



Assessing the risks of ‘infodemics’ in response to COVID-19 epidemics

Riccardo Gallotti ¹, Francesco Valle¹, Nicola Castaldo ¹, Pierluigi Sacco ^{2,3} and Manlio De Domenico ¹

During COVID-19, governments and the public are fighting not only a pandemic but also a co-evolving infodemic—the rapid and far-reaching spread of information of questionable quality. We analysed more than 100 million Twitter messages posted worldwide during the early stages of epidemic spread across countries (from 22 January to 10 March 2020) and classified the reliability of the news being circulated. We developed an Infodemic Risk Index to capture the magnitude of exposure to unreliable news across countries. We found that measurable waves of potentially unreliable information preceded the rise of COVID-19 infections, exposing entire countries to falsehoods that pose a serious threat to public health. As infections started to rise, reliable information quickly became more dominant, and Twitter content shifted towards more credible informational sources. Infodemic early-warning signals provide important cues for misinformation mitigation by means of adequate communication strategies.

The recent explosion of publicly shared, decentralized information production that characterizes digital societies¹ and in particular social media activity² provides an exceptional laboratory for the observation and study of complex social dynamics³, and potentially functions as a laboratory to understand, test and validate possible solutions to large-scale crises⁴. Pandemics are an instance of such crises, and the current outbreak of COVID-19 may therefore be thought of as a natural experiment to observe social responses to a major threat that may escalate to catastrophic levels and has already managed to seriously affect levels of economic activity and radically alter human social behaviours across the globe. In this study, we show that information dynamics tailored to alter individuals’ perceptions, and potentially their behavioural responses, is associated with a shift of collective attention⁵ towards false^{6,7} or inflammatory⁸ content, a phenomenon named infodemic (that is, an epidemic of information)^{9–12}, sharing similarities with more traditional epidemics and spreading phenomena^{13–15}.

Contrary to what could be expected in principle, this natural experiment reveals that, on the verge of a threatening global pandemic emergency due to SARS-CoV-2 (refs. ^{16–18}), human communication activity is largely characterized by the production of informational noise and even of misleading or false information¹⁹. This generates waves of unreliable and low-quality information with potentially dangerous impacts on society’s capacity to respond adaptively at all scales by rapidly adopting those norms and behaviours that may effectively contain the propagation of the pandemic²⁰. Spreading false or misleading information may prevent the timely and effective adoption of appropriate behaviours and of public health recommendations or measures²¹. Therefore, on the one hand, we face the threats of a pandemic, which spreads in the absence of effective therapies and valid countermeasures and calls for major efforts to model and anticipate the time course of its diffusion¹⁸. On the other hand, we can speak of an infodemic threat²², which proliferates when credible information sources fail to capture the attention and trust of some parts of the public, for whom alternative, low-quality sources are more appealing as they capture more social attention²³, better match their own beliefs or

prejudices²⁴, or sound more convincing, thanks to their typically straightforward messages²⁵.

The appeal of low-quality, misleading or manipulative information relies on simple, effective psychological mechanisms, such as curbing anxiety by denying or minimizing the seriousness of the threat; controlling fear and anger by targeting scapegoat individuals, groups or institutions as the ones responsible for the crisis; and delivering an illusory sense of control through the provision of ‘miracle’ remedies. Similarly to epidemics, infodemics could be thought of as outbreaks of false rumours and unreliable news^{26,27} with unexpected effects on social dynamics (Fig. 1), which can substantially increase epidemic spread. Infodemics call for suitable policy interventions built on state-of-the-art social and behavioural research²⁸.

As shown in Fig. 1, an infodemic is the result of the simultaneous action of multiple human and non-human sources of unreliable or misleading news in times of great abundance of circulating information. Note that, although this study does not directly deal with non-human accounts and their role in (mis-)information diffusion, we include them in the figure because they are known to be important contributors of noise in online social media^{7,8,29–31}. As users are repeatedly hit by a given message from different sources, this works as an indirect validation of its reliability and relevance, leading the user to spread it in turn and to become a vector of dangerously misleading information.

The COVID-19 crisis allows us to provide an evidence-based assessment of such risks and of the real-time interaction of infodemic and epidemic spread¹⁴. We focus our attention on the analysis of messages posted on Twitter³², an online social network characterized by heterogeneous connectivity³³ and topological shortcuts typical of small-world systems³⁴. Information spread on this type of network is well understood in terms of global cascades in a population of individuals who have to choose between complementary alternatives, while accounting for the behaviour and the relative size of the individuals’ social neighbourhoods³⁵, as well as for factors that characterize the popularity of specific content, such as the memory time of users and the underlying connectivity structure³⁶. However, the exact mechanisms responsible for the spread of false

¹CoMuNe Lab, Fondazione Bruno Kessler, Trento, Italy. ²IULM University, Milan, Italy. ³Fondazione Bruno Kessler, Trento, Italy. e-mail: pierluigi.sacco@iulm.it; mdedomenico@fbk.eu

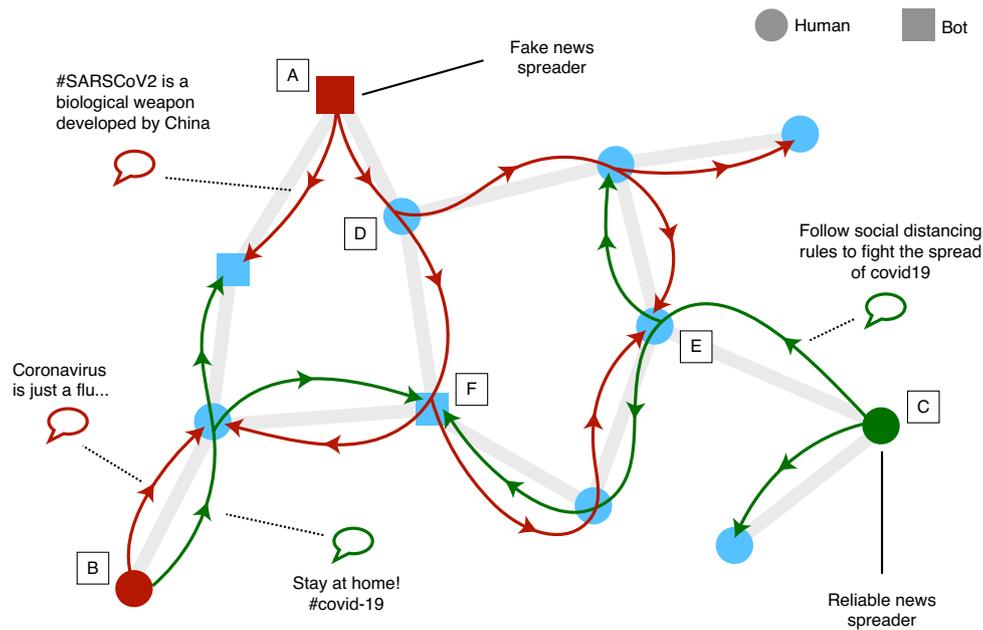


Fig. 1 | How infodemics work. Human (circles) and non-human (squares) accounts participate in the spread of news across a social network. Some users (A and B) create unreliable content, such as false or untrustworthy news or unsupported claims, while others (C) create content informed by reliable sources. When the topic attracts worldwide attention as in the case of COVID-19, the volume of information circulating makes it difficult to orientate oneself and to identify reliable sources. Indeed, some users (D) might be exposed to unreliable information only, while others (E and F) might receive contradictory information and become uncertain as to what information to trust. This is exacerbated when multiple spreading processes co-occur, and some users might be exposed multiple times to the same content or to different contents generated by distinct accounts.

information and inflammatory content, for example during political events^{8,30,37,38}, remain fundamentally unknown. Recently, it has been suggested that this challenging phenomenon might exist because, at a population level, the dynamics of multiple interacting contagions are indistinguishable from social reinforcement³⁹.

This feature reinforces the increasing consensus around the idea that infodemics of news consumption should be analysed through the lens of epidemiology^{9,40} to gain insights about the role of online activities in spreading reliable as well as unreliable news. To this end, we monitored Twitter activity and collected more than 112 million messages using a selection of words commonly used in the medical discourse about COVID-19, between 22 January and 10 March 2020 (see Methods for the details). The messages were in 64 languages from around the world, but because of our data filtering and enrichment procedures, the largest fraction of analysed messages point to English-language sources. As a result, the findings reported in this study mostly capture the behaviour of the English-speaking portion of Twitter users, while in the majority of countries included in our analysis, English is not an officially spoken language. Additionally, Twitter demographics are not representative of the general population—there is overrepresentation of the highly educated, working-age male population. Moreover, limiting the focus to medical terminology clearly narrows the scope of our search and is a further limitation of our work. However, it allows us use terms such as ‘coronavirus’ and ‘covid19’ that are interculturally consistent and used in several languages not depending on local idiomatic usages and variants. We describe in detail the limitations of our dataset in the Discussion and Methods.

Results

Where available, we extracted URLs from messages, collecting approximately 20.7 million links (3.3 million unique) pointing to websites external to the platform. Each URL was then subjected to our source reliability rating method, inheriting the reliability of its source

(Methods, Table 1 and Supplementary Fig. 1). We successfully associated approximately 50% of URLs with a reliability rating by screening almost 4,000 expert-curated web domains; the remaining corpus pointed to disappeared web pages or to content not classifiable automatically (for example, videos on YouTube) and rarely shared sources.

Our method allowed us to overcome the limitations due to text mining of different languages for the analysis of narratives. However, this step in our analysis is predominantly based on sources in English, and this prevents us from covering and representing local discourses that mostly use local languages.

To better understand the diffusion of these messages across countries, we filtered messages that included geographic information. Approximately 0.84% of the collected posts were geotagged by the user, providing highly accurate information about their geographic location. By geocoding the information available in users’ profiles, we were able to extend the corpus of geolocated messages to approximately 50% of the total observed volume (Fig. 2 and Methods). We therefore analysed more than 60 million geolocated messages, containing more than 9 million news links.

For each message, we applied a distinction between verified and unverified users. Usually, verification is performed by the social platform to clearly identify accounts of public interest and certify that they are authentic. The number of followers K_u of a single user u defines the exposure (see Supplementary Note 1 for further details), in terms of potential visualizations at first-order approximation, of a single message m posted by user u at time t . Let $M_u(t, t + \Delta t)$ indicate the set of messages posted by user u in a time window of length Δt . Since there are two different classes of users—verified (V) and unverified (U) accounts—we define the partial exposure (E) due to a single class C_i ($i = V, U$) as

$$E_i(t, t + \Delta t) = \sum_{u \in C_i} \sum_{m \in M_u(t, t + \Delta t)} K_u \quad (1)$$

Table 1 | Description of the nine categories of news in our classification

Category	Harm Score	Description	Count	Type	Total
Science	1	Domains providing content validated via scientific scrutiny.	150	Reliable	1,434
Mainstream media	2	Domains providing content that is generally subjected to professional fact checking and abides by the rules of media accountability.	1,284		
Satire	3	Domains providing content that is intentionally and explicitly aiming at providing a distorted representation of events as a form of humour and/or social critique.	177	Unreliable	2,264
Clickbait	4	Domains providing content that generally distorts or intentionally misrepresents information to capture attention.	47		
Political	7	Domains providing content that presents a partisan representation and interpretation of facts to support a political position over rival ones.	697		
Fake or hoax	8	Domains providing manipulative and fabricated content with the purpose of misleading public opinion on socially relevant issues and provoking inflammatory responses.	917		
Conspiracy and junk science	9	Domains providing systematically manipulative and fabricated content with the purpose of legitimizing implausible conceptualizations of facts and knowledge through argumentative methods that coarsely mimic those of scientific reasoning but without any sound logical or factual basis, targeting individuals or social groups as covert instigators or perpetrators of harmful actions.	426		
Other	5	Domains pointing to general content that cannot be easily classified, such as videos on YouTube.	160	Unknown	194
Shadow	6	Domains related to URL shortening that cannot be classified a priori but would require further URL expansion.	34		

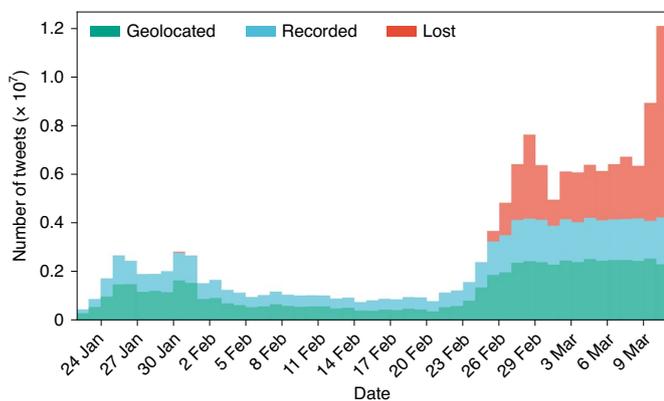


Fig. 2 | The evolution of Twitter activity about the COVID-19 pandemic.

We observe a first increase in collective attention after the outbreak in Wuhan, China (between 24 January and 2 February 2020), and a second strong rise after the epidemics began to spread in northern Italy (20 February 2020 onwards). The fraction of geolocated messages (messages with shared locations, or geonamed, indicated in green) is constantly approximately 50% of the total volume recorded (indicated in blue). From 26 February, we reached the limit of the fraction of data shared by Twitter (Methods), missing an increasing fraction of Tweets (indicated in red).

Note that different users of the same class might have overlapping social neighbourhoods: those neighbours might be reached multiple times by the messages coming from distinct users of the same class; therefore, our measure of exposure accounts for this effect. Note also that our measure provides a lower bound to the number of exposed users, because we do not track higher-order transmission pathways: a user might read a news item included in a message but not share it further. There is no way to account for such users.

The assumption that all followers of a specific user u will be reached by posted messages is clearly unrealistic. In Supplementary Note 1, we provide a mathematical extension of the definition of exposure from equation (1), which allows one to relax this assumption on the basis of a recent study⁷ and a mean-field model, without altering the quantitative analysis presented in this study.

Finally, for each message, we identified the presence of links pointing to external websites, and for each link, we verified whether it came from a trustworthy source or not (Methods). The reliability r_m of a single message m is either 0 or 1, because we discarded all web links that could not be easily assessed (such as ones shortened by third-party services) or that pointed to external platforms (such as YouTube) where it is not possible to automatically classify the reliability of the content. The news reliability of messages produced by a specific class of users (R_i) is therefore defined as

$$R_i(t, t + \Delta t) = \sum_{u \in C_i} \sum_{m \in M_u(t, t + \Delta t)} r_m \tag{2}$$

Unreliability can be defined similarly by replacing r_m with $1 - r_m$. Exposure and reliability are useful descriptors but do not fully suffice to assess the risk of infodemics. For this reason, we developed an Infodemic Risk Index (IRI), which quantifies the rate at which a generic user is exposed to unreliable news produced by a specific class of users (partial IRI, equation (3)) or by any class of users (IRI, equation (4)):

$$pIRI_i(t, t + \Delta t) = \frac{\sum_{u \in C_i} \sum_{m \in M_u(t, t + \Delta t)} K_u (1 - r_m)}{\sum_i E_i(t, t + \Delta t)} \tag{3}$$

$$IRI(t, t + \Delta t) = \sum_i pIRI_i(t, t + \Delta t) \tag{4}$$

Both indices are well defined and range from 0 (no infodemic risk) to 1 (maximum infodemic risk). Note that we can calculate

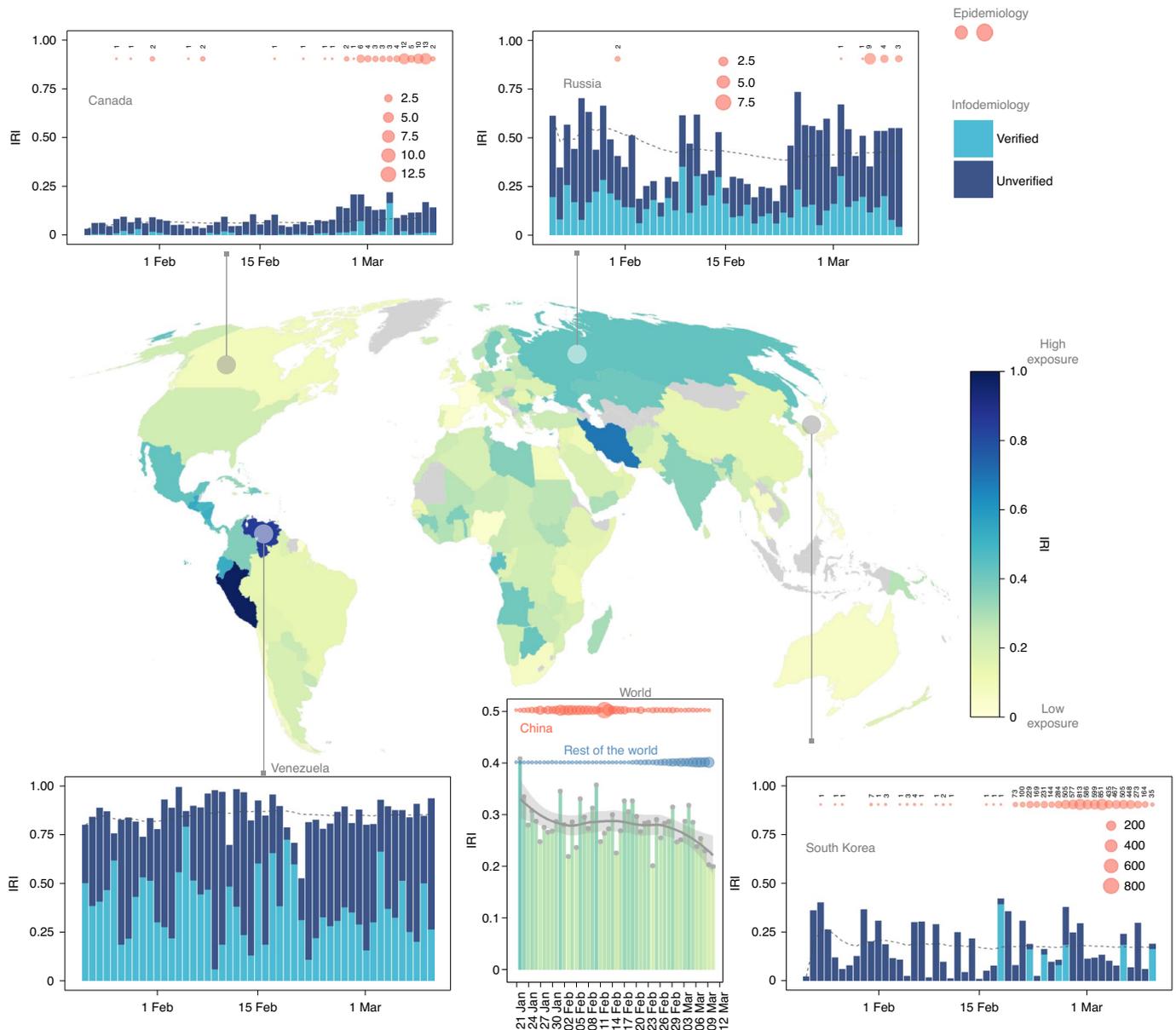


Fig. 3 | Mapping infodemic risk worldwide. The infodemic risk of each country, aggregated over time, is colour-coded on the map. The panels show the evolution of risk over time for a sample of countries; the bars indicate the partial contributions of verified and unverified users to the overall risk and the dashed lines represent the cumulative mean of the IRI at a given day d (computed as the ratio between the cumulative sum of the daily IRI in the days between 22 January and d , and the number of days between these two dates). Risk evolution for the whole world is also shown, demonstrating an overall decrease of risk over time (bottom middle panel, where the grey line represents a LOESS regression with $R^2 = 0.29$). The markers horizontally aligned at the top of each panel indicate the daily confirmed epidemiological cases, with their number encoded by the markers' sizes (Venezuela does not contain epidemiological markers as no confirmed cases were reported at the time of the analysis). Map made with public domain Natural Earth data.

all the infodemic descriptors introduced above at a desired level of spatial and temporal resolution.

Figure 3 shows how countries characterized by different levels of infodemic risk present very different profiles of news sources, which appear not to be strictly correlated with the level of socio-economic development (Supplementary Fig. 2). In low-risk countries such as Canada and South Korea, the level of infodemic risk remains small throughout the period of study, apart from isolated spikes mostly associated with unverified sources. As the epidemic spreads to important levels, infodemic risk further decreases, signalling an increasing focus of the public towards reliable news sources. By contrast, in a high-risk country such as Venezuela, the infodemic

is pronounced throughout the period of observation, and in addition to the expected activity from unverified sources, even verified sources contribute to a large extent to the infodemic. Finally, in a relatively high-risk country such as Russia, infodemic risk is erratic with sudden, very pronounced spikes, and again verified sources play a major role. Here too, information about the epidemic is fragmented and mostly unreliable.

Overall, the global level of infodemic risk tends to decrease as COVID-19 spreads globally, suggesting that epidemic spread leads people to look for relatively more reliable sources. It also suggests that verified influencers with many followers started to point to more reliable news (Supplementary Figs. 3 and 4 and

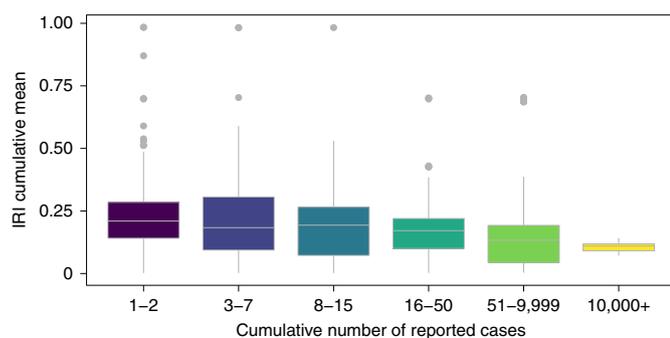


Fig. 4 | Reduction of infodemic risk after COVID-19 reaches countries.

Aggregated view of the evolution of the IRI for increasing numbers of reported cases. For each day and each of the 162 countries considered in our analysis, we compute the cumulative mean of the IRI at a given day d (computed as the ratio between the cumulative sum of the daily IRI in the days between 22 January and d , and the number of days between these two dates). We aggregate days and countries with a similar cumulative number of reported cases, using bins of increasing size to compensate for the limited number of countries that reached high levels of contagion at the time of the analysis and reporting the average value on the x axis. This allows us to describe the drop in IRI as the number of cases grows in a country using box plots. In box plots, the centre lines represent the medians, the boxes the range between the 25th and 75th percentile, and the whiskers the range between the smallest and largest data point, excluding outliers, which are represented as circles. Therefore, the difference between two boxes is statistically significant when each middle line lies outside of the other box. On the basis of the results of both a one-way ANOVA (F statistic (degrees of freedom), 18.86 (5); $P < 0.001$; effect size, $F = 0.05$; 95% confidence interval, (0.03, 0.06); the data distributions were assumed to be normal, but this was not formally tested) and Kruskal–Wallis rank sum tests $F 137.14$ (5); $P < 0.001$; effect size, $F = 0.0677$; 95% confidence interval, (0.0501, 0.0918); no assumptions are needed to use this non-parametric test), there is evidence of a statistically significant effect ($P < 0.001$ for both tests) of the number of reported cases on the IRI cumulative mean. In Supplementary Fig. 4, we provide further tests illustrating the significant difference between each pair of boxes except pairs 3–7 with 1–2 and with 8–15 and pair 16–50 with 51–9,999, where the differences are not statistically significant.

Supplementary Note 2), possibly shifting the state of the infodemic towards a clearer information landscape where it is easier to orientate and to identify unreliable facts.

In the case of Italy, where the epidemic struck the country heavily within the window of observation of the current study, we observe in coincidence with the first verified domestic contagions a sudden, clear increase in national Google searches for the best-known Italian virologists as they gained substantial visibility on national mainstream media (Supplementary Note 2). Our data do not allow us to establish a causal relationship between the sudden increase in popularity and media exposure of such experts and the shift in focus from unreliable to reliable sources in online social media conversations. However, it is likely that a spillover effect has occurred, contributing at least partly to this shift, as Italian Twitter is known to be very reflective of trending personalities and topics from the mainstream media⁴¹. This overall pattern, linking the local spread of the epidemics to the diffusion of more reliable information, is confirmed in terms of measures of infodemic risk aggregated daily and at the country level (Fig. 4 and Supplementary Figs. 5 and 6). This pattern is particularly pronounced with the escalation of the epidemic, suggesting that the effect could be mediated by levels of perceived social alarm.

In principle, countries with high infodemic risk could also present more reliability issues in terms of reporting of epidemic data,

thus altering the perceptions of the public and indirectly misleading them in their search for reliable information. In fact, there have been cases of countries with high infodemic risk where political leaders have actively spread misleading information and openly questioned the necessity to accurately track and measure the development of the epidemic diffusion, as well as the reliability of fact-checking sources^{42–45}. Our results, though, do not provide direct supporting evidence for this possibility, and this remains an open question for future research.

The dynamic profiles of infodemic risk in countries with similar risk levels may also be very different. Figure 5 compares Italy with the United States. In the case of Italy, the risk is mostly due to the activity of unverified sources, but we notice that with the outbreak of the epidemic, the production of misinformation collapses, and there is a sudden shift to reliable sources. In the United States, misinformation is mainly driven by verified sources, and it remains essentially constant even after the epidemic outbreak. Notice also how infodemic risk varies substantially across US states. As in our time window the United States lagged widely behind Italy in terms of epidemic progression, it remains to be seen whether a similar readjustment can be observed for the United States later on. Figure 5 shows, however, that the relationship between the reduction of infodemic risk and the spread of the epidemic seems to be a rather general trend, as the relationship between the number of confirmed cases and infodemic risk is (nonlinearly) negative, confirming the result shown in Fig. 4. Figure 5 also shows how the evolution of infodemic risk among countries with both high message volume and considerable epidemic contagion tends to be very different. The IRI maintained its relatively high level not only in countries such as Iran but also in the United States, Germany, the Netherlands, Sweden and Norway. Conversely, in other countries such as Italy, South Korea and Japan, the IRI substantially dropped with the progression of the epidemics.

Discussion

Our findings show that, in a highly digital society, the epidemic and the infodemic dimensions of COVID-19 co-evolve. The infodemic dimension is driven by a heterogeneous set of actors who pursue largely undisclosed goals.

Given the lack of pharmacological interventions to combat COVID-19, responsible behaviours driven by reliable information at all scales are key for the mitigation of adverse effects. It may therefore be important to develop integrated public health approaches, where the biological and informational dimensions of an epidemic are equally recognized, taken into account and managed through careful policy design.

Here, we have shown that in the context of the COVID-19 crisis, complex infodemic effects are indeed at work, with remarkable variations across countries, and the level of socio-economic development is not the key discriminant to separate countries with high versus low infodemic risk. In fact, we find that there are G8 countries with remarkable infodemic risk (for example, Russia and Germany) and developing countries with far lower risk levels (for example, Thailand and the Philippines). This means that, especially in countries where infodemic risk is high, the eventual speed and effectiveness of the containment of COVID-19 could depend on a prompt policy switch in communication strategies and in the effective countervailing of the most active sources of unreliable news. The escalation of the epidemics leads people to progressively pay attention to more reliable sources, thus potentially limiting the impact of infodemics, but the actual speed of adjustment may make a major difference in determining the social outcome (and in particular between a controlled epidemic and a global pandemic).

Our study is characterized by important limitations. A key limitation of any data collection from social media content is that each social medium has a specific demographic that is not representative

representative data worldwide relying only on online sources, yet the collection of offline sources presents other substantial limitations. In fact, before the advent of social media, it would have been unthinkable to carry out analyses of social phenomena at this scale in real time. Our focus on Twitter means that our reference population tends to be highly educated, working age and male, and our filter selection and source reliability database exacerbate this bias towards English-speaking users. One way to tackle this problem in future research is to extend data collection to several social media platforms at once, but there is a clear trade-off between intensively collecting large volumes of data on a single platform and extensively collecting data from multiple platforms with smaller volumes for each. Moreover, joint collection from multiple biased sources remains biased in principle, although the overall bias becomes less controllable. We consider our approach as a first step, with clear limitations, which may provide a benchmark for more comprehensive future approaches.

There are several important questions and goals for future research. We highlight four: (1) a better understanding of the role of artificial agents (bots) in infodemics, (2) the development of truly multilingual corpora and source reliability databases, (3) the extension of text mining to multiple social media platforms while maintaining the highest possible volumes of mined content from each source, and (4) building a representative sample of the global population through a suitable integration of online and offline sources. These are formidable challenges, but their urgency and relevance do not need much argumentation. We look forward to the future developments of what promises to be an emerging discipline with key theoretical and policy implications.

Methods

Data collection. We followed a methodology for collecting social media data consolidated over the years. We focused on Twitter, which is well known for providing access to publicly available messages upon specific requests through their application programming interface (API). We identified a set of hashtags and keywords gaining collective attention since the first recorded cases of COVID-19: coronavirus, ncov, #Wuhan, covid19, covid-19, sarscov2 and covid. This set includes the official names of the virus and the disease, including the early tentative ones, as well as the name of the city where the first cases of COVID-19 were recorded. We estimate the recall rate for these keywords to be higher than 16% and probably in the 40%–60% range at the time of recording (see Supplementary Note 3 for more details). We used the Filter API to collect the data in real time from 24 January 2020 to 10 March 2020 and the Search API to collect the data between 21 January 2020 and 24 January 2020. Our choice allowed us to monitor, without interruptions and regardless of the language, all the tweets posted about COVID-19 since 22 January 2020, when China reported more than 6,000 cases, calling for the attention of the international community. The Stream API has the advantage of providing all the messages satisfying our selection criteria and posted to the platform in the period of observation, provided that their volume is not larger than 1% of the overall (unfiltered) volume of posted messages. Above 1% of the overall flow of information, the Filter API provides a sample of filtered tweets and communicates an estimate of the number of lost messages. Note that this choice is the most reliable to date: in fact, it was recently shown that biases affecting the Sample API (which samples data on the basis of rate limits), for instance, are not found in the REST and Filter APIs⁴⁷. In Supplementary Note 4, we show how this problem does not affect our data.

We estimate that until 24 February 2020, we lost approximately 60,000 tweets out of millions, capturing more than 99.5% of all messages posted (Fig. 2). The global attention towards COVID-19 increased the volume of messages after 25 February 2020; however, Twitter restrictions allowed us to get no more than 4.5 million messages per day, on average. We have estimated a total of 161.2 million tweets posted until 10 March 2020; we have successfully collected 112.6 million of them.

Geocoding. The user's self-declared location field was used for geocoding with ArcGIS API. For approximately 56% of users, we had a response in terms of latitude and longitude. However, a large portion of these answers (about 10%) were associated with a small number (~1,600) of wrongly attributed locations that were removed (reaching the 50% ratio indicated in the main text). These errors were mostly caused by the use of non-toponyms in the location field such as 'Home' or 'Somewhere', or other pieces of information (such as Instagram and website URLs), which were wrongly associated with real locations. We identified these errors by isolating single locations associated with a large number of different unique

user-defined location strings. Finally, we also filtered out names of continents that were correctly geocoded but do not match the country-based granularity we set for our analysis. The reliability of our method was tested by comparing geocoded and georeferenced data for the United States (Supplementary Note 5).

Source reliability rating. We collected manually checked web domains from multiple publicly available databases, including scientific and journalistic ones. Specifically, we considered data shared by the sources listed in refs. ^{48–56}.

The databases adopted different labelling schemes to classify web domains. We therefore first had to develop a unifying classification scheme, reported in Table 1, and map all existing categories into a unique set of categories. Note that we have also mapped those categories into a coarse-grained classification scheme, distinguishing between reliable and unreliable.

We found a total of 4,988 domains, reduced to 4,417 after removing hard duplicates across databases. Note that a domain is considered a hard duplicate if its name and its classification coincide across databases.

A second level of filtering was applied to domains that are classified differently across databases (for example, xyz.com might be classified as FAKE/HOAX in a database and as SATIRE in another database). To deal with these cases, we adopted our own classification method by assigning to each category a Harm Score (HS) between 1 and 9. When two or more domains were soft duplicates, we kept the classification with the highest HS, as a conservative choice. This phase of processing reduced the overall database to 3,920 unique domains.

The HS classifies sources in terms of their potential contribution to the manipulative and misinformative character of an infodemic. As a general principle, the more systematic and intentionally harmful the knowledge manipulation and data fabrication, the higher the HS. "Science" or "Scientific" content has the lowest level of HS due to the rigorous process of validation carried out through scientific methods. "Mainstream media" content has the second lowest level of HS due to its constant scrutiny in terms of fact checking and media accountability. "Satire" is an unreliable source of news, but due to its explicit goal of distorting or misrepresenting information according to specific cultural codes of humour and social critique, it is generally identified with ease as an unreliable source. "Clickbait" is a more dangerous source (and thus ranks higher in HS) due to its intent to pass fabricated or misrepresented information for facts, with the main purpose of attracting attention and online traffic (that is, for mostly commercial purposes), but without a clear ideological intent. "Other" is a general-purpose category that contains diverse forms of (possibly) misleading or fabricated content, not easily classifiable but probably including bits of ideologically characterized content pursuing systematic goals of social manipulation, and thus ranking higher in HS. "Shadow" is a similar category to the previous one, where links are anonymized and often temporary (for example, bit.ly and dlvr.it), thereby adding an extra element of unaccountability and manipulation that translates into a higher level of HS. Known vanity URL shorteners such as usnytl.ms for the *New York Times* and wpo.st for the *Washington Post* are automatically associated with the source. "Political" is a category where we find an ample spectrum of content with varying levels of distortion and manipulation of information, also including mere selective reporting and omission, whose goal is to build consensus on a polarized political position against others; this category therefore directly aims at conditioning the public discourse and opinion making, with a higher HS than the previous categories. The majority of web domains listed in this category overlap with 'left' and 'right' categories as defined by the MediaBiasFactCheck source, while domains labelled as left-centre and right-centre are considered Mainstream media. "Fake or hoax" contains entirely manipulated or fabricated inflammatory content that is intended to be perceived as realistic and reliable and whose goal may also be political but fails to meet the basic rules of plausibility and accountability, thus reaching an even higher level of HS. Finally, the highest level of HS is associated with "Conspiracy and junk science"—that is, to strongly ideological, inflammatory content that aims at building conceptual paradigms that are entirely alternative and oppositional to tested and accountable knowledge and information, with the intent of building self-referential bubbles where fidelized audiences are simply refusing a priori any kind of knowledge or information that is not legitimized by the alternative source itself or by recognized affiliates, as is typical in sects of a religious or other nature.

A third level of filtering concerned poorly defined domains—for example, the ones explicitly missing top-level domain names (such as ".com" or ".org")—as well as the domains not classifiable by means of our proposed scheme. This action reduced the database to the final number of 3,892 entries (Table 1 and Supplementary Fig. 1).

Finally, in Supplementary Note 6 we also provide quantitative results excluding effects due to the shift of misinformation towards untracked domains during the time frame of our analysis. In Supplementary Note 7, we further provide a comparison between MediaBiasFactCheck and other databases.

Data limitations and possible selection biases. The process of gathering and integrating vast sources of user-generated data provides us with the opportunity of analysing complex collective phenomena in almost real time. At the same time, it is subject to a number of limitations inherent in user-generated content data⁴⁵ selection biases that might influence the analysis at different levels. In this section, we discuss these limitations in detail, as well as how they affect our results.

Use of Twitter as a data source (population bias). All Twitter-based research has to cope with the intrinsic demographic limitations of Twitter's penetration: our results apply mostly to well-educated males (65% of Twitter users⁵⁷) between the ages of 18 and 34 (58% of Twitter users, according to Statista GmbH⁵⁸). Although our results must be interpreted in the light of these demographic limitations, we believe that our work represents a first step in establishing a robust research agenda for the study of infodemic risk. Future research should expand our knowledge by working on different demographics from different data sources.

Furthermore, as the COVID-19 public health emergency spread and raised international concern, Twitter (as well as Facebook and Google) took actions against the diffusion of unreliable/misleading news by attempting to prioritize reliable sources over unreliable ones. In Supplementary Note 8, we show how this action seems not to have influenced our measures.

Use of words written with Latin characters in the Twitter Filter API (data filtering bias). Latin characters, and particularly English, are widespread and often used for hashtags in messages in languages not using the same alphabet. However, the fact that we used a set of terms shared by Western languages (including English, Spanish, French, Portuguese, German, Italian and others) to select tweets in the Filter API may exacerbate the Twitter bias towards highly educated individuals in countries where local languages do not use Latin characters.

Use of a limited and static list of words in the Twitter Filter API (data filtering bias). As discussed above, our analyses do not focus on reconstructing the whole communication network related to the topic; instead, they focus on estimating the fraction and impact of unreliable news. Therefore, our rationale behind the word choice was to include the most commonly used keywords to ensure that, if the discourse abruptly changed its key terms, we were still tracking them. This might lower the recall rate, as new terms might be progressively emerging. In particular, our dataset only partially includes '#stayathome' or '#stayatwork' messages, but ultimately our focus is on understanding whether news related to key medical pandemic hashtags is reliable or not, and to what extent this news reliability correlates with the epidemic wave. For this reason, we chose a set of words commonly used in medical discourse, using query expansion when it was crucial for collecting medical-related data (for example, when the name of the virus and of the disease changed to SARS-CoV-2 and COVID-19, respectively, from the previous 2019-nCoV).

An alternative would have been to use automated query expansion techniques to enlarge the set of terms used for filtering. Unfortunately, there is not yet an agreement on a standard methodology, as each design leads to a different source of bias. For example, a possible method would have been to build a hashtag co-occurrence network periodically and to expand the list using more central nodes in such networks. However, query expansions might have increased the sample at the expense of introducing further bias in our analysis, as it would have been done, day by day, on a considerably different user base. While our choice does not provide a complete picture of the social dynamics during the pandemic, it was specifically designed for the task of gathering tweets containing links to medically related news sources, reliable or not, which is the focus of our paper.

Use of Western-centric fact-checking sites (data enrichment bias). To enhance the specificity and robustness of our multilingual Twitter dataset sample, we collected fact-checking information data from several different and independent sources. Since the World Wide Web is strongly English centric, this collection of sources provides an overabundance of information about content in English. The English-centric nature of the resources helping us identify unreliable news sources probably exacerbates the intrinsic Twitter demographic limitations towards well-educated English-speaking users, a bias that could not be amended by any more complete database.

To assess this limitation, we collected statistics from Amazon Alexa (www.alexa.com/topsites/countries) about web traffic (the top 50 most visited websites) for all countries across the globe, matching these lists with the list of domains we used to classify reliable and unreliable sources. Remarkably, for 127 countries we have at least one domain in the reliable top-50 news source, and for 21 (iso2 codes: AE, AR, BB, BE, CA, DK, FR, KE, MX, NG, PA, PE, PH, PR, PT, QA, SD, SE, TT, US and VE) we have at least one domain in the top-50 websites labelled as unreliable (split equally between politically biased and fake or hoax websites). In fact, this is a lower bound, because Alexa provided only major domains, disregarding subdomains that we instead classified as well. This large presence among the very top tier of websites suggests that our results are robust for multilingual/multicultural analysis.

In our opinion, however, it is not entirely correct to say that fact-checking sites suffer from a Western-centric bias. It is the very notion of institutional sources of fact-checking and certification of media bias that is today still largely Western centric. An eloquent picture is provided by Reporters Without Borders' Press Freedom Index (<https://rsf.org/en/ranking>), where it is clearly shown that today, apart from the Western world and a few isolated non-Western countries (South Korea, Costa Rica, Jamaica, Uruguay, South Africa, a few small Western African states and micro states), the media environment of all other countries cannot be considered free, and in such conditions, the possibility of thorough, transparent

fact-checking is basically impossible. So, whereas we acknowledge that our study suffers from other sources of bias, we are not sure that this particular source should be classified as such: we are simply considering the only functioning, relatively reliable sources of fact-checking available.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The datasets generated during the current study are available from the corresponding author on reasonable request. The aggregated information, compliant with all privacy regulations, is publicly available online at the Infodemics Observatory (<http://covid19obs.fbk.eu/>) and at OSF (<https://doi.org/10.17605/OSF.IO/N6UPX>).

Code availability

The custom code that supports the findings of this study is available from the corresponding author upon request and available alongside the data in the permanent repository indicated above.

Received: 11 April 2020; Accepted: 5 October 2020;

Published online: 29 October 2020

References

1. Benkler, Y. *The Wealth of Networks: How Social Production Transforms Markets and Freedom* (Yale Univ. Press, 2006).
2. Fuchs, C. *Social Media: A Critical Introduction* (SAGE, 2014).
3. Giglietto, F., Rossi, L. & Bennato, D. The open laboratory: limits and possibilities of using Facebook, Twitter, and YouTube as a research data source. *J. Technol. Hum. Serv.* **30**, 145–159 (2012).
4. Ojo, A. & Mellouli, S. Deploying governance networks for societal challenges. *Gov. Inf. Q.* <https://doi.org/10.1016/j.giq.2016.04.001> (2016).
5. De Domenico, M. & Altman, E. G. Unraveling the origin of social bursts in collective attention. *Sci. Rep.* **10**, 4629 (2020).
6. Vosoughi, S., Roy, D. & Aral, S. The spread of true and false news online. *Science* **359**, 1146–1151 (2018).
7. Shao, C. et al. The spread of low-credibility content by social bots. *Nat. Commun.* **9**, 4787 (2018).
8. Stella, M., Ferrara, E. & De Domenico, M. Bots increase exposure to negative and inflammatory content in online social systems. *Proc. Natl Acad. Sci. USA* **115**, 12435–12440 (2018).
9. Eysenbach, G. Infodemiology: the epidemiology of (mis)information. *Am. J. Med.* **113**, 763–765 (2002).
10. Eysenbach, G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. *J. Med. Internet Res.* **11**, e11 (2009).
11. Eysenbach, G. Infodemiology and infoveillance tracking online health information and cyberbehavior for public health. *Am. J. Prev. Med.* **40**, S154–S158 (2011).
12. Zarocostas, J. How to fight an infodemic. *Lancet* **395**, 676 (2020).
13. Pastor-Satorras, R., Castellano, C., Van Mieghem, P. & Vespignani, A. Epidemic processes in complex networks. *Rev. Mod. Phys.* **87**, 925–979 (2015).
14. De Domenico, M., Granell, C., Porter, M. A. & Arenas, A. The physics of spreading processes in multilayer networks. *Nat. Phys.* **12**, 901–906 (2016).
15. Brockmann, D. & Helbing, D. The hidden geometry of complex, network-driven contagion phenomena. *Science* **342**, 1337–1342 (2013).
16. Huang, C. et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395**, 497–506 (2020).
17. Zhu, N. et al. A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med.* **382**, 727–733 (2020).
18. Chinazzi, M. et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* **368**, 395–400 (2020).
19. Lazer, D. M. J. et al. The science of fake news. *Science* **359**, 1094–1096 (2018).
20. Rapp, D. N. & Salovich, N. A. Can't we just disregard fake news? The consequences of exposure to inaccurate information. *Policy Insights Behav. Brain Sci.* **5**, 232–239 (2018).
21. Waszak, P. M., Kasprzycka-Waszak, W. & Kubanek, A. The spread of medical fake news in social media—the pilot quantitative study. *Health Policy Technol.* **7**, 115–118 (2018).
22. Leung, G. M. & Leung, K. Crowdsourcing data to mitigate epidemics. *Lancet Digit. Health* [https://doi.org/10.1016/S2589-7500\(20\)30055-8](https://doi.org/10.1016/S2589-7500(20)30055-8) (2020).
23. Altay, S., de Araujo, E. & Mercier, H. 'If this account is true, it is most enormously wonderful': interestingness-if-true and the sharing of true and false news. Preprint at *PsyArXiv* <https://doi.org/10.31234/osf.io/tdfh5> (2020).
24. Vicario, M. D., Quattrociocchi, W., Scala, A. & Zollo, F. Polarization and fake news. *ACM Trans. Web* **13**, 10 (2019).

25. Britt, M. A., Rouet, J.-F., Blaum, D. & Millis, K. A reasoned approach to dealing with fake news. *Policy Insights Behav. Brain Sci.* **6**, 94–101 (2019).
26. *Weekly Epidemiological Record* Vol. 95, 16 (WHO, 2020); <https://www.who.int/wer/2020/wer9516/en/>
27. Tangcharoensathien, V. et al. A framework for managing the COVID-19 infodemic: methods and results of an online, crowdsourced WHO technical consultation. *J. Med. Internet Res.* <https://doi.org/10.2196/19659> (2020).
28. Lunn, P. D. et al. Using behavioral science to help fight the Coronavirus. *J. Behav. Public Adm.* <https://doi.org/10.30636/jbpa.31.147> (2020).
29. Ferrara, E., Varol, O., Davis, C., Menczer, F. & Flammini, A. The rise of social bots. *Commun. ACM* **59**, 96–104 (2016).
30. Bessi, A. & Ferrara, E. Social bots distort the 2016 U.S. Presidential election online discussion. *First Monday* <https://doi.org/10.5210/fm.v21i11.7090>(2016).
31. Ferrara, E. Disinformation and social bot operations in the run up to the 2017 French presidential election. *First Monday* <https://doi.org/10.5210/fm.v22i8.8005> (2017).
32. Kwak, H., Lee, C., Park, H. & Moon, S. What is Twitter, a social network or a news media? In *Proc. 19th International Conference on World Wide Web* 591 (ACM, 2010).
33. Barabasi, A. L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
34. Watts, D. J. & Strogatz, S. H. Collective dynamics of ‘small-world’ networks. *Nature* **393**, 440–442 (1998).
35. Watts, D. J. A simple model of global cascades on random networks. *Proc. Natl Acad. Sci. USA* **99**, 5766–5771 (2002).
36. Gleeson, J. P., O’Sullivan, K. P., Baños, R. A. & Moreno, Y. Effects of network structure, competition and memory time on social spreading phenomena. *Phys. Rev. X* **6**, 021019 (2016).
37. Aral, S. & Eckles, D. Protecting elections from social media manipulation. *Science* **365**, 858–861 (2019).
38. Stella, M., Cristoforetti, M. & De Domenico, M. Influence of augmented humans in online interactions during voting events. *PLoS ONE* **14**, e0214210 (2019).
39. Hébert-Dufresne, L., Scarpino, S. V. & Young, J.-G. Macroscopic patterns of interacting contagions are indistinguishable from social reinforcement. *Nat. Phys.* <https://doi.org/10.1038/s41567-020-0791-2> (2020).
40. Eysenbach, G. How to fight an infodemic: the four pillars of infodemic management. *J. Med. Internet Res.* **22**, e21820 (2020).
41. Marchetti, R. & Ceccobelli, D. Twitter and television in a hybrid media system. *Journalism Pract.* **10**, 626–644 (2016).
42. Yen, H., Braun, S. & Woodward, C. AP fact check: Trump’s alternate reality on COVID-19 threat. *Associated Press* <https://apnews.com/0aa783aa734b2ac3d984c5116b3e8039> (20 July 2020).
43. Broad, W. J. Putin’s long war against American science. *The New York Times* <https://www.nytimes.com/2020/04/13/science/putin-russia-disinformation-health-coronavirus.html> (13 April 2020).
44. Iran’s reaction to coronavirus has become a danger for the world. *The Washington Post* https://www.washingtonpost.com/opinions/global-opinions/irans-moment-of-truth-on-coronavirus/2020/03/03/f82548fe-5cca-11ea-b29b-9db42f7803a7_story.html (3 March 2020).
45. Coronavirus: world leaders’ posts deleted over fake news. *BBC News* <https://www.bbc.com/news/technology-52106321> (31 March 2020).
46. Olteanu, A. et al. Social data: biases, methodological pitfalls, and ethical boundaries. *Front. Big Data* **2**, 13 (2019).
47. Pfeffer, J., Mayer, K. & Morstatter, F. Tampering with Twitter’s Sample API. *EPJ Data Sci.* **7**, 50 (2018).
48. Zimdar, M. My fake news list went viral but made up stories are only part of the problem. *The Washington Post* <https://www.washingtonpost.com/postever-ything/wp/2016/11/18/my-fake-news-list-went-viral-but-made-up-stories-are-only-part-of-the-problem/> (18 November 2016).
49. Silverman, C. Inside the partisan fight for your news feed. *BuzzFeed News* <https://www.buzzfeednews.com/article/craigsilverman/inside-the-partisan-fight-for-your-news-feed> (8 August 2017).
50. *Fake News Watch* (2015); <https://web.archive.org/web/20180213181029/http://www.fakenewswatch.com/>
51. Politifacts guide to fake news and what they peddle. *Politifacts.com* <https://www.politifact.com/article/2017/apr/20/politifacts-guide-fake-news-websites-and-what-they/> (20 April 2017).
52. The black list. La lista nera del web. *Bufole.net* <https://www.bufole.net/the-black-list-la-lista-nera-del-web/> (2018).
53. Starbird, K. et al. Ecosystem or echo-system? Exploring content sharing across alternative media domains. In *12th International AAAI Conference on Web and Social Media* 365–374 (AAAI, 2018).
54. Fletcher, R. et al. *Measuring the Reach of ‘Fake News’ and Online Disinformation in Europe* (Reuters Institute, 2018); <https://reutersinstitute.politics.ox.ac.uk/our-research/measuring-reach-fake-news-and-online-disinformation-europe>
55. Grinberg, N. et al. Fake news on Twitter during the 2016 US presidential election. *Science* **363**, 374–378 (2019).
56. *MediaBiasFactCheck* (2020); <https://mediabiasfactcheck.com/>
57. *Distribution of Twitter Users Worldwide as of July 2020, by Gender* (Statista, 2020); <https://www.statista.com/statistics/828092/distribution-of-users-on-twitter-worldwide-gender/>
58. *Distribution of Twitter Users Worldwide as of July 2020, by Age Group* (Statista, 2020); <https://www.statista.com/statistics/283119/age-distribution-of-global-twitter-users/>

Acknowledgements

We received no specific funding for this work. We acknowledge the support of the FBK’s Digital Society Department and the FBK’s Flagship Project CHUB (Computational Human Behavior). We thank all FBK’s Research Units for granting us privileged access to extraordinarily high-performance computing for the analysis of massive infodemic data. We thank J. Baumgartner for sharing data between 21 January and 24 January 2020. We acknowledge the WHO Information Network for Epidemics (WHO EPI-WIN) for useful discussions and the scientific members of the WHO ad hoc online consultation on managing the COVID-19 infodemic for very inspiring insights and conversations.

Author contributions

M.D.D. conceived the study. M.D.D. and F.V. collected the data. R.G., N.C. and F.V. analysed the data. M.D.D., P.S. and R.G. interpreted the data and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41562-020-00994-6>.

Correspondence and requests for materials should be addressed to P.S. or M.D.

Peer review information Primary handling editor: Stavroula Kousta.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

Data analysis

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The datasets generated during the current study are available from the corresponding author on reasonable request. Aggregated information, compliant with all privacy regulations, are publicly available online at the Infodemics Observatory (<http://covid19obs.fbk.eu/>) and on a permanent repository (DOI 10.17605/OSF.IO/N6UPX).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Quantitative analysis of automatically gathering and spatially and temporally aggregated user-generated data from Twitter.
Research sample	Over 100 Million Twitter messages
Sampling strategy	Twitter Search API with keywords : coronavirus, ncov, #Wuhan, covid19, covid-19, sarscov2, covid
Data collection	python scripts
Timing	22 jan - 10 mar 2020
Data exclusions	Messages whose user position could not be geocoded. Messages with no URLs shared
Non-participation	NA
Randomization	NA

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging