Supplementary Information for: Assessing the risks of "infodemics" in response to COVID-19 epidemics

Riccardo Gallotti[1], Francesco Valle[1], Nicola Castaldo[1], Pierluigi Sacco[2,3,4*] & Manlio De Domenico[1*]

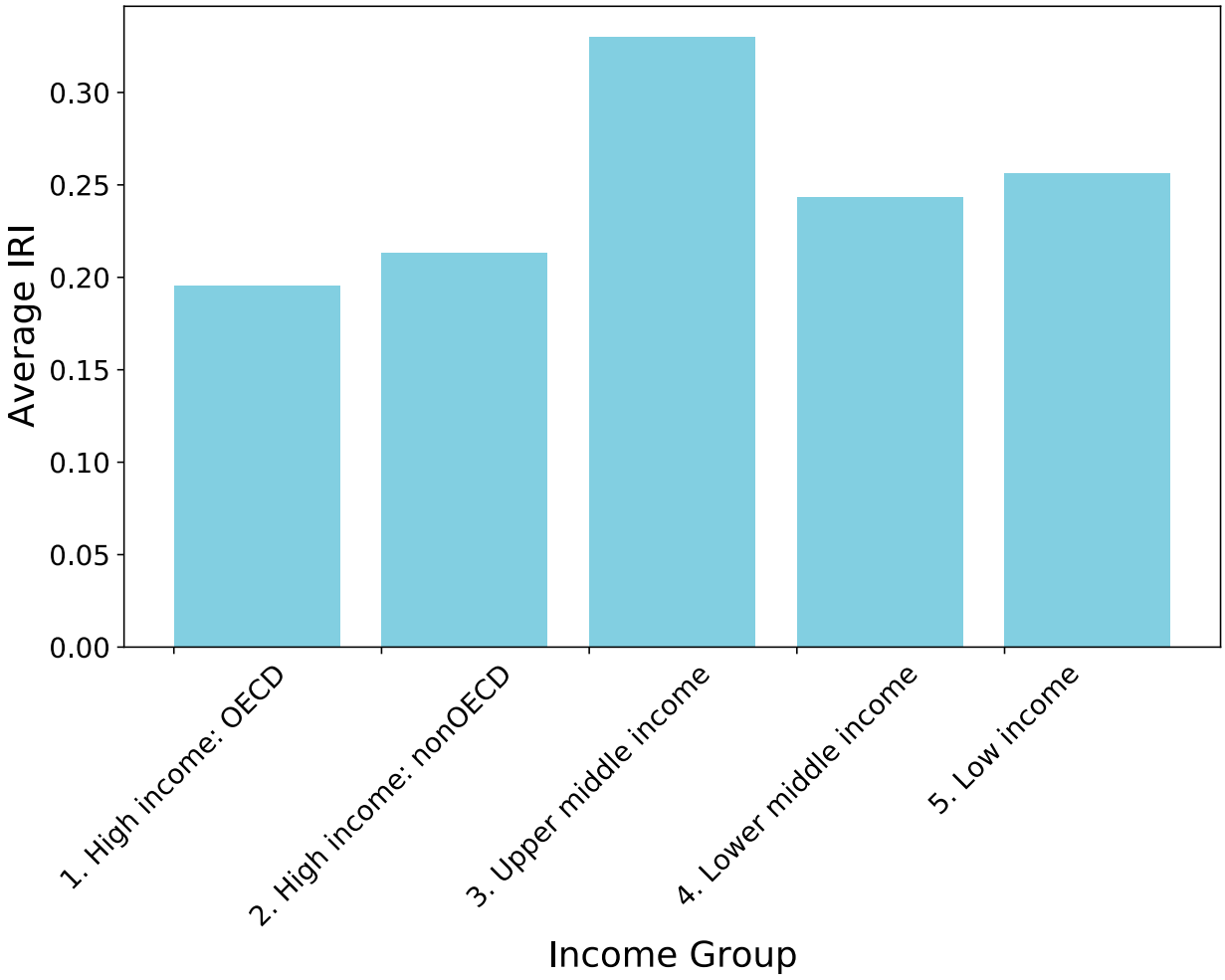
[1]*CoMuNe Lab, Fondazione Bruno Kessler, Via Sommarive 18, Povo, 38123, Trento, Italy*

[2]*IULM University, Via Carlo Bo, 1, 20143 Milan, Italy*

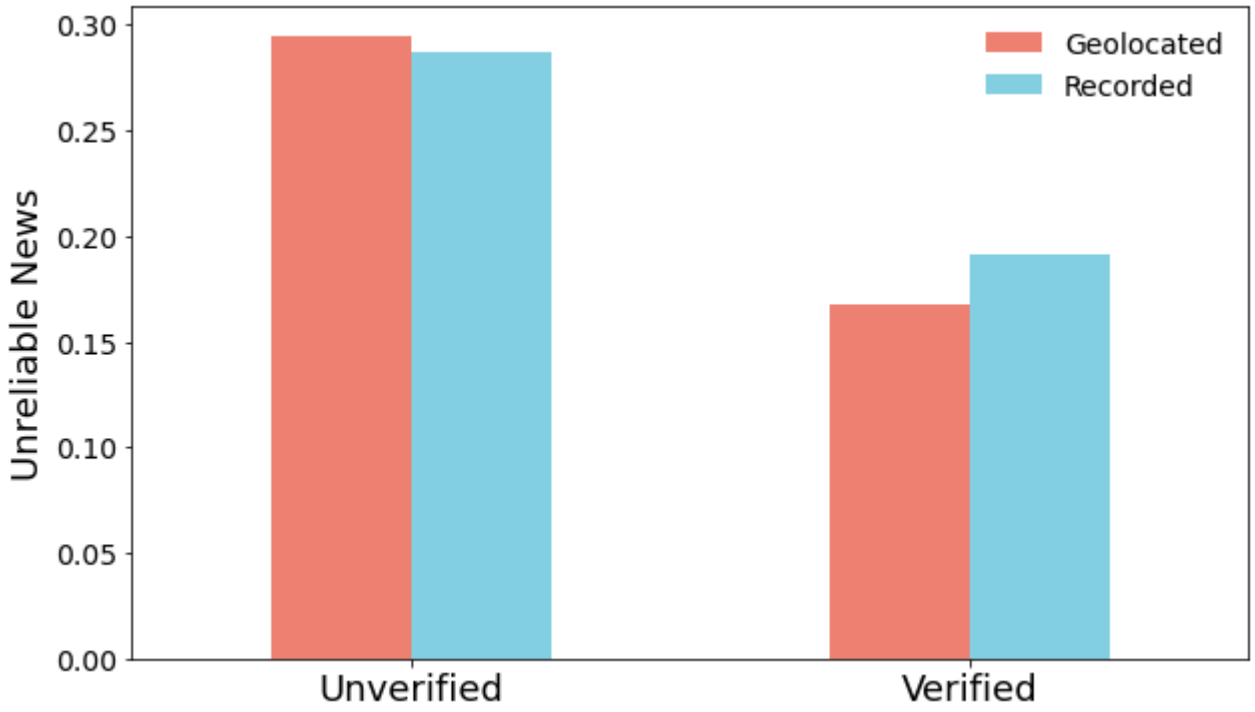
[3]*Berkman-Klein Center for Internet & Society, Harvard University, 23 Everett St 2, Cambridge MA 02138 USA*

[4]*Fondazione Bruno Kessler, Via Santa Croce, 77, 38122 Trento, Italy*

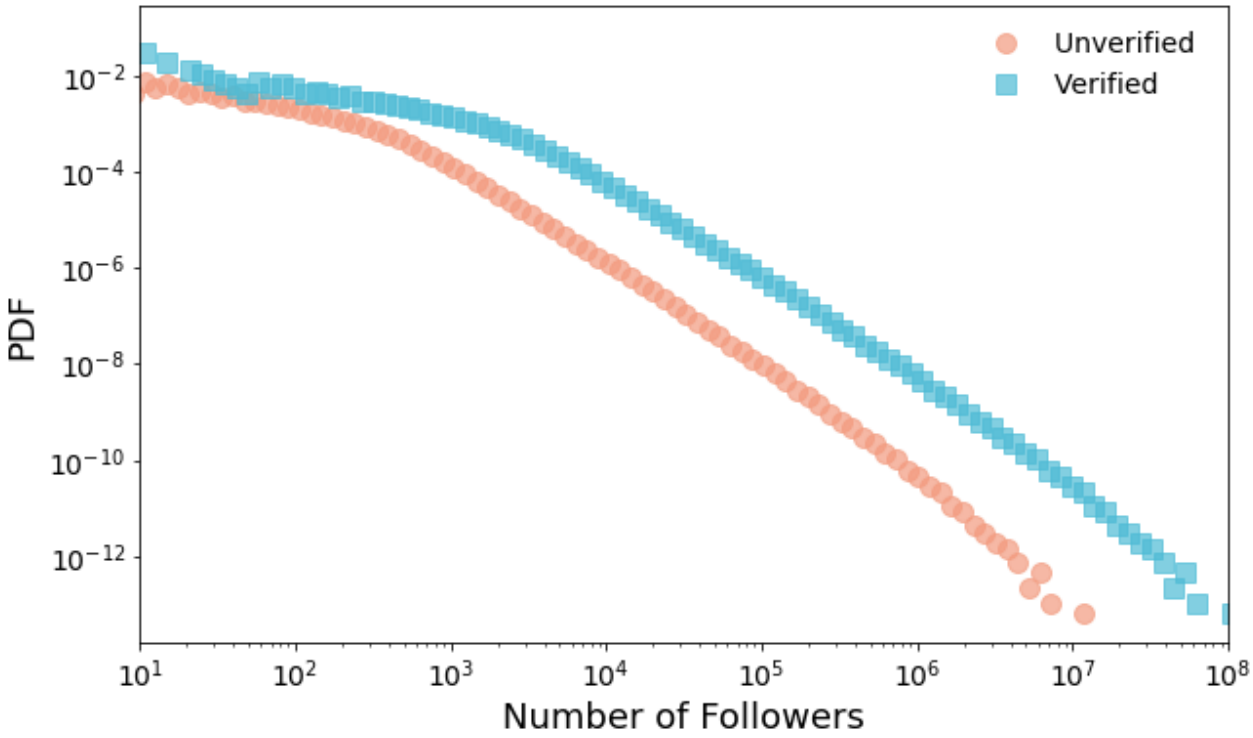
*Corresponding author: mdedomenico@fbk.eu, pierluigi_sacco@fas.harvard.edu



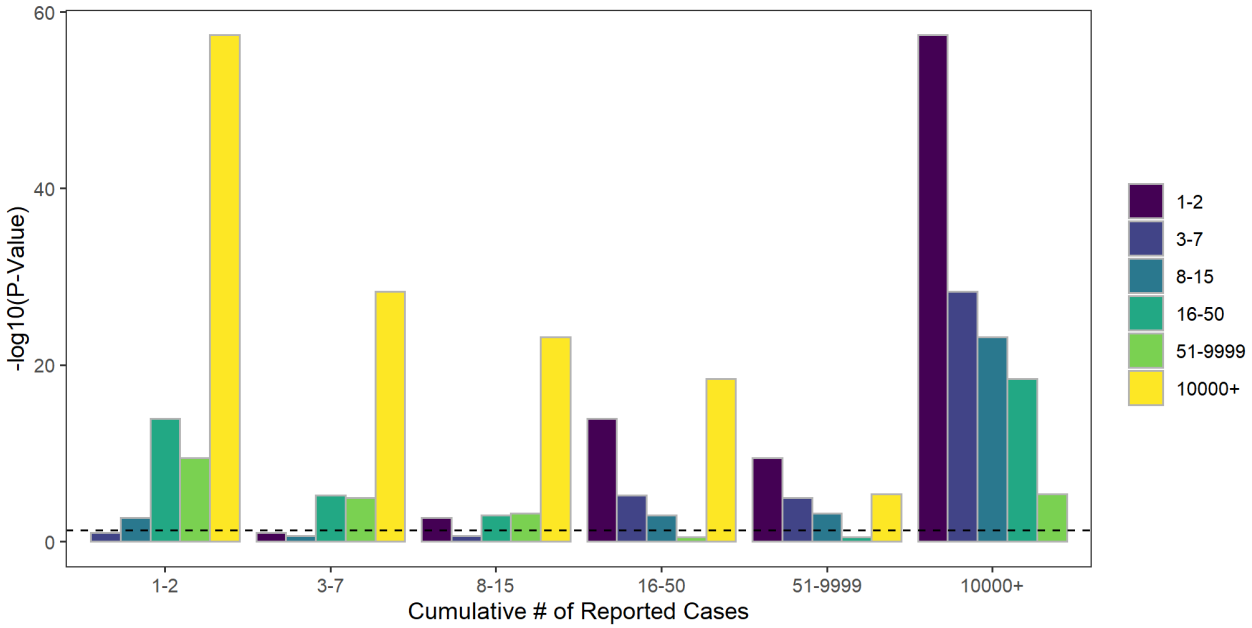
Supplementary Fig. 1: Average Infodemic Risk index for countries with different income levels. Combining IRI values, aggregated at country level from worldwide Twitter data, with the information on country clusters according to national income levels as provided by the Natural Earth map (www.naturalearthdata.com), we find no clear correlation between income levels and diffusion of misinformation. The highest values of the Infodemic Risk Index are found in countries with upper-middle income levels.



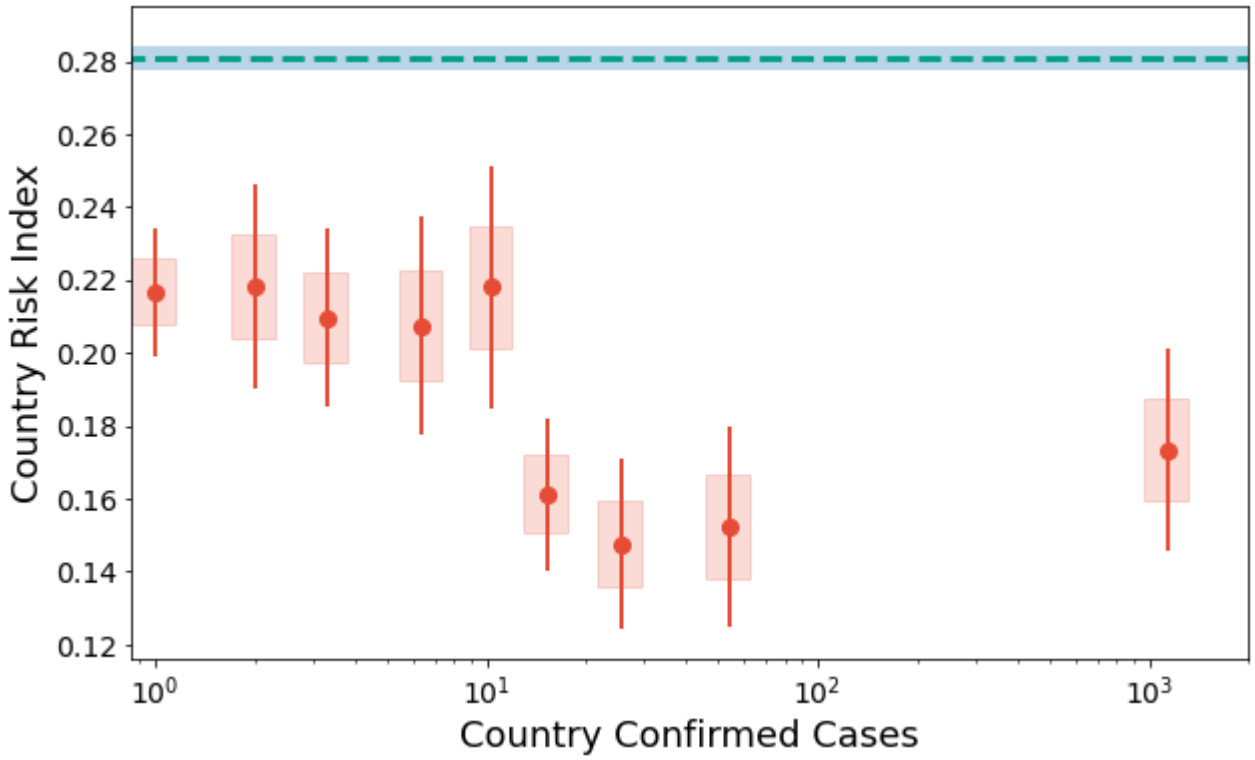
Supplementary Fig. 2: Rates of unreliable news shared by different types of accounts. Unverified users share a larger fraction of unreliable news than verified users. We observe some differences between the values estimated by means of the geolocated data used in this paper (red), vs. the values associated with the whole sample of recorded Tweets (blue). The difference, possibly due to the correlation between the propensity to share unreliable news and the choice of customizing the 'location' field of one's own profile, amounts to a slight overestimate of the volume of unreliable news from unverified accounts, and to a slight underestimate of unreliable news from verified accounts. As verified accounts participate in the IRI more intensively, we can expect that the value of a more accurate IRI estimate which includes all users would be slightly larger.



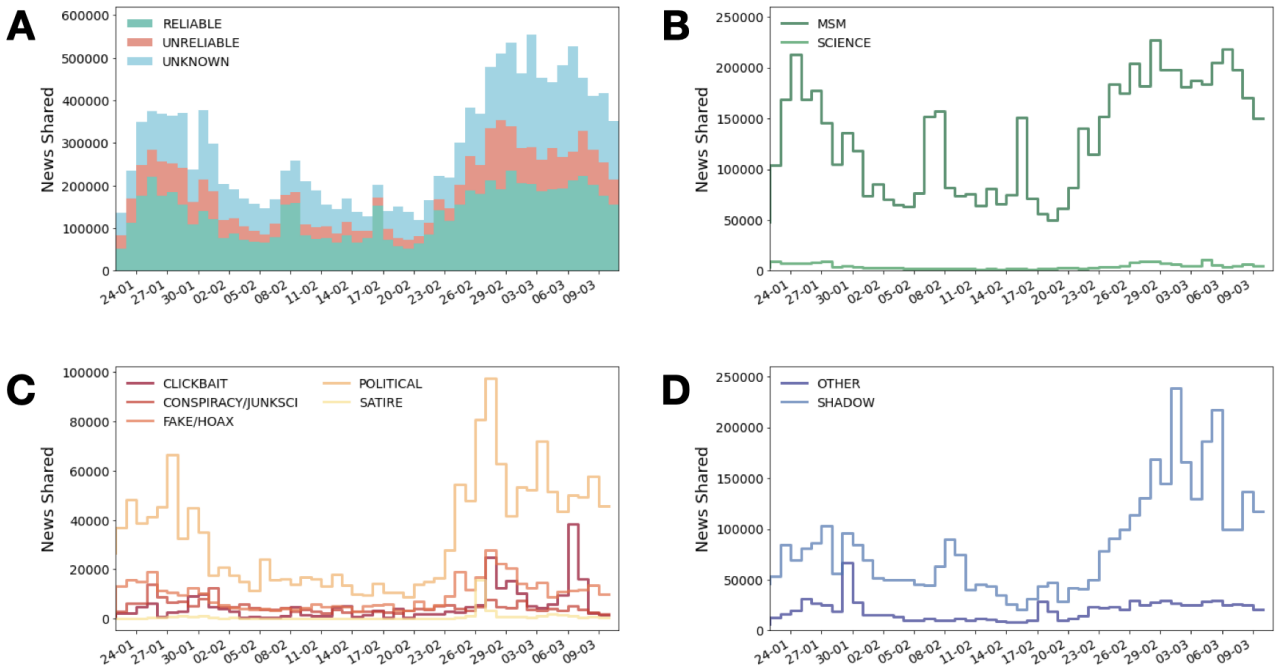
Supplementary Fig. 3: The probability distribution of the number of followers for the two classes of users considered in this study. All distributions display a fat-tail, but different categories of users have a different outreach. Unverified profiles have a significantly smaller number of followers than verified ones. The average values are: 957 for unverified users (circles), and 182k for verified ones (squares).



Supplementary Fig. 4: This figure illustrates the result of the (two tailed) t-test conducted among the different groups represented in Figure 3 of the manuscript. With this test, we check whether the difference between the mean of the cumulative average IRI of a pair of groups (represented by different colors and values in the x-axis) is statistically different from zero. On the y-axis we display the negative logarithm of the resulting p-value. The dotted line represents the value corresponding to a 5% significance level. Values under that threshold mean that there is no evidence that the difference between the mean of two different groups is statistically different from zero. Conversely, values above the threshold indicate that the groups' mean values of cumulative mean IRI are significantly different, confirming the statistical significance of the results discussed in this paper.



Supplementary Fig. 5: A second aggregated view of the evolution of the risk index for increasing numbers of confirmed cases. Differently from Fig. 3, here we compute the average risk index for a single day in each country (instead of the cumulative value). We first compute the average value for days with no confirmed cases, which is 0.28 (blue line in figure). We then aggregate all days with confirmed cases in homogeneous bins and compute the average values (red points). We observe that reporting of the first case is associated with a drop in the risk index, followed by a second significant drop in correspondence of 10 confirmed cases, supporting what was observed in Fig. 3. All shaded areas represent the s.e.m of the average risk index, error bars encode the 95% confidence intervals.

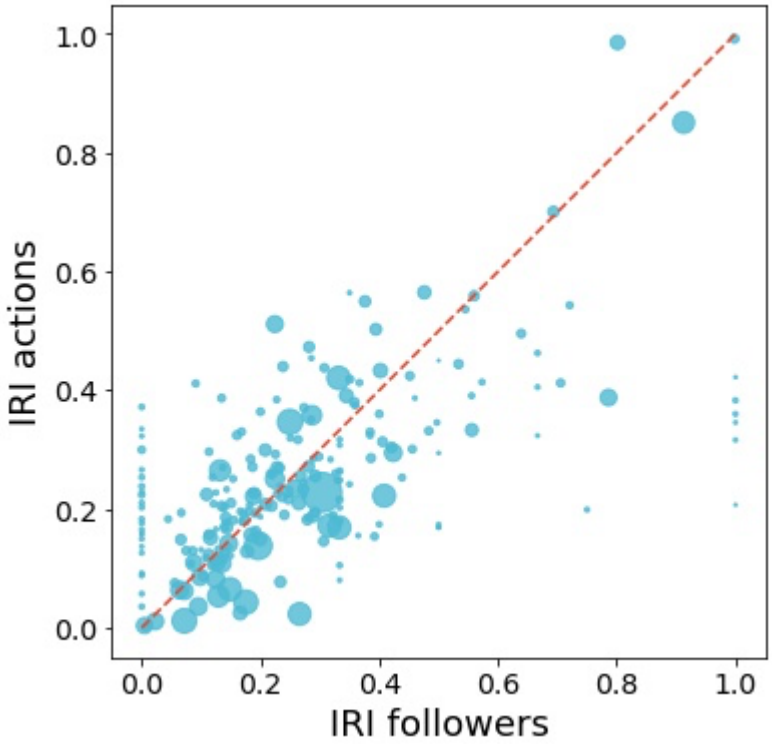


Supplementary Fig. 6: Temporal distribution of news shared on Twitter about COVID19 **A**) Timeseries stratified by the category used in our source reliability rating (see Methods). This analysis demonstrates that reliable sources are more represented than unreliable ones: however, they circulate in different ways and reach different targets, a feature that is captured by the infodemic risk index introduced in this study. **B**) Reliable news includes MSM and SCIENCE. **C**) Unreliable news includes CLICKBAIT, CONSPIRACY/JUNKSCIENCE, FAKE/HOAX, POLITICAL and SATIRE. **D**) Unknown sources include OTHER, i.e. URLs which point to general content (like YouTube videos), and SHADOW, which indicates shortened URLs which could not be fully expanded (e.g., because they pointed to removed web pages).

Supplementary Notes

Suppl. Note 1. Estimate of the robustness of exposure measures

In this paper, we estimated the IRI by using it as a proxy for the exposure to unreliable news experienced by the followers of the user who shared a link to the news content. A possible alternative would be to count the number of actions (retweet, replies, quotes) in response to those tweets. However, that would not be possible in the days following the 25th of February when we reached the 1% limit of the data provided by Twitter, and thus we are not recording all the relevant actions. To verify that our method could be compared with direct measures of exposure, we computed for each country the average IRI for the days between the 22nd of January and the 22nd of February, considering both the ratio of followers reached by unreliable news (as in the paper) and the ratio of actions (retweets, replies, quotes) associated to tweets pointing to unreliable news. As shown in the plot below, where the size of the circle represents the number of tweets in the database for that country, the two measures mostly align. If we consider all 248 countries in our database, we have R^2=0.39, but it goes to R^2=0.61 if we restrict to the 142 countries with at least 1000 tweets.



Supplementary Fig. 7: Comparison of indirect and direct exposure measures.

In more detail, the definition of exposure rests upon an underlying assumption that all the followers of a given account will see the posted message. In the definition given in the manuscript (eqs. 1-

4), one may further introduce a coefficient ϵ to modulate exposure, and assume that only a fraction of the followers will see the message. There is no reason to expect that this parameter ϵ, which is unknown and not directly measurable, will be different for all users. A very similar assumption has been considered plausible in a paper by one of the authors on the analysis of collective attention on Twitter (De Domenico, M., Altmann, E.G. Sci Rep 2020).
However, we would like to point out that this measure is only an intermediate one that is required only to operatively define the IRI: under the above assumption, the unknown coefficient ϵ would be present in both the numerator and denominator, therefore vanishing from the definition of the IRI and, consequently, not affecting its value.
We can further relax the above assumption. In fact, the original definition of exposure is based on the sum:

$$E_i(t, t+\Delta t) = \sum_{u \in C_i} \sum_{m \in M_u(t,t+\Delta t)} K_u$$

And in the more general case it changes into:

$$E_i(t, t+\Delta t) = \sum_{u \in C_i} \sum_{m \in M_u(t,t+\Delta t)} \epsilon_u K_u \approx \langle \epsilon \rangle \sum_{u \in C_i} \sum_{m \in M_u(t,t+\Delta t)} K_u$$

This definition incorporates the fact that the parameter can be different for all users. At the same time, we obtain the approximation by assuming that it is homogeneously distributed across users with a distribution that admits an average value $\langle \epsilon \rangle$: in this case we can use a mean-field approach to estimate the exposure, and it will correspond to rescaling the original definition of exposure by a factor $\langle \epsilon \rangle$. As in the previous scenario, this factor would appear both at numerator and denominator of the IRI, thus vanishing again without affecting our estimations.

The only doubt left at this point might be the assumption of homogeneity in the distribution of ϵ_u. Here, we are supported by the analysis proposed above in this note. Since we have considered the actual impact of a message (measured in terms of social reactions, e.g. replies and retweets), if the assumption of homogeneity was wrong, then different users (with broadly different numbers of followers) should present, on average, very peculiar patterns. Those peculiarities would be reflected in the version of the IRI (let's name it IRI2) calculated by using the number of social reactions instead of the number of followers.
Fortunately, this is not the case: the IRI2 is highly similar and correlated to the IRI (as shown in the scatterplot above), providing evidence that the above assumption, or at least one of its direct effects, is verified.

Suppl. Note 2. Behavioral change in Italy before and after the first domestic case

In principle, many possible factors could be behind the drop in IRI consequent to a local epidemic surge: the suspension of influencers spreading unreliable news; a general shift of the users towards more reliable news; or a takeover of the social conversation by users sharing more reliable news are three such instances.

To understand what really happened, we selected the Tweets associated to 17,900 users who tweeted in Italian language and within the Italian territory, and divided them in two parts: those posted before the 21st of February, and those posted after that date, when the first domestic contagion in the Italian territory was officially registered. In each of these subsets, we label the users with "R" if they only tweeted reliable news, with "U" if they only shared unreliable news, and with "M" if they shared both types of content.

In the Supplementary Tables 1 and 2 below, we show how, among the users who were already tweeting in the first phase, there was a net positive shift from U to R tweeting content in the second phase. Moreover, among users who didn't tweet in the first phase but started tweeting in the second, there was an overwhelming prevalence of R users over U ones.

	TOT	After: U	After: M	After: R	After: not seen
Before: U	1123	180	296	214	433
Before: M	960	56	666	162	76
Before: R	3153	86	447	1238	1382
Before: not seen	12664	1478	1014	10172	-

Supplementary Table 1: Change in type of news shared before and after the 21 February 2020 in Italy.

	TOT	After: U	After: M	After: R
Before	5236	1123 (21%)	960 (19%)	3153 (60%)
After (seen Before)	3345	322 (10%)	1409 (42%)	1614 (48%)
After (not seen Before)	12664	1478 (12%)	1014 (8%)	10172 (80%)

Supplementary Table 2: Aggregate type of news shared by users tweeting before or after the 21 February 2020 in Italy.

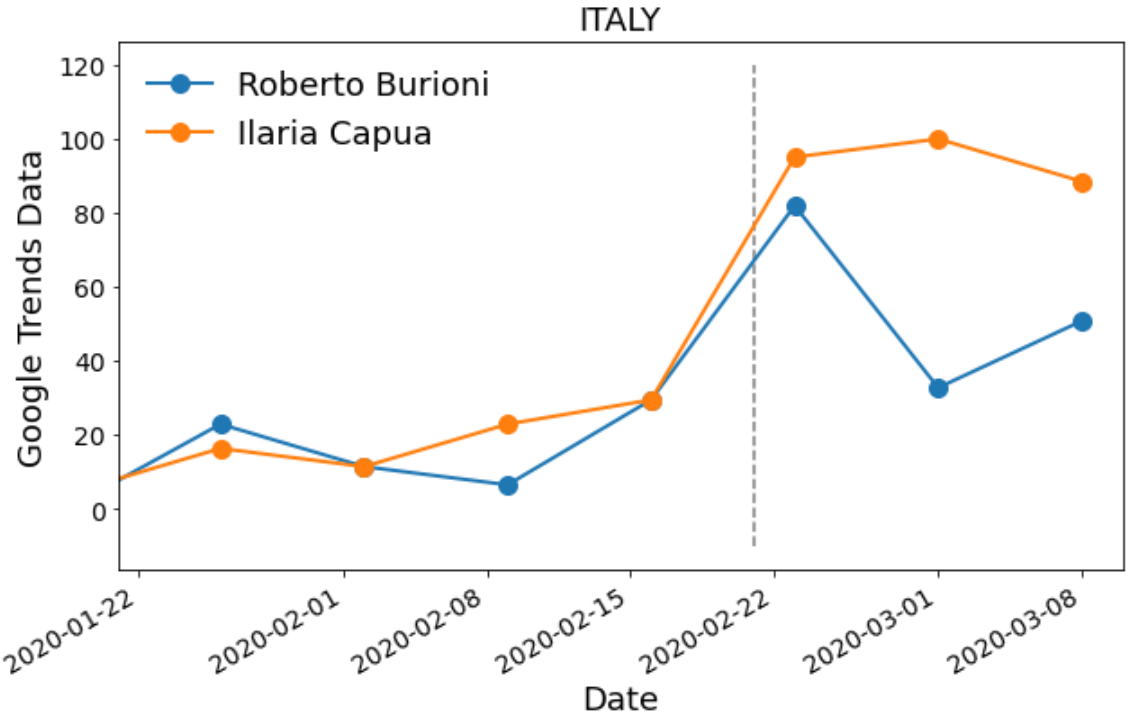
Both before and after the local outbreak of the epidemics, R users produce slightly more tweets per/capita than U users (+18% before, +25% after).

Similarly, both before and after R users have a slightly larger following than U users (difference of +36% before and +28% after, estimated using the log-average of all values >0).

In summary, at least for Italy, the shift is probably not driven by a few accounts, but by a composition of factors: an increased critical awareness by part of those who initially only shared unreliable news, and the entrance into the online discussion by previously unengaged users sharing more reliable content.

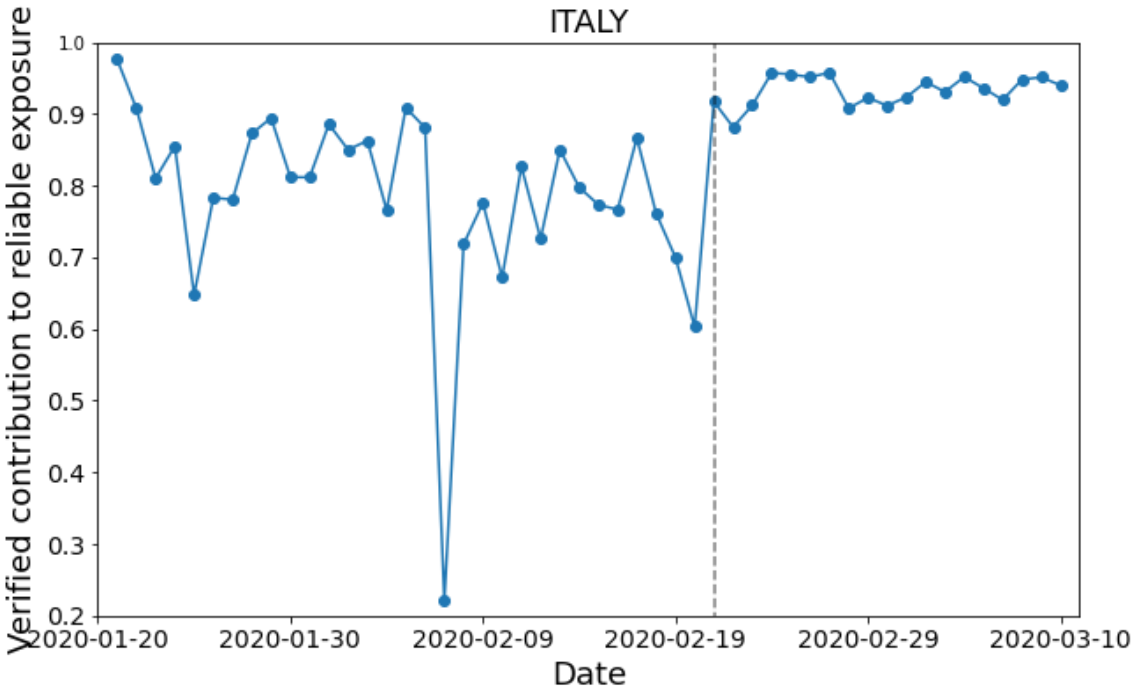
Now, a second question is whether this shift towards reliable content is associated with people actively looking for better sources or if the sources themselves, possibly represented by verified profiles, entered the discussion after the 21st of February.

Since Twitter data do not allow us to track the search for information, we answer here with an example taken from Google Trends. In the graph below, we compare the relative number of online searches for the two most famous Italian virologists, Roberto Burioni and Ilaria Capua.



Supplementary Fig. 8: Rise in Google searches for the two most famous Italian virologists after the first local Covid-19 cases on February 21.

This surge in search activity coincided again with the first domestic cases. Moreover, in Supplementary Fig. 9 we show, among the tweets in Italian and sent from the Italian territory, the fraction of reliable contents from verified accounts.



Supplementary Fig. 9: Fraction of reliable exposure associated with verified accounts in Italy.

Again, starting from the same dates, we have that verified accounts started steadily dominating overall exposure. We add that, as of 20 March 2020, Twitter has been actively verifying users providing reliable updates about Covid-19, but this policy change does not influence our results as it happened outside the range of dates covered in our analysis.

In conclusion, in Italy we observe both a shift of users towards the sharing of reliable news, and the appearance of new users sharing reliable information. These behaviors have possibly been triggered both by users' active searches for more reliable information and by the role of verified influencers.

Suppl. Note 3. Filter terms rationale and estimate of the recall rate

Our analyses do not focus on reconstructing the whole communication network on our research topic but, instead, on estimating the fraction and impact of unreliable news. Therefore, the rationale behind our search terms was to include the most commonly used keywords to ensure that, in case of an abrupt change in the reference terms, we were still tracking the phenomenon effectively. In particular, we have expanded our query to include the official naming selected by the WHO (e.g., from 2019-ncov to COVID-19). In hindsight, using only "coronavirus" might have been sufficient for the sake of our paper, as it is the most common term used in the news and social media to denote the officially named COVID-19 epidemics, and it kept its use across the whole period analyzed while providing a sufficiently large sample.

To estimate the recall rate of our filter of tweets associated to the terms coronavirus, ncov, #Wuhan, covid19, covid-19, sarscov2, covid, we gathered by means of the search API a random sample of 1000 tweets in three languages spoken in our team (Italian, English and Spanish) to manually evaluate the recall rate on the 26th of April at 11:50am UTC. We found the task of drawing the line between what was covid-related and what was not very challenging, and we therefore opted for a 3-values scale:

0 – not COVID-19 related;
1 – any 'doubtful' tweet, mostly related to the social and lifestyle consequences of the epidemics (e.g. personal lockdown and #stayathome comments or rants, ...);
2 – reliabtly associated with COVID-19 (e.g. from a medical, epidemiological, political, or hygiene perspective, with examples ranging from face mask stockage to the alleged responsibilities of Bill Gates).

The results compared with the result of a filter (IN/OUT) using the keywords selected in our paper appear in the confusion matrix (see Supplementary Table 3), where the (+1) numbers indicate tweets we could evaluate by parsing the keywords in URLs strings in a way that is not possible using Twitter's API.

	0	1	2
OUT	822	115(+3)	23(+1)
IN	0	4	22

Supplementary Table 3: Confusion matrix of the three values of manual classification (0,1,2) vs the filter associated with our choice of keywords.

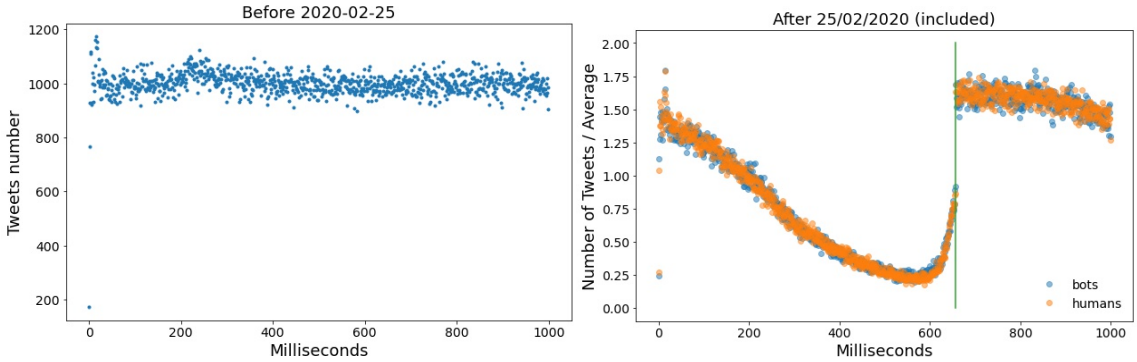
Therefore, our recall ranges between 22/(22+23) = 49% (if we only select tweets strictly associated with the epidemics) and (22+4)/(22+4+23+115) = 16% (if we include the social commentary we labelled as '1').

Since it is possible that this recall changes over time, we made a similar analysis on a set of 500 random tweets in Italian geolocated in Italy, gathered using the stream API on the 24th of February, again at 11:50 (unfiltered data recorded independently). Following the aforementioned criteria, the recall for this set is between 59% and 40%, which exceeds the recall for April both considering all 3 languages and the subset of 333 tweets in Italian (for which the recall ranges between 50% and 22%). This suggests that the recall expected during the range of dates included in our paper (22/01/2020 - 10/03/2020) is expected to be higher than what we find in our April sample, where the prevailing discussion focuses upon the social and lifestyle consequences of lockdown and social distancing.

Suppl. Note 4. How Twitter's Filter API behaves at saturation

We propose an in-depth analysis to better understand the functioning of the Stream API "at saturation" to illustrate how, unlike what is possible for the Search API, attempting to influence the results of stream API at the level of coverage associated with our request would be poorly effective. At the same time, we show that we do not see significant evidence of malicious behavior by bots.

As described in [46] the algorithm Twitter uses for building the sample for the search API selects tweets sent in a fixed range of milliseconds where a bot would have an opportunity of maliciously sneaking in. Due to this, bot-generated content could be over-represented in analyses of data from the search API. In particular, this range of milliseconds is 657-666 for the free 1% sample. The algorithm appears to be different for the stream API. We illustrate this in Supplementary Fig. 10, obtained by associating a random sample of 2 Million status IDs from our database to the corresponding millisecond, 1M before 25/2/2020 and 1M following that date.



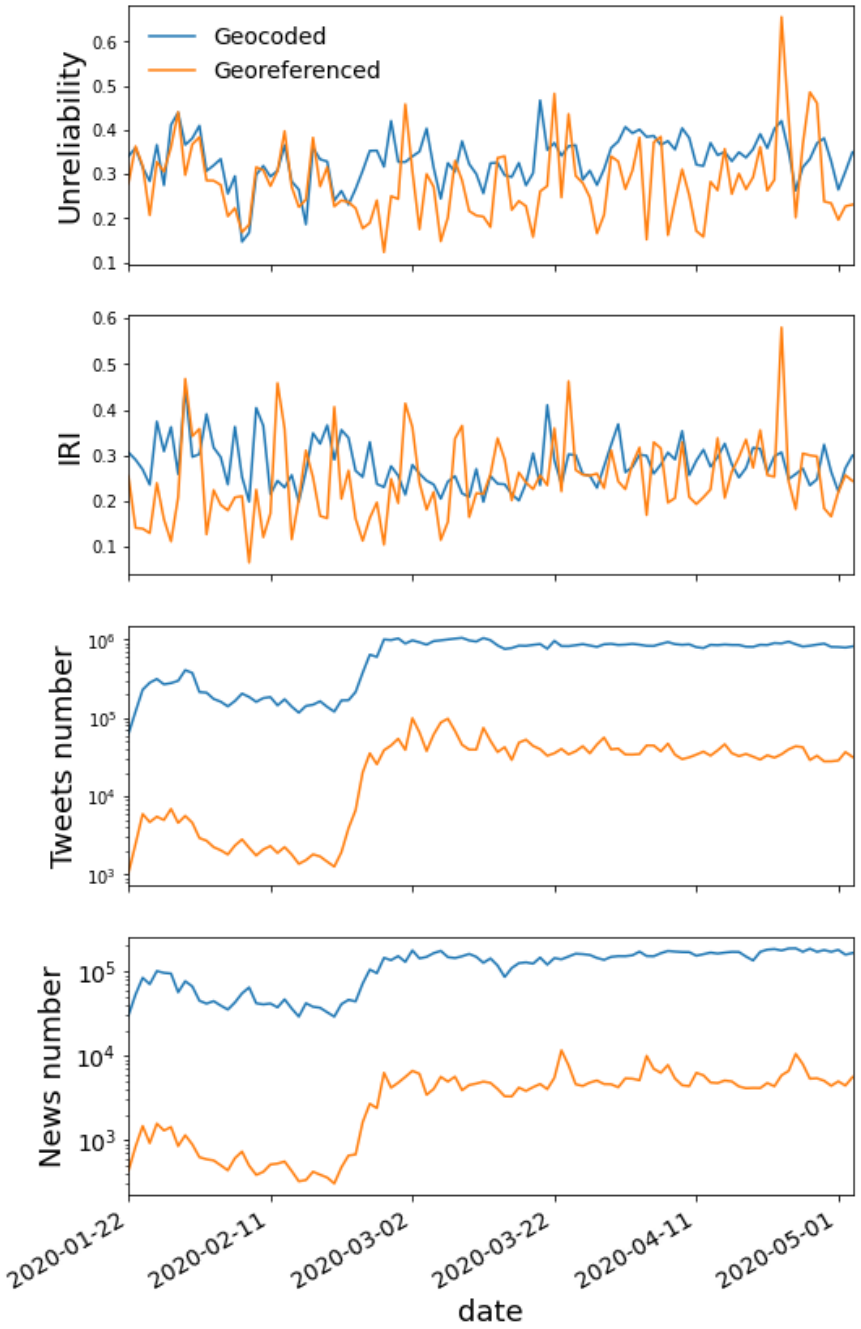
Supplementary Fig. 10: Distribution of the millisecond of the messages provided by the Stream API before and after reaching the 1% threshold on February 25, 2020.

On the left, we see a distribution of the milliseconds for tweets before the date when our stream reached the 1% cap. Here, all tweets pass the filter regardless of the millisecond in which they have been sent.
On the right, we see that indeed some sort of filter based on the millisecond is present when the cap is reached. The plot is the same as the left panel, but we separated the tweets associated with bots and humans and divided the associated number by the average for each subsample so that we could directly compare the two curves. Here we see the peculiar behavior of the stream filter, but at the same time we do not see any particularly focused behavior by bots. The filter indeed cannot be based here over a fixed interval, as for each second it might be sharing a fraction between 100% (if our request is below 1% of the total, as in the left panel) and 1% (if we somehow requested the entire Twitter stream) of the whole stream requested. From the plot, it seems like the filter is active again starting from the 'magic' number 657 (green line) and progressively yielding all the tweets associated with increasing values of milliseconds (modulo 1000) up to reaching the 1% threshold. This makes the filter much broader than the one of the sample API, and consequently harder to efficiently temper with and become over-represented. An optimal

tempering strategy corresponds to the same strategy used for the sample API: focusing on messages in the 657-666 interval. But this interval corresponds here only to the 1.6% of the selected tweets (as opposed to the 100% of the tweets selected in the sample API), reducing the impact of such strategy. For this reason, we are reasonably sure that sampling bias does not influence our results.

Suppl. Note 5. Comparison of time series obtained from geocoded data with those produced using only georeferenced tweets.

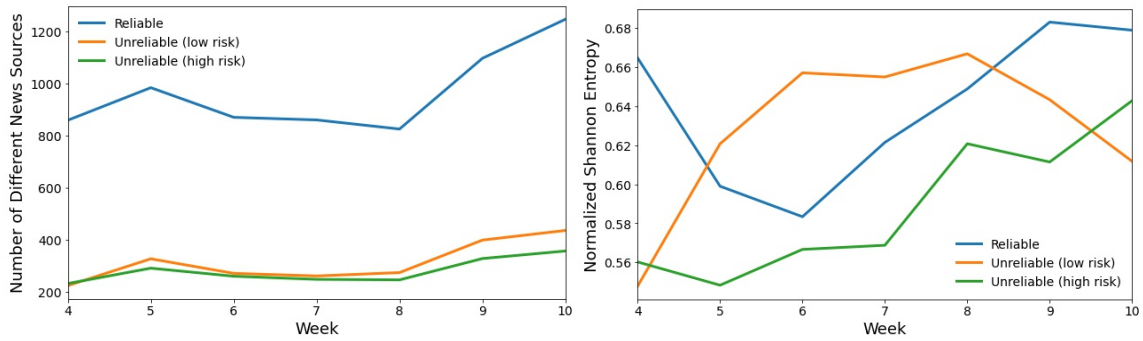
In this paper we analyzed ~60 millions of geocoded messages. However only 0.84% of messages originally included geolocation in the form of latitude-longitude or bounding box. In the majority of cases, what we have is the self-reported geolocation of the accounts responsible for these 60 million messages. To verify that our geocoding was not influencing our results, we compare the IRI computed for the country where we have the larger availability of data, that is the United States of America. In Supplementary Figure 11, we see how the IRI estimated by the large sample of geocoded users and the smaller subsample of messages that were georeferenced align strongly, especially in the last days of sampling when all kinds of data are more abundant, whereas the deviation observed in the early days might be possibly associated to an incorrect evaluation of exposure due to the small sample size of georeferenced data. Indeed, the measure of unreliability (which is simply the fraction of unreliable tweets not weighted by the number of followers) matches perfectly in the two datasets for the early days of the epidemics, and therefore the difference in IRI must be due to an imperfect estimate of exposure in the smaller dataset.



Supplementary Fig. 11: Similarities and differences between the information extracted from geocoded data and from the smaller fraction of georeferenced tweets in U.S.A.

Suppl. Note 6. The shift of misinformation towards untracked domains during the course of our analysis does not affect our results

The list of domains we used to classify reliable/unreliable sources borrows from prior work and might be partially outdated. Indeed, the MediaBiasFactCheck, which is continuously updated, is the source that allows us to label the greatest fraction of domains. To exclude that our observed drop in IRI is linked to the shift of misinformation towards different domains, we evaluated the number and variety of domains shared on twitter across the 7 complete weeks included in our analysis. We do not observe however any decreasing trend in either the number or variety of domains recorded in our dataset. This can be observed in the two plots of Supplementary Fig. 12.



Supplementary Fig. 12: Number of different news sources and Normalized Shannon Entropy for different kind of news.

On the left, we see how the number of different news source domains follow roughly the same growth pattern, which can be easily associated to the growing global attention towards the COVID-19 topic on Twitter. This first, rather trivial, observation is enriched in the right panel by a measure of entropy showing that such increasing number of sources is also associated with a progressively reduced concentration of the tweets on a smaller number of sources, in particular for high risk, unreliable sources (FAKE/HOAX and CONSPIRACY/JUNKSCIENCE). The combination of these two results contrasts with the possibility that our results might be influenced by a progressive loss of coverage in our domain classification during the course of the period of analysis, as the domains grow both in number and variety.

Suppl. Note 7. Further notes on the use of Media Bias Fact Check for the source reliability rating.

Our work has the advantage of having aggregated multiple databases, allowing us to compare available information. Among those, Media Bias Fact Check is the largest but, unlike others, it has not been curated by scientific or media experts. In the following, we provide statistics about the successful comparison between MBFC and other lists published in different years, which support our choice to make use of such resources in this paper.

MBFC shares 392 URLS with the list M. Zimdars, published in **2016**:
https://www.washingtonpost.com/posteverything/wp/2016/11/18/my-fake-news-list-went-viral-but-made-up-stories-are-only-part-of-the-problem/

Among those, 352 (89.7%) are matches in our reliable/unreliable classification, 12 are mismatches (11 of which amount in particular to differences in the POLITICAL/MSM classification) and 28 are 'unknown' in one of the two sources.

MBFC shares 99 URLS with the list of unreliable news websites published by the Poynter institute (**2017**): https://www.politifact.com/article/2017/apr/20/politifacts-guide-fake-news-websites-and-what-they/
97 (98%) of which are matches, and 2 mismatches.

MBFC shares 69 URLS with the list used in the paper by Starbird et al.:

"Ecosystem or echo-system? exploring content sharing across alternative media domains, ICWSM (**2018**)"

58 (84%) of which are matches whereas 8 are mismatches (6 in the POLITICAL/MSM classification, 2 in the CLICKBAIT/MSM classification), while 3 are 'unknown' in one of the two sources.

MBFC shares 78 URLS with the list of unreliable news websites used in the paper by Grinberg et al.:

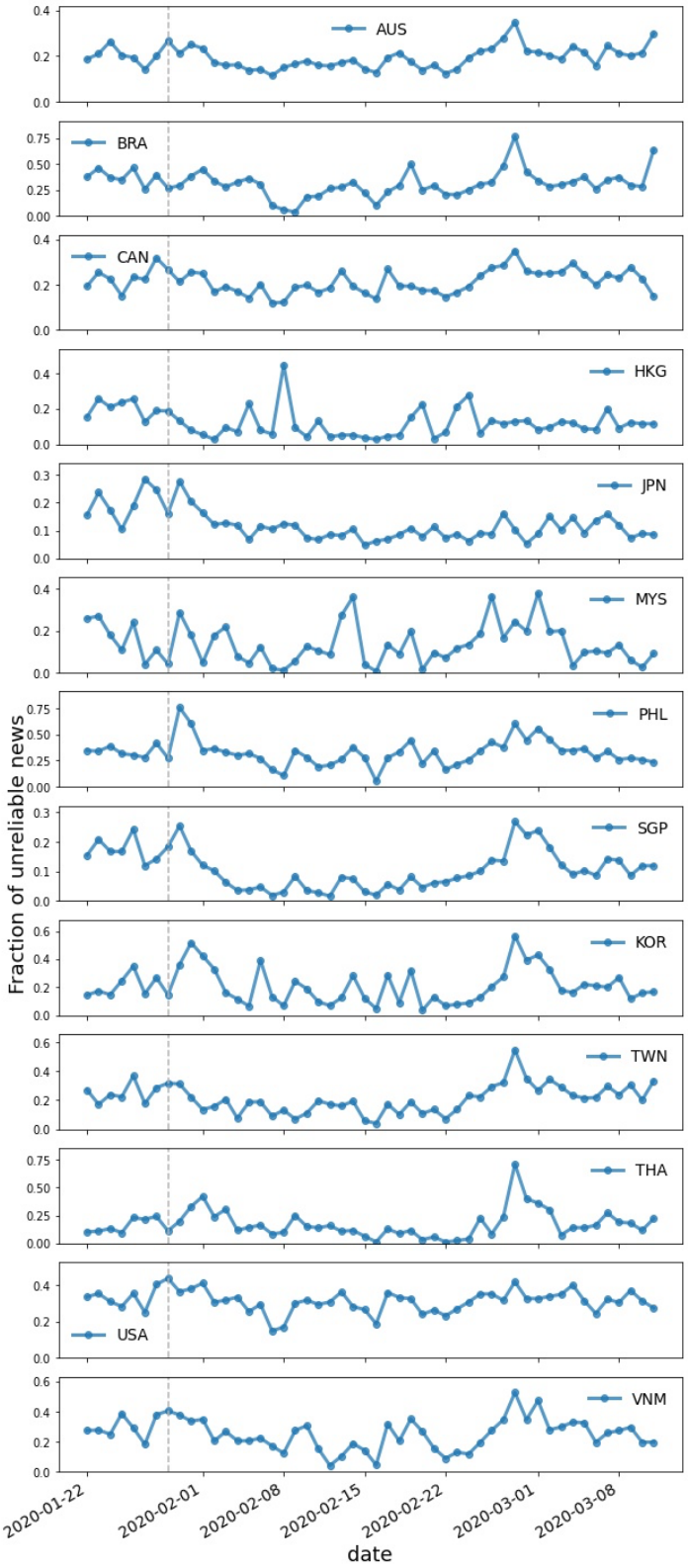
"Fake news on Twitter during the 2016 U.S. presidential election. Science 363, 374 (**2019**)"
76 (97.4%) of which are matches, and 2 mismatches.

Since these comparisons with expert-based classification suggest that the accuracy of MBFC is expected to be reasonably high, we consider it safe to include MBFC in our analysis.

Suppl. Note 8. Twitter's policies to promote authoritative content do not affect our results

During the onset of the infodemics, platforms like Twitter acted to prioritize the visibility of official sources over unreliable ones by introducing special search prompts that return only results from vetted sources, which could fall in the SCIENCE and MSM categories discussed in the Methods section. This shift in priority however does not explain the main findings of the paper. Indeed, we observe no clear shift associated to the changes in Twitter's policies. As indicated in covid19.twitter.com (last visited June 5th 2020), the 29th of January 2020 Twitter *"launched a new dedicated search prompt to ensure that when you come to the service for information about the #coronavirus, you're met with credible, authoritative information first."* They *"partnered with the national public health agency or the World Health Organization"* starting with a selected list of countries (Australia, Brazil, Canada, Hong Kong, Japan, Malaysia, New Zealand, Philippines, Singapore, South Korea, Taiwan, Thailand, United States, Vietnam) as indicated in this tweet: https://twitter.com/TwitterGov/status/1222582416201736192

In the Supplementary Fig. 13, we show how this date (indicated by the dashed line) is not associated with any drastic shift in the number of unreliable news circulating in those countries. The initial list of countries has been progressively expanded, and the 4th of March it was communicated that the list then included 50 countries, but since we cannot be sure about the date when the actual change had been made, we limited our time series analysis to the above mentioned 13 original countries. We finally report that, starting March 20th, Twitter is also actively verifying previously unverified experts, but this date falls outside the range of the analysis considered in our manuscript.



Supplementary Fig. 13: Fraction of unreliable news in the countries where Twitter's new policies were first implemented.