher the KL, the more unique or novel is Γ_r compared to Γ_b.

- Bhattacharyya Distance (BD): BD measures the dissimilarity between two probability distributions. The formula for calculating BD is:

$$BD(\Gamma_r, \Gamma_b) = -ln(\sum_{x \in X} \sqrt{\Gamma_r(x)\Gamma_b(x)}) \tag{S9}$$

  Where Γ_r and Γ_b correspond to the topic distribution of a retweeted rumor tweet and the background tweets respectively; X corresponds to the set of 200 topics. The higher the BD, the more distance between the topic distributions in Γ_r and Γ_b.

After calculating these three novelty metrics, we aggregated and averaged the scores for false and true rumors to produce a mean and a standard error for each of the novelty metrics. Figures 4c and 4e show the results. There were statistically significant differences between true and false rumors for all three novelty scores. False rumors had significantly higher IU, higher KL, and higher BD, indicating greater novelty compared to true rumors. As with all of our other calculations, the standard errors reported in the figures are cluster-robust standard errors, clustered on rumors.

S7.3 Evaluating LDA

We evaluated our LDA models using a Tweet semantic similarity tool called Tweet2Vec [54]. Tweet2Vec, which uses recent advances in deep neural networks, was designed specifically to be robust to the short text, noise and misspellings commonly found on Twitter. Tweet2Vec, which creates semantic embeddings for tweets, was evaluated using data from recent SEMEVAL competitions during which it outperformed the winners of the competition in the tweet semantic similarity task (see [54] for more detail).

To further evaluate our LDA model, we first labelled each tweet using the most prominent topic in that tweet's LDA topic distribution (called hard-labelling). We then used Tweet2Vec to generate semantic embeddings for each tweet. Next we

created a fully connected graph with each tweet being a node and the weights of the edges set to the cosine similarity between the tweet embeddings. We then use the Louvain community detection algorithm [72] to cluster this graph. So each tweet belongs to a cluster based on the outcome of the Louvain clustering on Tweet2Vec embeddings (called C) and has a label based on its LDA topic distribution (called L). There was a high correlation between the labels L and the cluster-ids C (Pearson r = 0.63, p 0.001).

This analysis independently verifies that tweets that are hard-labelled with the same topic by our LDA model are semantically much more similar to each other than to tweets assigned to different topics. We used Tweet2Vec for evaluating the LDA models because (i) it is specifically designed for Twitter and its idiosyncrasies with regard to short text length, jargon and misspellings, (ii) it is a completely different technique than LDA, and thus can serve as a good robustness check to our topic modeling.

S8 Robustness Analysis

We employed several techniques to ensure the robustness of our analysis. First, we used cluster-robust standard errors for all of our analysis to account for within-cluster (i.e. rumor-level) error correlations (see section S9 for more information on this). Second, to address a possible selection bias in examining only fact checked rumors, we collected, curated and independently fact checked a second dataset of rumors that had never been assessed by any of our fact checking organizations (see section S8.1 below for more detail). Third, to assess the potential effects of bots on our analysis, we used a state-of-the-art bot detection algorithm to identify and remove bot accounts (see section S8.3) and then compared results both with and without bot traffic (see section S8.3.2). Fourth, we tested the robustness of our bot detection methods by a) comparing our results to those produced by using a second, independent bot detection algorithm and b) testing the sensitivity of our analysis to different bot detection thresholds.

S8.1 Robustness: Selection Bias

As mentioned in section S2.1, the rumors in our analysis represent the entire sample of fact checked rumors from six fact checking organizations such as snopes.com (amongst others). To validate our results and generalizations and to

assess our vulnerability to selection bias, we manually collected a second dataset of rumors that were not investigated by any of the six fact checking organizations. These rumors were collected by three undergraduate students at MIT and Wellesley College. We instructed the students on how to detect and investigate the rumors. First, we employed a system that detects stories spreading on Twitter [73], running the system on English-language tweets from June to December 2016. We ran the system until it had detected 300 unique rumors (we picked a relatively low threshold because each of these rumors had to be investigated manually). The system reached this threshold after analyzing around 3 million original tweets (i.e., not counting retweets). We then asked the student annotators to investigate these rumors. Some of what the system had detected as rumors ended up not being rumors (false positives), or were already investigated by one of the fact checking organizations; these were discarded (131 were discarded). Overall, there were 169 unique rumors, corresponding to 13,240 rumors cascades (7,979 false, and 5,261 true). These were investigated by the student annotators. The annotators worked independently of each other (in fact they were not aware of each other), and were asked to score the rumors as true, false or mixed. The annotators investigated all 169 rumors with a Fleiss' kappa (κ) of 0.88. Table S15 below shows the inter-annotator agreement. Figure S16a shows the CCDF of the number of cascades for false and true rumors in our robustness dataset. As can be seen, the number of cascades per rumor in our robustness dataset is fairly even between false and true rumors. The three annotators all agreed on the veracity score for 90% of the rumors. We used these 11,099 rumors cascades (6,291 false, and 4,808 true) for which we had unanimity in the veracity labels as our first robustness check. We then expanded the dataset to include majority rule veracity labeling. Figure S16b shows the CCDF of the number of cascades for false and true rumors in our robustness dataset.

	Annotator 1	Annotator 2
Annotator 2	90%	
Annotator 3	92%	95%

Table S15: Agreement between annotators on veracity of rumors. $\kappa = 0.88$

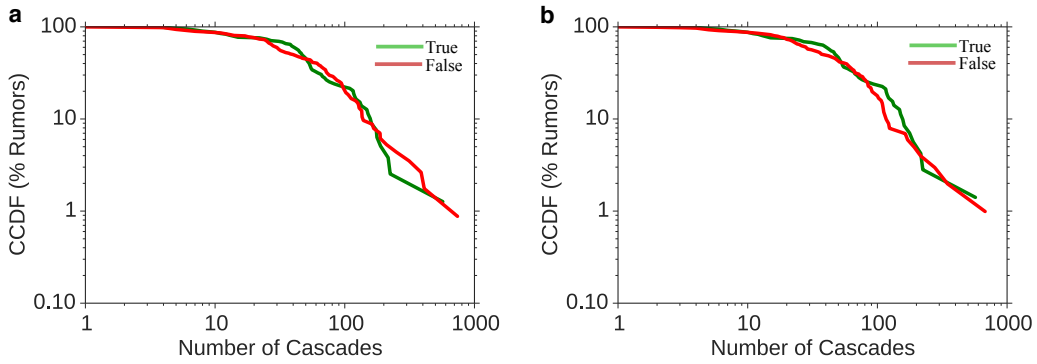

Figure S16: CCDF showing the number of cascades for false and true rumors in the robustness dataset. (a): all rumors, (b): rumors whose veracity was agreed upon by all three annotators.

S8.2 Analysis of Selection Bias

We ran the analysis described in section S3 on the two robustness datasets described above (unanimity labeling and majority-rule labeling) to see if our findings generalize to all rumors on Twitter. Figure S17 shows the results of our analysis on the unanimity labelled rumors and Figure S18 shows the same analysis for the majority-rule labelled rumors. The figures are analogous to Figure 2 in the main text. As with Figure 2, we show the cluster-robust standard errors in both Figures, S17 and S18. As can be seen, we found the same differences in the spread of true and false rumors in both robustness datasets as we did in our main analysis. This provides reassuring evidence that our analysis is not being affected by a selection bias from only examining fact checked rumors.

Tables S16, S17, S18 and S19 below show the mean (log), the cluster-robust standard errors (log) and the max and min of the depth, max-breadth, structural virality, and size for false and true rumors in the two robustness datasets (the tables show the mean and SE of the logged data). For each of these measurements, we also ran two-sample Kolmogorov-Smirnov (KS) tests to compare the distribution of these measures between false and true rumors cascades in each of the datasets. The results of the tests are reported in the table legends. Note that structural virality is only defined for cascades of sizes greater than one (this explains the smaller N in Table S18 compared to the other tables).

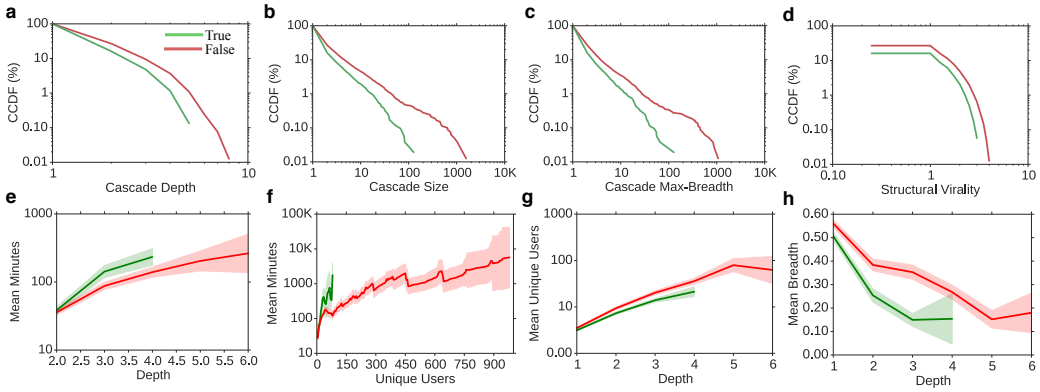

Figure S17: Difference between false and true rumor cascades in the validity dataset with unanimous labeling.

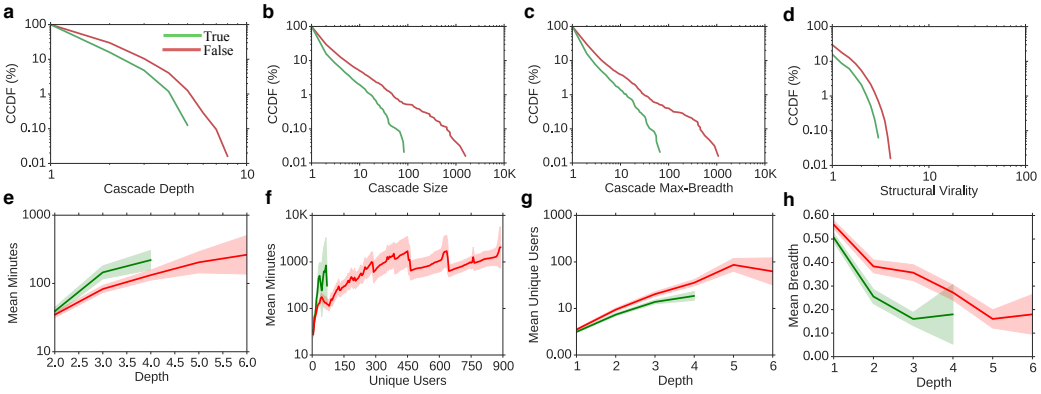

Figure S18: Difference between false and true rumor cascades in the validity dataset with majority rule labeling.

	N	Mean (log)	Robust-SE (log)	Min	Max
False-Unanimity	7,979	0.103	0.0068	0	8
True-Unanimity	5,261	0.058	0.0056	0	5
False-Majority Rule	6,291	0.115	0.0081	0	8
True-Majority Rule	4,809	0.058	0.0060	0	5

Table S16: Statistics on the depth of cascades. KS-test for false and true cascades for Unanimity: $D = 0.108$, $p \sim 0.0$ and Majority Rule: $D = 0.139$, $p \sim 0.0$.

	N	Mean (log)	Robust-SE (log)	Min	Max
False-Unanimity	7979	0.152	0.0102	1	1,075
True-Unanimity	5261	0.082	0.0088	1	128
False-Majority Rule	6,291	0.169	0.0125	1	1.075
True-Majority Rule	4,809	0.081	0.0094	1	66

Table S17: Statistics on the max-breadth of cascades. KS-test for false and true cascades for Unanimity: $D = 0.108$, $p \sim 0.0$ and Majority Rule: $D = 0.139$, $p \sim 0.0$.

	N	Mean (log)	Robust-SE (log)	Min	Max
False-Unanimity	2,142	0.149	0.0059	1.0	4.068
True-Unanimity	846	0.125	0.0077	1.0	3.185
False-Majority Rule	1,876	0.148	0.0065	1	4.068
True-Majority Rule	764	0.126	0.0084	1	3.185

Table S18: Statistics on the structural virality of cascades (of size 2 or greater). KS-test for false and true cascades for Unanimity: $D = 0.066$, $p \sim 0.0095$ and Majority Rule: $D = 0.063$, $p \sim 0.0243$.

	N	Mean (log)	Robust-SE (log)	Min	Max
False-Unanimity	7,979	0.165	0.0117	1	1,565
True-Unanimity	5,261	0.088	0.0097	1	128
False-Majority Rule	6,291	0.183	0.0143	1	1,565
True-Majority Rule	4,809	0.087	0.0104	1	83

Table S19: Statistics on the size of cascades. KS-test for false and true cascades for Unanimity: $D = 0.100$, $p \sim 0.0$ and Majority Rule: $D = 0.131$, $p \sim 0.0$.

S8.3 Robustness: Bot Traffic

The prevalence of bots on Twitter has been well studied [74, 75] and we wanted to ensure that our conclusions were robust to the presence of bots. We approached this problem in two ways. First, as explained in section S2.2.2, we used a bot-detection algorithm to identify and remove all accounts operated by bots. All analyses reported in the paper were conducted on a bot-free dataset. Second, we measured the effects of bots on our analysis by reanalyzing the data using all accounts (including the detected bots) and comparing the results to our original analysis.

S8.3.1 Detecting Bots

As explained in section S2.2.2, we used a state-of-the-art bot detection algorithm developed by Varol et al. [55], called *BotOrNot* to identify and remove bot accounts from our dataset. There is a publicly available API implemented by the authors that allows anyone to query a Twitter account.[4] The API returns a bot-likelihood score, between 0 and 1. We removed all accounts that had a bot-likelihood score of .5 or higher, which corresponded to 13.2% of the accounts. The threshold of .5 was recommended by the authors of the bot detection algorithm [55]. At this value, almost all "simple" bots and a large percentage of "sophisticated" bots get captured with a very small percentage of false positives.

BotOrNot uses more than 1,000 features from an account. These features fall under 6 main categories: network, user, friends, temporal, content, and sentiment. The network features relate to information diffusion patterns, the user features capture the account meta-data provided by Twitter, the friends features capture various aspects of the account's social graph, temporal features capture the timing patterns of the activity of the account, the content features are based on the linguistic cues in the tweets posted by the account, and finally sentiment features capture the emotions conveyed in the account's tweets. The algorithm was trained on 31K manually verified accounts (15K bot and 16K legitimate accounts), using a Random Forrest classifier. According to Varol et al., the algorithm has an AUC (Area Under ROC Curve) of 0.95, measured via cross validation.

Figure S19a shows the distribution of the bot-likelihood scores for all the accounts. The percentage of the accounts identified as bots (bot-likelihood score $>$ .5) were fairly evenly divided between false and true rumors, with a slight skew towards false rumors. Overall, 14.0% of the accounts in false rumors and 10.0% of the accounts in true rumors were identified as bots. Figure S19b shows the distribution of the bot-likelihood score for accounts involved in false and true rumors. As can be seen, the two distributions look very similar, but they are in fact statistically different from each other (shown using a KS-test: $D = 0.107$, $p \sim 0.0$). Figure S20 shows the CCDF of the bot-likelihood scores of accounts involved in false and true rumor cascades. These results indicate that bots are slightly more likely to participate in the spread of false rumors than in the spread of true rumors, but they do not address whether the added velocity and depth of the spread of false rumors can be attributed to the presence of bots.

[4] https://truthy.indiana.edu/botornot/

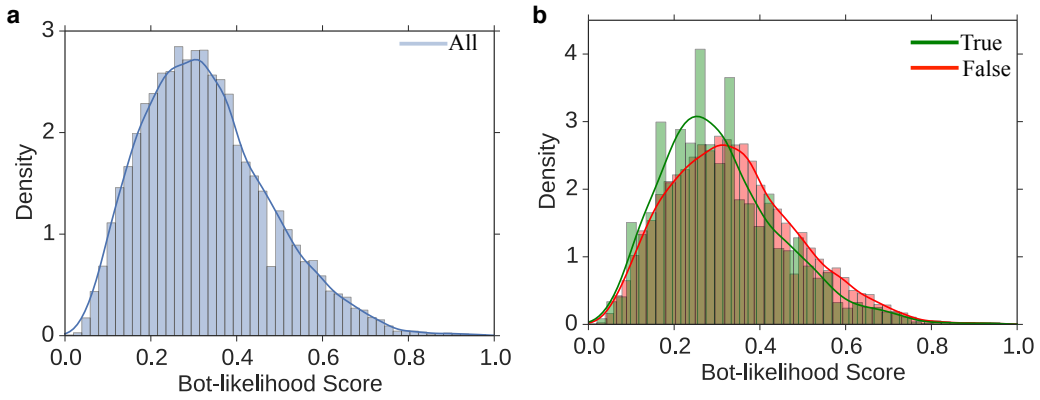

Figure S19: The bot-likelihood distribution of (a) all accounts, and (b) accounts involved in false and true cascades.

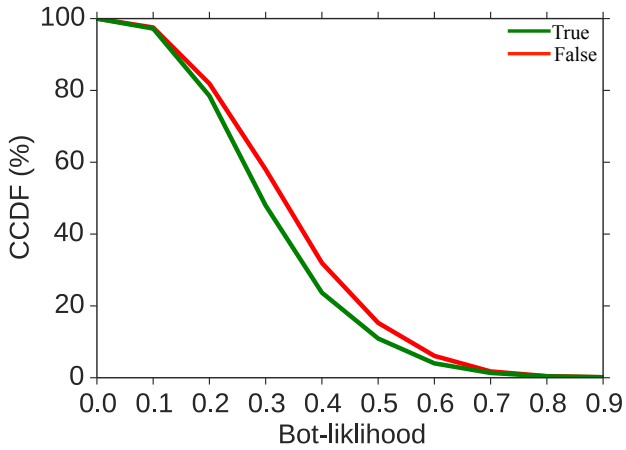

Figure S20: CCDF showing the bot-likelihood distributions for accounts involved in false and true rumors.

S8.3.2 Analysis

In order to better understand the potential effects of bots in our analysis, we ran the analysis described in section S3 on our full dataset (including the bot traffic that was excluded from the main analysis).

Tables S20, S21, S22, and S23 below show the mean (log), the cluster-robust standard errors (log) and the min and max of the depth, max-breadth, structural virality, and size for false and true rumors. For each of these measurements, we also ran a two-sample Kolmogorov-Smirnov (KS) test to compare the distribution

of these measures between false and true cascades. The results of the tests are reported in the caption of the tables. Note that as before, structural virality is only defined for cascades of size two or greater (thus the lower N in Table S22 compared to other tables).

There are three takeaways from this analysis: first, though the removal of bots did affect the results, all findings regarding the differences between true and false rumor cascades hold, even when bots are not removed. Second, the dataset with bots scores higher on all measures (e.g., depth and size) for both true and false cascades when compared to the dataset without bots. This indicates that bots are accelerating the spread of both true and false rumors. Removing bots decreased cascade size by $\sim$26%, depth by $\sim$21%, max-breadth by $\sim$26% and structural virality by $\sim$9%. Third, however, there are no meaningful differences between the increases in diffusion attributable to bots across true and false cascades. These findings imply that the presence of bots is not driving our results since they seem to effect true and false cascades similarly. The results also imply that false news spreads farther, faster, deeper and more broadly than the truth because humans, not robots, are more likely to spread it.

	N	Mean (log)	Margin (%)	Robust-SE (log)	Min	Max
False	101,179	0.197	-21%	0.0172	0	25
True	28,985	0.126	-21%	0.0316	0	12

Table S20: Statistics on the depth of cascades, including bots. The margin shows the difference when bots are removed. KS-test for false and true cascades: $D = 0.162$, $p \sim 0.0$

	N	Mean (log)	Margin (%)	Robust-SE (log)	Min	Max
False	101,179	0.391	-26%	0.0372	1	31,279
True	28,985	0.228	-25%	0.0686	1	1,717

Table S21: Statistics on the max-breadth of cascades, including bots. The margin shows the difference when bots are removed. KS-test for false and true cascades: $D = 0.162$, $p \sim 0.0$

	N	Mean (log)	Margin (%)	Robust-SE (log)	Min	Max
False	48,717	0.206	-9%	0.0064	1.0	10.51
True	9,251	0.181	-9%	0.0236	1.0	5.84

Table S22: Statistics on the structural virality of cascades (of size 2 or greater), including bots. The margin shows the difference when bots are removed. KS-test for false and true cascades: $D = 0.103$, $p \sim 0.0$

	N	Mean (log)	Margin (%)	Robust-SE (log)	Min	Max
False	101,179	0.419	-26%	0.0389	1	49,097
True	28,985	0.246	-24%	0.0725	1	1,873

Table S23: Statistics on the size of cascades, including bots. The margin shows the difference when bots are removed. KS-test for false and true cascades: $D = 0.161$, $p \sim 0.0$

S8.3.3 Secondary Analysis

We also ran a more extreme version of the analysis where we removed all retweets that trace their origin to a bot account, even if the retweets themselves were made by non-bot accounts.

Tables S24, S25, S26, and S27 below show the mean (log), the cluster-robust standard errors (log) and the min and max of the depth, max-breadth, structural virality, and size of false and true cascades. For each of these measurements, we also ran two-sample Kolmogorov-Smirnov (KS) tests to compare their distributions across true and false cascades. The results of the tests are reported in the table captions. Note that, as before, structural virality is only defined for cascades of size two or greater (thus the lower N in Table S26 compared to other tables).

From this analysis we see that even in this extreme bot-removal analysis, where all cascades originated by bots are removed, our findings still hold. Similar to the previous analysis, we see that bots are accelerating the spread of both true and false rumors. When we removed cascades originated by bots, both true and false cascades were similarly affected; compared to the dataset containing cascades with just bots removed, in both cases we saw a reduction in cascade size by $\sim$23%, depth by $\sim$19%, max-breadth by $\sim$24% and structural virality by $\sim$10%; compared to the dataset containing all cascades, in both cases we saw a reduction in cascades' size by $\sim$42%, depth by $\sim$36%, max-breadth by $\sim$42% and structural virality by $\sim$18%. We also saw a reduction in the spread of false rumors and true

rumors, with false rumors being reduced slightly more than the velocity of the spread of true rumors when bots are removed, with an average reduction of 19% for false rumors and 11% for true rumors. This further validates and confirms the conclusions of our bot analysis in the previous section (Section S8.3.2).

	N	Mean (log)	Margin (%)	Robust-SE (log)	Min	Max
False	73,294	0.127	-36%	0.0099	0	22
True	22,121	0.081	-36%	0.0110	0	11

Table S24: Statistics on the depth of cascades, excluding all cascades originated by bots. KS-test for false and true cascades: $D = 0.111$, $p \sim 0.0$

	N	Mean (log)	Margin (%)	Robust-SE (log)	Min	Max
False	73,294	0.218	-44%	0.0196	1	21,190
True	22,121	0.136	-40%	0.0235	1	1,390

Table S25: Statistics on the max-breadth of cascades, excluding all cascades originated by bots. KS-test for false and true cascades: $D = 0.111$, $p \sim 0.0$

	N	Mean (log)	Margin (%)	Robust-SE (log)	Min	Max
False	23,403	0.169	-18%	0.0082	1.0	9.91
True	4,598	0.147	-19%	0.0159	1.0	5.53

Table S26: Statistics on the structural virality of cascades (of size 2 or greater), excluding all cascades originated by bots. KS-test for false and true cascades: $D = 0.101$, $p \sim 0.0$

	N	Mean (log)	Margin (%)	Robust-SE (log)	Min	Max
False	73,294	0.237	-43%	0.0215	1	43,146
True	22,121	0.147	-40%	0.0257	1	1,599

Table S27: Statistics on the size of cascades, excluding all cascades originated by bots. KS-test for false and true cascades: $D = 0.110$, $p \sim 0.0$

S8.3.4 Bot Sensitivity

In this section, we measure the effect of bot detection sensitivity on our analysis. As mentioned, for our analysis we set the bot detection likelihood threshold to be 0.5 as this is the threshold that performs the best and is recommended by the authors of the algorithm. Here, we measure the effects of more liberal thresholds (lower than 0.5) on our analysis. We reran the bot detection algorithm for thresholds of 0.1, 0.2 ,0.3, and 0.4. At these thresholds, 96%, 79%, 54%, and 27% of accounts are identified as bots respectively. We then reran our analysis for thresholds of 0.3, and 0.4 to understand the effect of bot detection sensitivity on our findings (there is not enough data to rerun the analysis on thresholds of 0.1 and 0.2).

From this analysis we see that even for the very liberal bot detection and removal thresholds of 0.4 and 0.3 (where 27%, and 54% of the accounts respectively are identified as bots), all our findings still hold. This shows that our results are not dependent on the sensitivity of the bot detection algorithm.

Sensitivity at 0.4 Tables S28, S29, S30, and S31 below show the mean (log), the cluster-robust standard errors (log) and the min and max of the depth, max-breadth, structural virality, and size of false and true cascades. For each of these measurements, we also ran two-sample Kolmogorov-Smirnov (KS) tests to compare their distributions across true and false cascades. The results of the tests are reported in the table captions.

	N	Mean (log)	Margin (%)	Robust-SE (log)	Min	Max
False	66,407	0.101	-49%	0.0068	0	18
True	21,277	0.075	-40%	0.0104	0	11

Table S28: Statistics on the depth of cascades, excluding bots (sensitivity set at 0.4). KS-test for false and true cascades: $D = 0.055$, $p \sim 0.0$

	N	Mean (log)	Margin (%)	Robust-SE (log)	Min	Max
False	66,407	0.159	-59%	0.0123	1	14,044
True	21,277	0.126	-45%	0.0205	1	1,156

Table S29: Statistics on the max-breadth of cascades, excluding bots (sensitivity set at 0.4). KS-test for false and true cascades: $D = 0.055$, $p \sim 0.0$

	N	Mean (log)	Margin (%)	Robust-SE (log)	Min	Max
False	17,284	0.146	-29%	0.0085	1.0	9.05
True	4,374	0.139	-23%	0.0142	1.0	5.30

Table S30: Statistics on the structural virality of cascades (of size 2 or greater), excluding bots (sensitivity set at 0.4). KS-test for false and true cascades: $D = 0.0441$, $p \sim 0.0$

	N	Mean (log)	Margin (%)	Robust-SE (log)	Min	Max
False	66,407	0.174	-58%	0.0140	1	40,711
True	21,277	0.137	-44%	0.0226	1	1,485

Table S31: Statistics on the size of cascades, excluding bots (sensitivity set at 0.4). KS-test for false and true cascades: $D = 0.0533$, $p \sim 0.0$

Sensitivity at 0.3 Tables S32, S33, S34, and S35 below show the mean (log), the cluster-robust standard errors (log) and the min and max of the depth, max-breadth, structural virality, and size of false and true cascades. For each of these measurements, we also ran two-sample Kolmogorov-Smirnov (KS) tests to compare their distributions across true and false cascades. The results of the tests are reported in the table captions.

	N	Mean (log)	Margin (%)	Robust-SE (log)	Min	Max
False	52,593	0.086	-56%	0.0058	0	15
True	19,161	0.070	-44%	0.0103	0	9

Table S32: Statistics on the depth of cascades, excluding bots (sensitivity set at 0.3). KS-test for false and true cascades: $D = 0.023$, $p \sim 0.0$

	N	Mean (log)	Margin (%)	Robust-SE (log)	Min	Max
False	52,593	0.128	-67%	0.0094	1	11,309
True	19,161	0.116	-49%	0.0201	1	973

Table S33: Statistics on the max-breadth of cascades, excluding bots (sensitivity set at 0.3). KS-test for false and true cascades: $D = 0.023$, $p \sim 0.0$

	N	Mean (log)	Margin (%)	Robust-SE (log)	Min	Max
False	13,436	0.138	-31%	0.0077	1.0	8.15
True	3,689	0.130	-28%	0.0119	1.0	4.67

Table S34: Statistics on the structural virality of cascades (of size 2 or greater), excluding bots (sensitivity set at 0.3). KS-test for false and true cascades: $D = 0.057$, $p \sim 0.0$

	N	Mean (log)	Margin (%)	Robust-SE (log)	Min	Max
False	52,593	0.139	-67%	0.0107	1	35,139
True	19,161	0.117	-52%	0.0222	1	1,053

Table S35: Statistics on the size of cascades, excluding bots (sensitivity set at 0.3). KS-test for false and true cascades: $D = 0.0213$, $p \sim 0.0$

S8.3.5 Alternative Bot Detection Algorithm

As a final test of the robustness of our results with respect to bot activity on Twitter, we reran our analysis using another bot detection algorithm. This algorithm was trained on features suggested by Almaatouq et al. [75] to identify spammer accounts in our dataset.

Overall, around 11.1% of the accounts were identified as bots using Almaatouq et al.'s detection algorithm, corresponding to 11.9% of the accounts in false rumors and 10.2% of the accounts in true rumors.

Tables S36, S37, S38, and S39 below show the mean (log), the cluster-robust standard errors (log) and the min and max of the depth, max-breadth, structural virality, and size of false and true cascades. For each of these measurements, we also ran two-sample Kolmogorov-Smirnov (KS) tests to compare their distributions across true and false cascades. The results of the tests are reported in the table captions.

From this analysis we see that our findings hold when using another independent, but well designed, detection algorithm. Similar to when using the primary bot detection algorithm, we see that bot accounts are accelerating the spread of both true and false rumors.

	N	Mean (log)	Margin (%)	Robust-SE (log)	Min	Max
False	80,375	0.149	-24%	0.0135	0	23
True	22,455	0.083	-34%	0.0114	0	12

Table S36: Statistics on the depth of cascades, excluding bots (using the alternative detection algorithm). KS-test for false and true cascades: $D = 0.163$, $p \sim 0.0$

	N	Mean (log)	Margin (%)	Robust-SE (log)	Min	Max
False	80,375	0.273	-30%	0.0283	1	28,859
True	22,455	0.143	-37%	0.0256	1	1,559

Table S37: Statistics on the max-breadth of cascades, excluding bots (using the alternative detection algorithm). KS-test for false and true cascades: $D = 0.163$, $p \sim 0.0$

	N	Mean (log)	Margin (%)	Robust-SE (log)	Min	Max
False	30,060	0.183	-11%	0.0075	1.0	10.01
True	4,740	0.152	-16%	0.0166	1.0	5.72

Table S38: Statistics on the structural virality of cascades (of size 2 or greater), excluding bots (using the alternative detection algorithm). KS-test for false and true cascades: $D = 0.158$, $p \sim 0.0$

	N	Mean (log)	Margin (%)	Robust-SE (log)	Min	Max
False	80,375	0.295	-30%	0.0279	1	46,196
True	22,455	0.155	-37%	0.0301	1	1,646

Table S39: Statistics on the size of cascades, excluding bots (using the alternative detection algorithm). KS-test for false and true cascades: $D = 0.162$, $p \sim 0.0$

S8.4 Goodness-of-fit Analysis

We measured the goodness-of-fit of the logistic regression model described in section S6 and shown in Figure 4b in the main text in several ways. We used the deviance goodness-of-fit statistic, which is well-suited to logistic regression. The total deviance goodness-of-fit statistic of our model was $3.4649e + 06$, with $3,724,190$ degrees of freedom, corresponding to a p-value of 1. This means there are no grounds to reject the null hypothesis that the model is well specified. Moreover, the log-likelihood of our model was $-1.7170e + 06$, with a log-likelihood

ratio chi-squared test p-value of 0.000, meaning that the observed relationships are unlikely to have been found due to chance.

S9 Cluster-robust Standard Errors

All standard errors and regression models reported in this paper are cluster-robust. This is important for data that are grouped into clusters (such as ours). It is unlikely that the propagation dynamics of the cascades of the same rumor are unrelated (though it is reasonable to assume that whatever the effects might be, they effect the cascades of the same rumor similarly). As explained, the clusters in our dataset are the unique rumors. By using cluster-robust methods, we ensure that any correlation within the rumor clusters is accounted for. Not controlling for possible error correlation within clusters would most likely yield standard errors that are misleadingly small. To learn more about cluster-robust inference please refer to Cameron and Miller's [59] excellent article on this subject. As expected, when we do not cluster the standard errors the magnitudes and directions of the coefficients remain the same, but the precision of the estimates increases.

S10 Complementary Cumulative Distribution Function

We have used the complementary cumulative distribution function (CCDF) extensively in our analysis. As the name suggests, the CCDF is the complement of the cumulative distribution function (CDF), that is, it measures how often a distribution function is above a particular level (whereas the CDF measures how often a distribution function is below or equal to a particular level). Thus, the CCDF, $\overline{F}$, of distribution function of X, evaluated at x, is the probability that X will take a value more than x:

$$\overline{F}_X(x) = P(X > x) \tag{S10}$$

In our case, all probabilities are calculated empirically from our data.

References and Notes

1. L. J. Savage, The theory of statistical decision. *J. Am. Stat. Assoc.* **46**, 55–67 (1951). doi:10.1080/01621459.1951.10500768

2. H. A. Simon, *The New Science of Management Decision* (Harper & Brothers Publishers, New York, 1960).

3. R. Wedgwood, The aim of belief. *Noûs* **36**, 267–297 (2002). doi:10.1111/1468-0068.36.s16.10

4. E. Fehr, U. Fischbacher, The nature of human altruism. *Nature* **425**, 785–791 (2003). doi:10.1038/nature02043 Medline

5. C. E. Shannon, A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948). doi:10.1002/j.1538-7305.1948.tb01338.x

6. S. Bikhchandani, D. Hirshleifer, I. Welch, A theory of fads, fashion, custom, and cultural change as informational cascades. *J. Polit. Econ.* **100**, 992–1026 (1992). doi:10.1086/261849

7. K. Rapoza, "Can 'fake news' impact the stock market?" *Forbes*, 26 February 2017; www.forbes.com/sites/kenrapoza/2017/02/26/can-fake-news-impact-the-stock-market/.

8. M. Mendoza, B. Poblete, C. Castillo, "Twitter under crisis: Can we trust what we RT?" in *Proceedings of the First Workshop on Social Media Analytics* (Association for Computing Machinery, ACM, 2010), pp. 71–79.

9. A. Gupta, H. Lamba, P. Kumaraguru, A. Joshi, "Faking Sandy: Characterizing and identifying fake images on Twitter during Hurricane Sandy," in *Proceedings of the 22nd International Conference on World Wide Web* (ACM, 2010), pp. 729–736.

10. K. Starbird, J. Maddock, M. Orand, P. Achterman, R. M. Mason, "Rumors, false flags, and digital vigilantes: Misinformation on Twitter after the 2013 Boston Marathon bombing," in *iConference 2014 Proceedings* (iSchools, 2014).

11. J. Gottfried, E. Shearer, "News use across social media platforms," Pew Research Center, 26 May 2016; www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/.

12. C. Silverman, "This analysis shows how viral fake election news stories outperformed real news on Facebook," *BuzzFeed News*, 16 November 2016; www.buzzfeed.com/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook/.

13. M. De Domenico, A. Lima, P. Mougel, M. Musolesi, The anatomy of a scientific rumor. *Sci. Rep.* **3**, 2980 (2013). doi:10.1038/srep02980 Medline

14. O. Oh, K. H. Kwon, H. R. Rao, "An exploration of social media in extreme events: Rumor theory and Twitter during the Haiti earthquake 2010," in *Proceedings of the International Conference on Information Systems* (International Conference on Information Systems, ICIS, paper 231, 2010).

15. M. Tambuscio, G. Ruffo, A. Flammini, F. Menczer, “Fact-checking effect on viral hoaxes: A model of misinformation spread in social networks,” in *Proceedings of the 24th International Conference on World Wide Web* (ACM, 2015), pp. 977–982.

16. Z. Zhao, P. Resnick, Q. Mei, “Enquiring minds: Early detection of rumors in social media from enquiry posts,” in *Proceedings of the 24th International Conference on World Wide Web* (ACM, 2015), pp. 1395–1405.

17. M. Gupta, P. Zhao, J. Han, “Evaluating event credibility on Twitter,” in *Proceedings of the 2012 Society for Industrial and Applied Mathematics International Conference on Data Mining* (Society for Industrial and Applied Mathematics, SIAM, 2012), pp. 153–164.

18. G. L. Ciampaglia, P. Shiralkar, L. M. Rocha, J. Bollen, F. Menczer, A. Flammini, Computational fact checking from knowledge networks. *PLOS ONE* **10**, e0128193 (2015). doi:10.1371/journal.pone.0128193 Medline

19. A. Friggeri, L. A. Adamic, D. Eckles, J. Cheng, “Rumor cascades,” in *Proceedings of the International Conference on Weblogs and Social Media* (Association for the Advancement of Artificial Intelligence, AAAI, 2014)

20. M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, W. Quattrociocchi, The spreading of misinformation online. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 554–559 (2016). doi:10.1073/pnas.1517441113 Medline

21. A. Bessi, M. Coletto, G. A. Davidescu, A. Scala, G. Caldarelli, W. Quattrociocchi, Science vs conspiracy: Collective narratives in the age of misinformation. *PLOS ONE* **10**, e0118093 (2015). doi:10.1371/journal.pone.0118093 Medline

22. Friggeri *et al*. (*19*) do evaluate two metrics of diffusion: depth, which shows little difference between true and false rumors, and shares per rumor, which is higher for true rumors than it is for false rumors. Although these results are important, they are not definitive owing to the smaller sample size of the study; the early timing of the sample, which misses the rise of false news after 2013; and the fact that more shares per rumor do not necessarily equate to deeper, broader, or more rapid diffusion.

23. S. Goel, A. Anderson, J. Hofman, D. J. Watts, The structural virality of online diffusion. *Manage. Sci.* **62**, 180–196 (2015).

24. L. Itti, P. Baldi, Bayesian surprise attracts human attention. *Vision Res.* **49**, 1295–1306 (2009). doi:10.1016/j.visres.2008.09.007 Medline

25. S. Aral, M. Van Alstyne, The diversity-bandwidth trade-off. *Am. J. Sociol.* **117**, 90–171 (2011). doi:10.1086/661238

26. J. Berger, K. L. Milkman, What makes online content viral? *J. Mark. Res.* **49**, 192–205 (2012). doi:10.1509/jmr.10.0353

27. D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003).

28. S. Aral, P. Dhillon, “Unpacking novelty: The anatomy of vision advantages,” Working paper, MIT–Sloan School of Management, Cambridge, MA, 22 June 2016; https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2388254.

29. T. M. Cover, J. A. Thomas, *Elements of Information Theory* (Wiley, 2012).

30. T. Kailath, The divergence and Bhattacharyya distance measures in signal selection. *IEEE Trans. Commun. Technol.* **15**, 52–60 (1967). doi:10.1109/TCOM.1967.1089532

31. R. Plutchik, The nature of emotions. *Am. Sci.* **89**, 344–350 (2001). doi:10.1511/2001.4.344

32. S. M. Mohammad, P. D. Turney, Crowdsourcing a word-emotion association lexicon. *Comput. Intell.* **29**, 436–465 (2013). doi:10.1111/j.1467-8640.2012.00460.x

33. S. M. Mohammad, S. Kiritchenko, Using hashtags to capture fine emotion categories from tweets. *Comput. Intell.* **31**, 301–326 (2015). doi:10.1111/coin.12024

34. S. Vosoughi, D. Roy, "A semi-automatic method for efficient detection of stories on social media," in *Proceedings of the 10th International AAAI Conference on Weblogs and Social Media* (AAAI, 2016), pp. 707–710.

35. C. A. Davis, O. Varol, E. Ferrara, A. Flammini, F. Menczer, "BotOrNot: A system to evaluate social bots," in *Proceedings of the 25th International Conference on World Wide Web* (ACM, 2016), pp. 273–274.

36. For example, this is an argument made in recent testimony by Clint Watts—Robert A. Fox Fellow at the Foreign Policy Research Institute and Senior Fellow at the Center for Cyber and Homeland Security at George Washington University—given during the U.S. Senate Select Committee on Intelligence hearing on "Disinformation: A Primer in Russian Active Measures and Influence Campaigns" on 30 March 2017; www.intelligence.senate.gov/sites/default/files/documents/os-cwatts-033017.pdf.

37. D. Trpevski, W. K. Tang, L. Kocarev, Model for rumor spreading over networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **81**, 056102 (2010). doi:10.1103/PhysRevE.81.056102 Medline

38. B. Doerr, M. Fouz, T. Friedrich, Why rumors spread so quickly in social networks. *Commun. ACM* **55**, 70–75 (2012). doi:10.1145/2184319.2184338

39. F. Jin, E. Dougherty, P. Saraf, Y. Cao, N. Ramakrishnan, "Epidemiological modeling of news and rumors on Twitter," in *Proceedings of the 7th Workshop on Social Network Mining and Analysis* (ACM, 2013).

40. J. Cheng, L. A. Adamic, J. M. Kleinberg, J. Leskovec, "Do cascades recur?" in *Proceedings of the 25th International Conference on World Wide Web* (ACM, 2016).

41. V. Qazvinian, E. Rosengren, D. R. Radev, Q. Mei, "Rumor has it: Identifying misinformation in microblogs," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, ACL, 2011).

42. S. Vosoughi, M. Mohsenvand, D. Roy, Rumor gauge: Predicting the veracity of rumors on Twitter. *ACM Trans. Knowl. Discov. Data* **11**, 50 (2017). doi:10.1145/3070644

43. W. Xu, H. Chen, "Scalable rumor source detection under independent cascade model in online social networks," in *2015 11th International Conference on Mobile Ad-hoc and Sensor Networks (MSN)* (IEEE, 2015).

44. T. Takahashi, N. Igata, “Rumor detection on Twitter,” in *2012 Joint 6th International Conference on Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS)* (IEEE, 2012).

45. C. Castillo, M. Mendoza, B. Poblete, “Information credibility on Twitter,” in *Proceedings of the 20th International Conference on World Wide Web* (ACM, 2011).

46. R. M. Tripathy, A. Bagchi, S. Mehta, “A study of rumor control strategies on social networks,” in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (ACM, 2010).

47. J. Shin, L. Jian, K. Driscoll, F. Bar, Political rumoring on Twitter during the 2012 U.S. presidential election: Rumor diffusion and correction. *New Media Soc.* **19**, 1214–1235 (2017).

48. P. Ozturk, H. Li, Y. Sakamoto, “Combating rumor spread on social media: The effectiveness of refutation and warning,” in *2015 48th Hawaii International Conference on System Sciences (HICSS)* (IEEE, 2015).

49. A. Bessi, F. Petroni, M. Del Vicario, F. Zollo, A. Anagnostopoulos, A. Scala, G. Caldarelli, W. Quattrociocchi, Homophily and polarization in the age of misinformation. *Eur. Phys. J. Spec. Top.* **225**, 2047–2059 (2016). doi:10.1140/epjst/e2015-50319-0

50. A. Bessi, A. Scala, L. Rossi, Q. Zhang, W. Quattrociocchi. The economy of attention in the age of (mis)information. *J. Trust Manage.* **1**, 12 (2014). doi:10.1186/s40493-014-0012-y

51. A. Mitchell, J. Gottfried, J. Kiley, K. E. Matsa, “Political polarization & media habits,” Pew Research Center; www.journalism.org/2014/10/21/political-polarization-media-habits/.

52. J. L. Fleiss, Measuring nominal scale agreement among many raters. *Psychol. Bull.* **76**, 378–382 (1971). doi:10.1037/h0031619

53. Q. Le, T. Mikolov, “Distributed representations of sentences and documents,” in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)* (Journal of Machine Learning Research, 2014).

54. S. Vosoughi, P. Vijayaraghavan, D. Roy, “Tweet2vec: Learning tweet embeddings using character-level cnn-lstm encoder-decoder,” in *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (ACM, 2016).

55. C. A. Davis, O. Varol, E. Ferrara, A. Flammini, F. Menczer, “Botornot: A system to evaluate social bots,” in *Proceedings of the 25th International Conference Companion on World Wide Web* (ACM, 2016).

56. J. Maddock, K. Starbird, R. M. Mason, “Using historical Twitter data for research: Ethical challenges of tweet deletions,” in *CSCW 2015 Workshop on Ethics for Studying Sociotechnical Systems in a Big Data World* (ACM, 2015).

57. S. Goel, D. J. Watts, D. G. Goldstein, “The structure of online diffusion networks,” in *Proceedings of the 13th ACM conference on Electronic Commerce* (ACM, 2012).

58. J. M. Wooldridge, Cluster-sample methods in applied econometrics. *Am. Econ. Rev.* **93**, 133–138 (2003). doi:10.1257/000282803321946930

59. A. C. Cameron, D. L. Miller, A practitioner's guide to cluster-robust inference. *J. Hum. Resour.* **50**, 317–372 (2015). doi:10.3368/jhr.50.2.317

60. P. Vijayaraghavan, S. Vosoughi, D. Roy, "Twitter demographic classification using deep multi-modal multi-task learning," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 2: Short Papers)* (ACL, 2017).

61. A. Gupta, H. Lamba, P. Kumaraguru, A. Joshi, "Faking Sandy: Characterizing and identifying fake images on Twitter during Hurricane Sandy," in *Proceedings of the 22nd International Conference on World Wide Web* (ACM, 2013).

62. S. M. Mohammad, P. D. Turney, "Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon," in *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text* (ACL, 2010).

63. S. M. Mohammad, "# emotional tweets," in *Proceedings of the First Joint Conference on Lexical and Computational Semantics* (ACL, 2012).

64. S. Bird, E. Klein, E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit* (O'Reilly Media, ed. 1, 2009).

65. J. W. Pennebaker, M. E. Francis, R. J. Booth, Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* **71**, 2001 (2001).

66. M. Mendoza, B. Poblete, C. Castillo, "Twitter under crisis: Can we trust what we RT?" in *Proceedings of the First Workshop on Social Media Analytics* (ACM, 2010).

67. L. Zeng, K. Starbird, E. S. Spiro, "Rumors at the speed of light? Modeling the rate of rumor transmission during crisis," in *2016 49th Hawaii International Conference on System Sciences (HICSS)* (IEEE, 2016).

68. W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, X. Li, "Comparing Twitter and traditional media using topic models," in *European Conference on Information Retrieval (ECIR)* (ECIR, 2011).

69. S. Aral, P. Dhillon, "Unpacking novelty: The anatomy of vision advantages," Working paper, MIT–Sloan School of Management, Cambridge, MA, 22 June 2016.

70. T. M. Cover, J. A. Thomas, *Elements of Information Theory* (Wiley, ed. 2, 2012).

71. S. Kullback, R. A. Leibler, On information and sufficiency. *Ann. Math. Stat.* **22**, 79–86 (1951). doi:10.1214/aoms/1177729694

72. V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**, P10008 (2008).

73. S. Vosoughi, D. Roy, "A semi-automatic method for efficient detection of stories on social media," in *10th International AAAI Conference on Web and Social Media* (AAAI, 2016).

74. E. Ferrara, O. Varol, C. Davis, F. Menczer, A. Flammini, The rise of social bots. *Commun. ACM* **59**, 96–104 (2016). doi:10.1145/2818717

75. A. Almaatouq, E. Shmueli, M. Nouh, A. Alabdulkareem, V. K. Singh, M. Alsaleh, A. Alarifi, A. Alfaris, A. Pentland, If it looks like a spammer and behaves like a spammer, it must be a spammer: Analysis and detection of microblogging spam accounts. *Int. J. Inf. Secur.* **15**, 475–491 (2016). doi:10.1007/s10207-016-0321-5