nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-022-01353-3

Check for updates

The supply and demand of news during COVID-19 and assessment of questionable sources production

Pietro Gravino[1]✉, Giulio Prevedello[1], Martina Galletti[1] and Vittorio Loreto[1,2,3]

Misinformation threatens our societies, but little is known about how the production of news by unreliable sources relates to supply and demand dynamics. We exploit the burst of news production triggered by the COVID-19 outbreak through an Italian database partially annotated for questionable sources. We compare news supply with news demand, as captured by Google Trends data. We identify the Granger causal relationships between supply and demand for the most searched keywords, quantifying the inertial behaviour of the news supply. Focusing on COVID-19 news, we find that questionable sources are more sensitive than general news production to people's interests, especially when news supply and demand mismatched. We introduce an index assessing the level of questionable news production solely based on the available volumes of news and searches. We contend that these results can be a powerful asset in informing campaigns against disinformation and providing news outlets and institutions with potentially relevant strategies.

The Covid-19 crisis evidenced once more that disinformation stands as one of the major plagues of the Information Age. In the past decades, many national and international institutions started to implement a plethora of strategies to tackle this issue[1] and mitigate its effects. Still, the mechanisms underlying the role and phenomenology of disinformation are largely unclear.

It is only in recent times that the complex ecosystem of information has massively attracted the interest of the scientific community[2]. Disinformation went under investigation[3], from its very definition[4] to its psychological mechanisms[5,6], and its spreading and exposure dynamics[7,8]. Detection and forecast of disinformation were also among the relevant topics explored by the scientific community[9]. These studies raised questions about how to identify statistical markers in the news content[10] or about diffusion mechanisms[11]. Another important debate in the literature revolves around the contrast to common narratives about disinformation, such as those concerning its fast-spreading pattern[12], its link with partisanship[6] or the most suitable sharing-prevention strategies[13].

A meaningful part of the research effort focused on the impact of disinformation on diverse fields of human activities, such as consumers' behaviour[14], political elections[15], sustainability[16] or health[17]. Particularly during the Covid-19 pandemic, the effect of disinformation on social behaviours became so compelling that the term 'Infodemic' made a comeback from the SARS epidemic of 2003[18]. Infodemic refers to the spreading of many pieces of information about the SARS-CoV-2 virus, some correct others false, to the extent that it overwhelms people and hinders their understanding of the phenomenon. The consequences were disastrous[19] and led to dangerous behaviours that further aggravated the pandemic crisis.

While disinformation is always under the spotlight, the complex ecosystem of information, which is the substrate for disinformation, attracted much less interest. It is important to stress that the infosphere relies on the subtle interplay of two types of actors: news producers on the one hand and news consumers on the other. In this structure, information supply and demand stand in a market-like relationship. The study of their interplay is essential to unveil the mechanisms of information dynamics. It also provides a broader view in which disinformation can be contextualized and analysed.

The news supply can be identified with the overall news production, mainly consisting of officially recognized newspapers. As such, the extensive literature concerning journalism and media information is investigated: news linguistics[20,21]; journalism economics[22]; information coverage[23,24], often focusing on particular countries or topics[25]; content of news[26], its quality assessment[27] and delivery[28], considering different media sources[29]. Other works investigated the impacts of news and its consumption on, for example, reading behaviour[30], finance[31] and political opinions[32].

Similarly, news demand has also been studied deeply and from several angles, covering, for example: demographic groups of audience[33] and their behaviour[34]; consumer needs[35] and assessment of news[36]; the interplay between news consumption and production[37], and the adaptation of media to news consumer behaviours enabled by technology[38–40]; the drivers for consumption of news[41] and misinformation[42].

However, news demand is more difficult to pinpoint methodologically than news supply. In the literature, surveys and lab studies are usual procedures of investigation[29,30,43,44], but unlike general news production, they cannot scale up to the population level. Thus, different solutions must be adopted. The tracking of reading behaviours, for example, had been used to study the demands and interests of readers[45]. However, such a methodology is biased by the very existence of news since the interest for topics not covered by news cannot be recorded.

An independent way to track people's interests that gained popularity in the scientific community is the Google Trends service (https://trends.google.com/)[46]. It provides an index proportional to the number of searches made with the Google Search engine, enabling the quantitative comparison of searched queries. As Google's algorithm aims at delivering the information that best relates with the input query[47], Google Trends has been mainly used in the past decade as a marker of people's behaviours in different

[1]Sony Computer Science Laboratories, Paris, France. [2]Physics Department, Sapienza University of Rome, Rome, Italy. [3]Complexity Science Hub Vienna, Vienna, Austria. ✉e-mail: pietro.gravino@sony.com

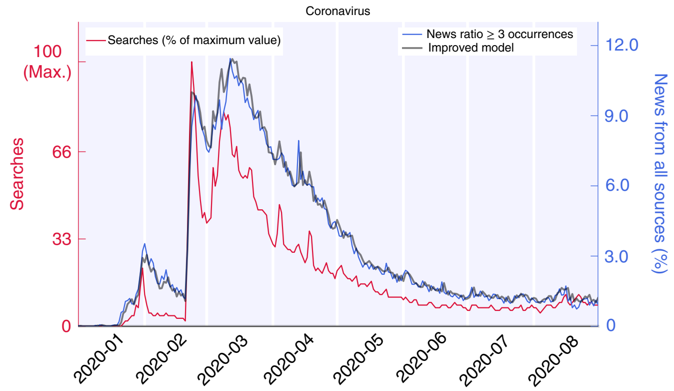

Fig. 1 | Temporal behaviour of the fractions of Searches and News from All Sources in Italy. The Searches (red, left *y* axis) and News from All Sources (blue, right *y* axis) for the keyword 'coronavirus' were recorded from 6 December 2019 to 31 August 2020. Searches are reported as a percentage of the maximum observed in the monitored period. News from All Sources is represented by the daily fraction of articles containing at least three keyword occurrences (see Methods). The improved model (black line) leverages the past News from All Sources and Searches, together with present Searches, to infer the dynamics of News from All Sources.

contexts, such as finance[48,49], epidemiology[50–52] or socio-economic indicators[53,54]. Interestingly, its intrinsic value as a proxy for people's demand for information was considered[55], but not extensively.

Intents underlying the queries could be categorized into: informational, to seek for information; navigational, to reach a certain website; and transactional, to perform a web-mediated activity such as shopping[56]. In particular, these motivations have been differentiated using the query itself by assuming that searches of keywords that are newsworthy are made to demand information[57]. While the limitations of Google's searches were considered, evidencing diversity of interaction among users and even misuse of the platform[58,59], we also found that in the context of news, the Google Trends index has been adopted for forecasting[60] and acknowledged by information media as a source of attention for news[61].

Given the plethora of research on the supply of news on one side and on the demand for information on the other, we focused on the relations between the two. In this sense, the news ecosystem dynamics can be studied as a single complex system, and we could leverage this comprehension to deepen our understanding of related phenomena, such as the production of news from questionable sources. Through its results and methodology, this paper aims to pave the way for such a discussion.

Here we investigate, using a unified framework, the supply and demand for information and analyse their dynamical interplay, with the final goal of understanding the main mechanisms of the information ecosystem dynamics and extracting hints about the determinants of questionable sources production during the COVID-19 outbreak. To this end, we focused on the general production of news in Italy, from early December 2019 to the end of August 2020, as the reference for the news supply. For the same period and country, the Google Trend index served as a proxy for the general public's information demand.

We adopted Vector Auto-Regression (VAR) models to study the interplay between news demand and supply, evidencing different causal relationships for distinct subjects. We presented an improved modelling scheme that allows for a quantitative description of the dependencies in the time-series evolution for information demand and supply.

The framework also permitted the study and comparison of the supply of sources identified as questionable by professional fact-checkers within the general information system during the COVID-19 upsurge, highlighting behavioural differences in reactivity and modelling efficacy for COVID-19-related news from general and unreliable media. Furthermore, we observed that for the same period, the semantic misalignment between COVID-19-related information demand and supply from all sources is higher than the misalignment between COVID-19-related information demand and supply from questionable sources.

These discrepancies could be exploited to aggregate a questionable sources activity indicator independent from annotations. We contend that this index could provide a reliable and independent assessment tool for the news supply's health status.

Results

Dynamics of news supply and demand. Information systems feature two main drivers: news supply and news demand. As a reference for the news supply, we looked at the whole Italian production of information from every single news outlet, termed News from All Sources or N_{AS}, from early December 2019 to the end of August 2020. For the same period, the Google Trend index served as a proxy for the news demand from the Italian general public, here termed Searches or S (see Methods for details).

To investigate the nature of the relation between supply and demand of news about a certain subject, six keywords referring to the most searched subjects in Italy over the entire observation period were selected: 'coronavirus', 'regionali', 'playstation', 'papa francesco', 'eurovision' and 'sondaggi' (Supplementary Fig. 1). The time series for N_{AS} and S for 'coronavirus' are reported in Fig. 1. For each keyword, the time series of the daily appearances in News from All Sources and the daily volume of queries in the Searches were simultaneously fitted by VAR linear modelling[62]. VAR models with different lag parameters that encapsulate the system's memory were considered, and the best parameters were identified via the Akaike criterion[63] (see Methods). For all keywords, best-fitting lags ranged between 2 and 4, suggesting a typical, short-memory timescale in the system (Supplementary Fig. 2).

Within the VAR framework, we performed the test for Granger-causality[62] to illustrate which time series, between N_{AS} and S, contributed more to the prediction of the other, and whether any contribution was significant. For the majority of keywords (that is, 'coronavirus', 'regionali', 'playstation', 'papa francesco'), the contribution of past Searches to present News from All Sources was most significant (see Supplementary Fig. 2). We could safely assume that S anticipates N_{AS} and use this assumption to improve the model of the temporal behaviour of the latter. We modified the VAR equation for the evolution of News from All Sources by inserting Searches' role. More precisely, we let $S(t)$ and $N_{AS}(t)$ be the respective values of Searches and News from All Sources at day t, and the new equation for the evolution of $N_{AS}(t)$ reads:

$$N_{\mathrm{AS}}(t)=\sum_{i=1}^{d}(\alpha_i N_{\mathrm{AS}}(t-i)+\beta_i S(t-i))+\beta_0 S(t), \qquad (1)$$

where the coefficients α_i, β_0 and β_i were fitted, while the Akaike criterion provides the optimal lag d. This 'improved model' closely reproduced the data, particularly in correspondence with the peaks (Fig. 1 for 'coronavirus' and Supplementary Fig. 3).

The model's parameters also provided a quantitative insight on the interplay between N_{AS} and S (Table 1):

- α_1 was always significantly positive, indicating a strong dependence of News from All Sources on the previous day's activity. This evidence is a sign of an inertial behaviour of the news supply.
- β_0, the weight of present Searches, was always significantly non-zero, supporting the assumption of the role of present Searches for the improved model.

Table 1 | The parameters and statistics from the improved model of equation (1) for the 4 selected keywords

Keyword	**'coronavirus'**	**Adjusted $R^2 = 0.995$ (0.991)**		**DoF = 266**
Coeff.	**Value**	**95% C.I.**	***T*-stat.**	***P* value**
α_1	0.82	(0.78, 0.86)	43	<0.001
β_0	0.071	(0.061, 0.081)	14	<0.001
β_1	−0.035	(−0.047, −0.022)	−5.4	<0.001
Keyword	**'regionali'**	**Adjusted R^2** = 0.89 (0.86)		**DoF = 263**
Coeff.	**Value**	**95% C.I.**	***T*-stat.**	***P* value**
α_1	0.65	(0.54, 0.77)	11	<0.001
α_2	0.26	(0.15, 0.37)	4.7	<0.001
β_0	0.0085	(0.0053, 0.012)	5.2	<0.001
β_1	0.0032	(−0.00076, 0.0072)	1.6	0.113
β_2	−0.0066	(−0.01, −0.0032)	−3.8	<0.001
Keyword	**'playstation'**	**Adjusted R^2** = 0.54 (0.29)		**DoF = 263**
Coeff.	**Value**	**95% C.I.**	***T*-stat.**	***P* value**
α_1	0.18	(0.066, 0.3)	3.1	0.002
α_2	0.19	(0.069, 0.3)	3.1	0.002
β_0	0.00056	(0.00034, 0.00078)	5	<0.001
β_1	0.00036	(3×10^{-5}, 0.00069)	2.1	0.033
β_2	−0.00069	(−0.00093, −0.00046)	−5.8	<0.001
Keyword	**'papa francesco'**	**Adjusted R^2** = 0.73 (0.63)		**DoF = 267**
Coeff.	**Value**	**95% C.I.**	***T*-stat.**	***P* value**
α_1	0.54	(0.46, 0.62)	14	<0.001
β_0	0.0039	(0.0031, 0.0046)	10	<0.001

For each keyword, the number of degrees of freedom (DoF) and the adjusted R^2 are reported, together with the adjusted R^2 (in brackets) from a trivial model with equation $N_{AS}(t) = \alpha N_{AS}(t-1)$, that is, a model where every day depends only on the day before. The fitness of the improved model is always larger than that of the trivial model. All tests are two-tailed and not adjusted for multiple comparisons.

- the last β parameters for two keywords (β_1 for 'coronavirus' and β_2 for 'regionali') were significantly negative. This result suggests that N_{AS} might depend on the different quotient of Searches, together with the volume of Searches itself.

- Of note, a direct comparison between α and β parameters was not possible, as S and N_{AS} were scaled differently (Google Trends does not disclose the absolute scale of the queries volume).

The different behaviours of questionable sources. The improved model (Table 2) quantifies the information supply dynamics and enables the comparison between News from All Sources and questionable news suppliers' production. We applied this methodology to the topic 'coronavirus' since it dominated the landscape of information (Supplementary Fig. 1) and due to the direct impact of disinformation on the response to the 2020 pandemic. To this end, we extended our analysis to news items that were produced by sources annotated as questionable by professional fact-checkers (see Methods). We named the production of this subset of outlets as News from Questionable Sources or N_{QS}. Such a supply is very scarce for keywords other than 'coronavirus', as reported in Table 3, these scarce keywords not being analysed with this methodology. In the following, we refer specifically to 'coronavirus'-related S, N_{AS} and N_{QS}.

We exploited the improved model in equation (1) to compare N_{AS} and N_{QS} through their best-fitting coefficients α and β. To this end, we paralleled the variable $N_{AS}(t)$, the daily proportion of 'coronavirus'-related News from All Sources at day t, and $N_{QS}(t)$, the daily proportion of 'coronavirus'-related News from Questionable Sources at day t (Table 4).

Table 2 | The parameters and statistics from the improved model of equation (1) fitting N_{AS}, N_{QS} and News from Online Sources N_{OS} having at least one occurrence of the keyword 'coronavirus' (see Methods)

All Sources		**Adjusted $R^2 = 0.995$**		**DoF = 241**
Coeff.	**Value**	**95% C.I.**	***T*-stat.**	***P* value**
α_1	0.860	(0.824, 0.890)	51	<0.001
β_0	0.460	(0.387, 0.526)	12	<0.001
β_1	−0.248	(−0.325, −0.159)	−5.7	<0.001
Questionable Sources		**Adjusted R^2** = 0.931		**DoF = 241**
Coeff.	**Value**	**95% C.I.**	***T*-stat.**	***P* value**
α_1	0.758	(0.681, 0.836)	51	<0.001
β_0	0.294	(0.125, 0.463)	12	<0.001
β_1	−0.081	(−0.261, 0.099)	−5.7	0.38
Online Sources		**Adjusted R^2** = 0.964		**DoF = 236**
Coeff.	**Value**	**95% C.I.**	***T*-stat.**	***P* value**
α_1	0.832	(0.775, 0.889)	51	<0.001
β_0	0.125	(0.125, 0.463)	1.39	0.17
β_1	0.096	(−0.092, 0.284)	1.01	0.31

All tests are two-tailed and not adjusted for multiple comparisons.

Compared with News from All Sources, 'coronavirus'-related News from Questionable Sources shows a lower inertia term, α_1. This difference is meaningful when compared with the confidence

Table 3 | The exact sample size (*n*) for each keyword, given as a discrete number for both N_{AS} data and N_{QS} data

Keyword	(*n*) for N_{AS}	(*n*) for N_{AS}	(*n*) for N_{QS}
	#occ. ≥1	#occ. ≥3	#occ. ≥1
‘coronavirus’	1,368,246	216,993	34,020
‘regionali’	260,263	19,859	2,744
‘playstation’	5,740	435	133
‘papa francesco’	48,526	4,207	581
‘eurovision’	1,926	371	26
‘sondaggi’	68,656	6,900	1,353

We used the metric with at least three occurrences (in the table indicated as ‘#occ. ≥3’) for the improved model described in equation (1). The most inclusive metric (in the table indicated as ‘#occ. ≥1’) was used when comparing N_{AS} data and N_{QS} data.

Table 4 | Cross-model coefficients comparison from improved model fitting of N_{AS}, N_{QS} and N_{OS} having at least one occurrence of the keyword ‘coronavirus’ (see Methods and, for the number of observations, Table 3)

Sources comparison	Coeff.	Δ	*P* value	Tail side
All vs Questionable	α_1	0.102	<0.001	left
	β_0	0.164	<0.001	left
	β_1	−0.290	<0.001	right
Online vs Questionable	α_1	0.07	<0.001	left
	β_0	−0.17	<0.001	right
	β_1	0.18	<0.001	left

Statistics Δ and *P* values were calculated using a bootstrap procedure to test the null hypothesis that one coefficient from the model of one source is smaller (left-tail test) or larger (right-tail test) than the same coefficient from the model of the other source (see Methods). *P* values are not adjusted for multiple comparisons.

intervals and has been further validated through a bootstrap analysis with replacement (see Methods). It is worth noting that this difference is not due to the online nature of the unreliable suppliers. Our analysis shows that online suppliers, named News from Online Sources or N_{OS}, present a significantly different inertial behaviour than News from Questionable Sources. The results from both tests are reported in Table 4. Unreliable supply also shows a non-significant β_1, indicating a greater reactivity to $S(t)$. These pieces of evidence and the lower prediction score (adjusted R^2) suggest that unreliable supply presents a different behaviour than N_{AS} to the point that it distorts the dynamics of the news ecosystem and leads to impaired modelling performance.

Another difference in the behaviour of N_{AS} and N_{QS} emerged at a semantic level. We focused on the most queried keywords searched together with ‘coronavirus’ in Google Search (see Methods). Each of these related queries provided a time series of news demand about a subdomain that co-occurs with, and therefore is semantically linked to, ‘coronavirus’. We quantified the co-occurrence of these terms with the ‘coronavirus’ keyword also in the news items for both N_{AS} and N_{QS}. In this way, we defined $S(t)$, $N_{\mathrm{AS}}(t)$, $N_{\mathrm{QS}}(t)$ as the daily semantic vectors for ‘coronavirus’-related Searches, News from All Sources and News from Questionable Sources, respectively. Each vector has 17 entries, one per subdomain (see Methods for details).

We calculated $S_{\mathrm{tot}}=\sum_t S(t)$ and sorted its components to rank the different subdomains by the total news demand over the period considered (Fig. 2). To assess the difference between general and questionable suppliers with respect to the matching of news demand for different subdomains, we challenged the components’ rankings of $N_{\mathrm{AS}_{\mathrm{tot}}} = \sum_t N_{\mathrm{AS}}(t)$ and $N_{\mathrm{QS}_{\mathrm{tot}}} = \sum_t N_{\mathrm{QS}}(t)$ against the corresponding ones of S_{tot} (Fig. 2).

Given the ‘coronavirus’-related keywords ranked from the Searches as a reference, News from Questionable Sources ranking shows fewer and minor mismatches compared with News from All Sources. We quantified this difference in behaviour through Spearman’s correlation (Supplementary Fig. 4). S_{tot} and $N_{\mathrm{AS}_{\mathrm{tot}}}$ components were positively correlated ($r=0.52$, two-tailed *P* value <0.031 with 17 observations), but S_{tot} and $N_{\mathrm{QS}_{\mathrm{tot}}}$ were more correlated ($r=0.67$, two-tailed *P* value <0.003 with 17 observations).

The semantic difference in the behaviour of News from Questionable Sources and News from All Sources holds not only at the aggregated level but also at a daily level. This was measured through the cosine distance $\mathrm{d}(\cdot,\cdot)$ on their daily vectors $S(t)$, $N_{\mathrm{AS}}(t)$ and $N_{\mathrm{QS}}(t)$ (see Methods). Again, Searches were taken as reference and we calculated its cosine distance from News from All Sources, $\mathrm{d}(S(t),N_{\mathrm{AS}}(t))$, and from News from Questionable Sources, $\mathrm{d}(S(t),N_{\mathrm{QS}}(t))$. The daily relative difference between these distances

$$\frac{\mathrm{d}(S(t),N_{\mathrm{QS}}(t))-\mathrm{d}(S(t),N_{\mathrm{AS}}(t))}{\mathrm{d}(S(t),N_{\mathrm{AS}}(t))} \tag{2}$$

resulted in negative values in most days *t* (Supplementary Fig. 5). In fact, both the mean (−0.13) and median (−0.15) were negative, indicating that the cosine distance of ‘Searches-News from Questionable Sources’ is generally smaller than that of ‘Searches-News from All Sources’. This result shows how News from Questionable Sources meets news demand better than News from All Sources.

Independent detection of questionable sources concentration. The observed differences between N_{AS} and N_{QS} dynamics can be exploited to assess the production of ‘coronavirus’-related news items from unreliable suppliers.

The difference in modelling N_{QS} and N_{AS} suggests that when the concentration of News from Questionable Sources on a topic increases, the N_{AS} dynamics, which includes N_{QS}, becomes perturbed. We hypothesize that this perturbation impairs the modelling performance of News from All Sources. To test this hypothesis, the improved model was fitted to N_{AS} locally on a time window of 14 d, sliding over the entire data time range (see Methods). For each window, centred in *t*, we computed the local modelling error defined as:

$$E(t)=(1-R^2(t))\langle N_{\mathrm{AS}}\rangle(t), \tag{3}$$

where $R^2(t)$, the R^2 score for the model fitted to the window, is weighted by $\langle N_{\mathrm{AS}}\rangle(t)$, the average share of News from All Sources produced in that time window.

Although formulated without exploiting fact-checkers’ annotations, $E(t)$ significantly correlates with the concentration of News from Questionable Sources on the ‘coronavirus’ subject, $N_{\mathrm{QS}}(t)/N_{\mathrm{AS}}(t)$ (Spearman’s $r=0.47$, two-tailed *P* value <0.001 with 217 observations, see Methods). This result supports the hypothesis that loss of predictability from the N_{AS} dynamics co-occurs with N_{QS} spikes. As a consequence, *E* could be a very promising proxy for the concentration of News from Questionable Sources production about the topic ‘coronavirus’.

The semantic difference between News from All Sources and News from Questionable Sources suggests that unreliable suppliers might react not only to the news demand but, in particular, to the ‘semantically unsatisfied’ news demand. We hypothesized that as N_{AS} becomes more semantically distant from *S*, N_{QS} would fill that gap. This hypothesis was tested over the entire time range by measuring the daily cosine distance between the semantic vectors

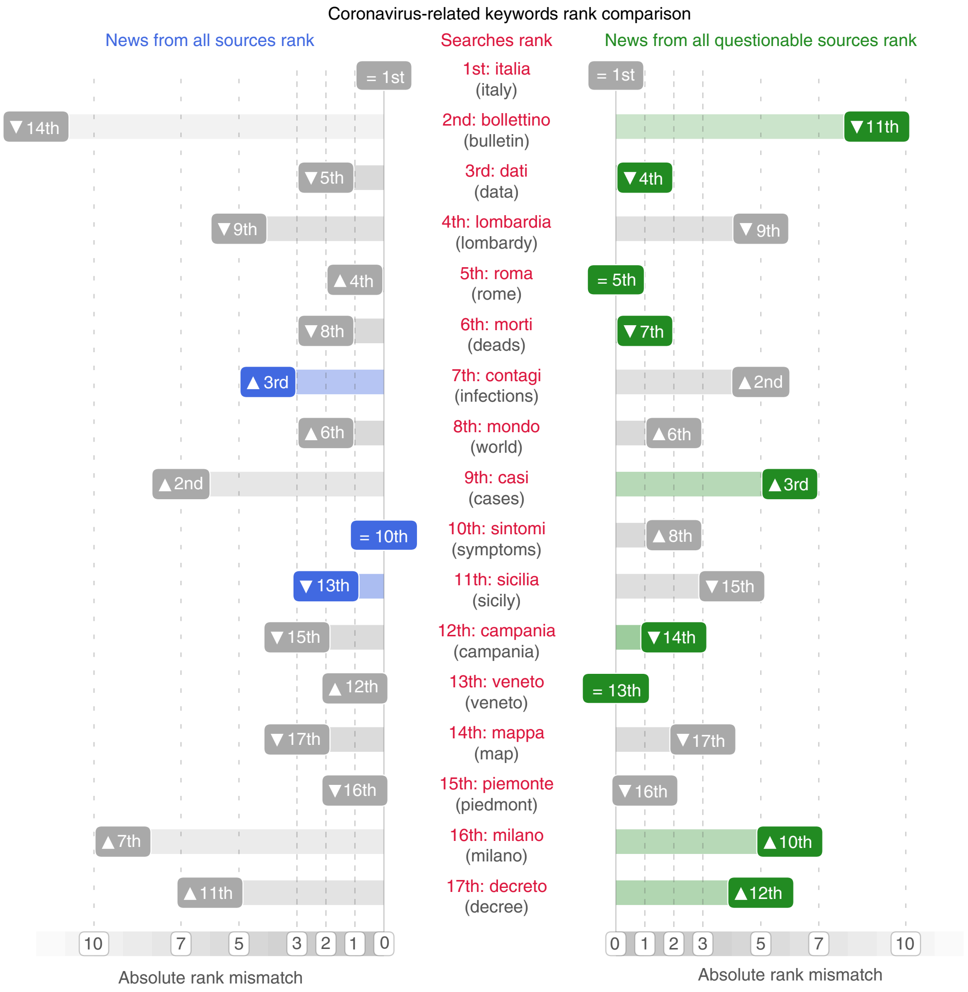

Fig. 2 | The ranked components of S_{tot}, representing 'coronavirus' subdomains sorted by total news demand over the observed time. Middle: S_{tot}, red text. On the sides of each keyword, a tag indicates the rank in $N_{AS_{tot}}$ for News from All Sources (left), and in $N_{QS_{tot}}$ for News from Questionable Sources (right). Tags are distanced from the centre by the amount of rank mismatch to Searches ranks. Tags are coloured to highlight the rank closest to the Searches rank: blue for News from All Sources and green for News from Questionable Sources.

of Searches and News from All Sources $K(t)=\mathrm{d}(S(t),N_{AS}(t))$. We then checked the correlation of $K(t)$ with the daily concentration of News from Questionable Sources, $N_{QS}(t)/N_{AS}(t)$, about 'coronavirus'. The correlation turns out to be positive and significant (Spearman's $r=0.58$, two-tailed P value <0.001 with 223 observations), supporting the hypothesis. This result allows us to adopt K as a second independent indicator for the concentration of News from Questionable Sources.

To test the effectiveness of our indicators E and K in assessing the concentration of news from unreliable sources on COVID-19, we merged them in a 'combined index' (see Methods) that could be considered as a questionable sources activity indicator. We fitted them linearly on a training set composed of approximately the first 25% of data from the time series providing the best linear combination of the two (Fig. 3).

The combined index was then tested against the validation set, achieving substantial accuracy (reduced chi-squared statistic of 0.945). All these findings suggest that the combined index provides a valuable measure for assessing the concentration of news from unreliable sources on COVID-19.

In principle, the methodology could also be applied to different topics to assess the health status of the news ecosystem at a more general level. Unfortunately, mainly due to the scarcity of fact-checkers' annotation for the other keywords (reported in Table 3), we could not perform the same analysis. However, to test whether the methodology can be, to some extent, generalized, we considered the set of the 4 keywords modelled. We aggregated them to create a synthetic macro-topic, for which they individually represented analogues of the related queries we have seen before. We judged the adoption of the first indicator, that is, the weighted modelling error for the local fitting, to be pointless since the macro-subject dynamics is largely dominated by the topic 'coronavirus'. This would have resulted in an indicator similar to the modelling of the 'coronavirus' component alone. We thus focused only on the second indicator, that is, the cosine distance between the semantic vectors of Searches and News, where the components of the vectors are now the values of News from All Sources, News from Questionable Sources, and Searches for the 4 keywords. The daily value of cosine distance between News from All Sources and Searches of the synthetic subject correlates positively and meaningfully with the concentration of News from Questionable Sources on the synthetic subject (Spearman's $r=0.44$, two-tailed P value <0.001 with 215 observations). Although the generalizability of the previous measurements must be tested for different topics as well as for different languages and different timeframes, this last result supports the plausibility of the application of our methodology in wider contexts.

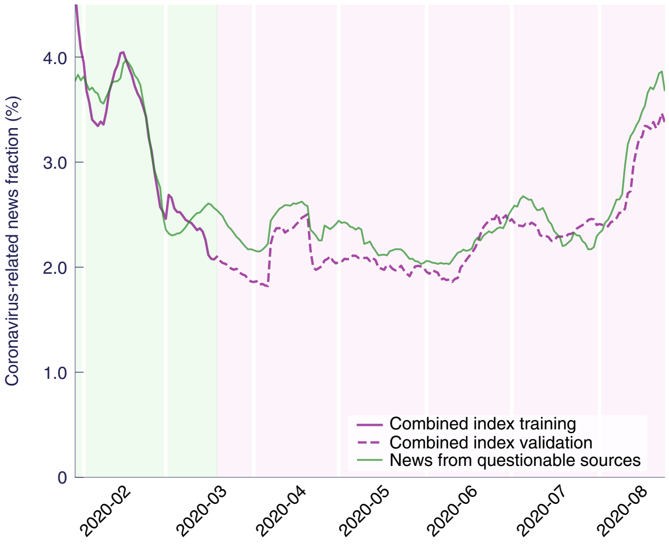

Fig. 3 | The combined index and the normalized time series of news whose sources were annotated as questionable (N_{QS}). The time series of N_{QS} was normalized using the total number of 'coronavirus'-related News compared with the combined index. The combined index is defined as a linear combination of the weighted modelling error for the local fitting of N_{AS} within the improved Vector Auto-Regression model and the cosine distance between the semantic vectors of Searches and News from All Sources. The parameters of the combination were fitted in the training set and then tested in the validation set. Green background represents the training set while pink background represents the validation set.

Discussion

Information quality is a fundamental challenge for the Information Age, especially during a pandemic. Studying the general news system and comparing it with the subset of news produced by sources labelled as questionable, we found that 'coronavirus'-related News from Questionable Sources production seems more reactive and precise than News from All Sources in addressing people's news demand. We exploited such a difference to develop an index for the vulnerability of specific topics to unreliable supplier takeover.

The analysis of Searches and News from All Sources for 'coronavirus' and a set of other 'coronavirus'-unrelated highly queried keywords exposed the interplay between news supply and demand: (1) a linear modelling scheme was effective in almost all cases; (2) the memory of the process seems to be very short (2–4 d) in all cases; (3) causality was more commonly directed from Searches to News from All Sources (for example, for 'coronavirus'). Due to these considerations, we developed an improved descriptive model to better describe the relationship between supply and demand for information. This modelling framework allowed us to discern how the inertia of news suppliers is one of the main traits of the dynamics for all the studied keywords. Also, the negative dependence on previous days' Searches observed in some cases suggests a dynamics where the trend of the interest is more important to news producers than the interest itself.

The comparison of 'coronavirus'-related News from All Sources and News from Questionable Sources, through the improved linear model's lens, exposed that News from Questionable Sources feature lower inertia and a different dependence on Searches, quantifying their more reactive behaviour. We can speculate that this behavioural difference could be a consequence of the different production environments of general suppliers and unreliable outlets. News from Questionable Sources are mainly produced by a large and well-established community of professional journalists, while News from Questionable Sources are the outcome of a scattered multitude of small, unorganized actors. The community size effect might be responsible for the different inertial behaviour observed. For example, a large community of mutually influencing journalists might take some time to reach a consensus on a new topic. In contrast, unreliable suppliers could freely publish just according to the most 'trending topics' without worrying about community codes of conduct.

The semantic analysis revealed another key difference between the production dynamics of News from All Sources and News from Questionable Sources. Looking at the shares of the most queried keywords co-occurring with 'coronavirus', we discovered that News from Questionable Sources is better aligned to Searches than News from All Sources, not only at a cumulative level but also daily over the entire observation period. This result suggests that people's interests are matched more precisely by unreliable suppliers than by general news producers. This difference might be explained by considering the different aims of the two communities. While they are both interested in answering people's demand for information, general news producers also strive for complete coverage of topics. In contrast, News from Questionable Sources focus on chasing people's attention.

We exploited the modelling and the semantic mismatch between News from All Sources and News from Questionable Sources to introduce two indexes to detect bursts of news production on 'coronavirus' from unreliable sources. It is worth mentioning how these indexes do not rely, in their definition, on any information about News from Questionable Sources. They are based instead only on the time series of News from All Sources and Searches. The first index is based on the modelling of News from All Sources and News from Questionable Sources, and exploits the goodness of fit of the modelling scheme. Since News from All Sources includes News from Questionable Sources, a higher presence of the latter could be revealed by a worse performance of the modelling scheme, quantified through the local weighted modelling error. A higher value of this indicator means that the normal relations between News from All Sources and Searches have been altered, presumably by the presence of News from Questionable Sources. The second index exploits the daily semantic misalignment between News from All Sources and Searches. In this case, a higher value of this indicator signals that semantic imprecision of News from All Sources leaves readers' interests unsatisfied, possibly fostering the production by unreliable sources.

The positive and meaningful correlation of both indicators with the concentration of News from Questionable Sources on 'coronavirus' supports two hypotheses. The first is that unreliable suppliers perturb the normal interplay between News from All Sources and Searches. The second is that unreliable suppliers are fuelled by the semantic misalignment between News from All Sources and Searches.

The two indices discussed above blend into a single combined index to assess the concentration of News from Questionable Sources on 'coronavirus'. We adopted a training set for its definition and a validation set to test its performance. The combined index is a good proxy for the activity of questionable sources and, although not a direct assessment of misinformation at the single news level, it can provide valuable insights to fight misinformation. Owing to its independence of annotations of News from Questionable Sources and its potential generalization to other topics, the combined index can be a powerful tool for journalists and editors on the one hand, and news monitoring authorities on the other, to detect, in real time, vulnerabilities to unreliable outlet production. Our results also suggest, as a possible strategy to tackle these vulnerabilities, a timely refocus of news supply to better meet the information demand of the public.

Information vulnerabilities are a major risk factor for our societies and they directly impact individuals in their behaviours and choices. For example, the solution to the coronavirus crisis heavily depends on individuals' behaviours, which in turn are directly affected by the news. The approach presented here, far from being conclusive, seeks to encourage a debate towards the understanding of the phenomenology of misinformation production as part of the information ecosystem's general dynamics. Additional studies will be needed to test the conclusions further and to generalize the results to different countries, languages, domains and time periods. Moreover, the diffusion layer should be added to the analysis of the dynamics of the infosphere, with particular attention to social media spreading of news. In our opinion, a paradigm shift in facing misinformation production is no longer an option. Instead, it is a pressing need, and we contend that the work we presented could contribute to the shift in scientific research towards a more concrete view, aiming to provide policymakers with knowledge and tools to prevent and fight misinformation supply.

Methods

Searches data. The information demand about a specific subject was obtained from Google Trends, a platform providing access to an anonymous sample of actual search requests made in the Google Search engine, from a selected location and time interval.

For each given keyword, Google Search returns a time series with values proportional to the number of times the keyword was searched each day. Since Google Search does not disclose the actual number of searches, the time-series values are rendered as percentages of the maximum number returned. As a result, data consist of integers within the interval 0–100. The time series of one keyword was referred to as the 'Searches' for that keyword and provided a measure of the interest it received.

The use of the 'pytrends' library for Python (https://github.com/GeneralMills/pytrends) enabled interaction with the Google Trends platform. The terms from Supplementary Fig. 1 were requested separately for the time ranging from 6 December 2019 to 31 August 2020 in Italy. These terms were 'coronavirus', 'regionali' (regional elections), 'playstation', 'papa francesco' (Pope Francesco), 'eurovision' (the European music contest) and 'sondaggi' (polls).

Google Trends also provided information about queries most searched with a specific keyword. In particular, the most popular queries related to the keyword 'coronavirus' (for example, 'coronavirus' news) were gathered. Such list is capped by Google Trends at a maximum of 25 related keywords, ordered by most searched to least, and denoted $q_1(t),\ldots,q_{25}(t)$, respectively, with t indicating the time and $q_0(t)$ being the time series of 'coronavirus' searches.

To compare the searches of a given keyword with its related keywords, it is necessary to put them on the same scale. To this end, searched items were queried in pairs. In this way, Google Trends normalized the two resulting time series to the highest of the maximums of the two. Given the two time series per request $(q_{i-1}(t), q_i(t))$, with $i = 1,\ldots,25$, a coefficient $\alpha_i = \max_t(q_{i-1}(t))/\max_t(q_i(t))$ was calculated. Thus, all the time series q_i could be set on the same scale of q_0, multiplying by $\prod_{j=1}^{i}\alpha_j$. This procedure was needed so as not to lose resolution on keywords with a small number of queries. Having queried for pairs $(q_0(t), q_i(t))$ would have resulted in a rounding at 0 performed by Google Trends.

'Coronavirus'-related queries were then aggregated by summing up their time series. Thus, 'coronavirus oggi' (coronavirus today), 'coronavirus notizie' (coronavirus news), 'coronavirus ultime' (coronavirus latest), 'coronavirus ultime notizie' (coronavirus latest news) and 'coronavirus news', were all aggregated into 'coronavirus news'. Subsequently, we removed all queries that returned the same search results as another query. These were 'coronavirus contagi' (coronavirus infections) and 'coronavirus in italia' (coronavirus in Italy), which are duplicates of 'contagi coronavirus' (coronavirus infections) and 'coronavirus italia' (coronavirus Italy), respectively. Also, the query 'corona' was excluded because it has other meanings in Italian, namely 'crown', and it is also a famous brand of beer. Finally, the list of queries associated with 'coronavirus', ordered by the amount of searches, was: 'news', 'italia' (Italy), 'lombardia' (Lombardy), 'sintomi' (symptoms), 'contagi' (infections), 'casi' (cases), 'morti' (deaths), 'bollettino' (bulletin), 'roma' (Rome), 'dati' (data), 'mondo' (world), 'mappa' (map), 'sicilia' (Sicily), 'veneto', 'campania', 'decreto' (decree), 'milano' (Milan) and 'piemonte' (Piedmont).

News data. To analyse the news supply, we investigated the data provided by AGCOM, the Italian Authority for Communications Guarantees, which granted us access to the content of a vast number of Italian news sources published online and offline from 6 December 2019 to 31 August 2020 in Italy. These data included articles from printed and digital newspapers and information agencies, TV, radio sites and scientific sources.

Moreover, the data had a specific annotation on questionable sources. AGCOM compiled a list of these outlets by merging the lists from independent fact-checking organizations such as bufale.net, butac.it, facta.news and pagellapolitica.it.

The protocols of these organizations for checking individual news consist of addressing only verifiable facts or numbers, comparing versions from different sources and tracking the history of the contents (for example, reverse searching pictures to check for possible misuse). More details on their procedures can be found on their websites (https://pagellapolitica.it/progetto/index or https://www.bufale.net/come-lavoriamo/) and on the Code of Principles they subscribe to (https://ifcncodeofprinciples.poynter.org).

For the classification of the sources, specific taxonomies have been developed. A source can be classified as questionable for different reasons. The most common are: (1) being a 'fake' version of an actual newspaper, such as the source 'Il Fatto Quotidaino' faking 'Il Fatto Quotidiano' by switching the letters in its domain; (2) supporting well-known conspiracy theories, such as 'Autismo Vaccini' (translated as autism vaccines); and (3) click-baiting websites, with fabricated news and exaggerated titles. More details can be found on the organizations' websites, where the lists of labelled sources are continuously updated (https://www.bufale.net/the-black-list-la-lista-nera-del-web/ or https://www.butac.it/the-black-list/).

The Authority verified the fact-checking organizations' methodologies and legitimacy through the recognition of international organizations such as the International Fact-checkers Organizations, the Duke Reporters' Lab or the European Digital Media Observatory. The Authority released the list of unreliable sources to its scientific partners only after it was used by independent scientific studies[11,64–67]. Also, the list provided by the Authority in 2020 was already adopted in one other study[68].

All the sources annotated as questionable in our dataset are listed in Supplementary List 1. The source-based methodology is well-known and well-established in the current literature on disinformation[8,69–71]. We followed the same approach, which is particularly well-suited to studying the behaviour of unreliable suppliers, as in the present study. As a sanity check, albeit without being comprehensive, we manually inspected only a small randomly chosen sample of the almost 40,000 news items available.

However, the source-based approach implies some limitations. Sources annotated as questionable might not publish just questionable news, and news from the annotated sources might be misinformation of different degrees. In our approach, the sources annotated as questionable are assumed to be questionable to the same extent and static for the observed period. In principle, the percentage of questionable news items might vary from one source to another and over time. Future studies will address these limitations. For the present work, we assumed that the source annotations are reliable enough to represent the questionable supply, at least at the aggregated level we considered.

We pre-processed the data for duplicates and incomplete logs elimination. In particular, we excluded items from Facebook and Twitter sources since our purpose is to monitor the direct production of news and social media usually copy contents created elsewhere. Also, an outlier was found in the pieces of news coming from a source called 'Non siamo soli', which were reported for only a few days and therefore excluded. After the cleaning, the News from All Sources data consisted of 6,806,881 items from 554 different news sources, while the News from Questionable Sources data consisted of 134,793 items. Each data entry has a unique ID and contains, among other information, the title and the content of the piece of news, its date, its source and the annotation of belonging to the questionable sources list.

Needing to imitate the rationale underlying Google Trends data, where daily search counts refer to the query of specific keywords, we sought to find counts of daily keywords also in the news data. To do so, given a keyword (for example, 'coronavirus'), we defined three different metrics: the piece of news containing the keyword at least once, those having the keyword at least three times, and finally, all the occurrences of a specific keyword. These three metrics were then normalized to the total number of news sources per day to level the press activity during weekends. For each model, we chose the metric with the best modelling performance. For the improved version of the VAR model described in equation (1) from the Results, the metric with at least three occurrences was selected, even if the other two showed similar performances. Instead, the most inclusive metric (at least one occurrence) was adopted when dealing with unreliable sources. This procedure was necessary to enhance the signal, given the low number of sources of News from Questionable Sources encountered. For consistency, News from All Sources was considered with the same metric (at least one occurrence) when comparing it with the News from Questionable Sources time series. The exact sample size (n) for each keyword is available in Table 3 as a discrete number for both sources of News from All Sources data and News from Questionable Sources data.

Following the same rationale, we adopted the first metric to filter for the keywords related to the 'coronavirus' subject described in the previous subsection. To do so, we selected the piece of news containing the keyword 'coronavirus' at least once and, in this subset, we counted the ones featuring the desired related keyword at least once. The values found were normalized to the total number of news pieces featuring the keyword 'coronavirus' at least once per day. We did this to get a proxy for the share of 'coronavirus' piece of information focused on the related keyword subdomain. We repeated this analysis for the subset of news mentioning the keyword 'coronavirus' at least once, coming from sources annotated as questionable. We then used the values extracted from this analysis to investigate the questionable supply in the 'coronavirus' context. The exact sample size (n) for each 'coronavirus'-related keyword is available in Table 5 as a discrete number for both News from All Sources and News from Questionable Sources.

Table 5 | The exact sample size (*n*) for each keyword, given as a discrete number for both N_{AS} data and N_{QS} data

Keyword	(*n*) for N_{AS}	(*n*) for N_{QS}
	#occ. ≥1	#occ. ≥1
'italia'	625,099	18,572
'bollettino'	56,294	1,731
'dati'	299,718	7,943
'lombardia'	166,449	4,165
'roma'	384,837	6,697
'morti'	176,950	5,936
'contagi'	469,594	13,665
'mondo'	290,069	6,889
'casi'	434,620	10,362
'sintomi'	122,293	3,767
'sicilia'	70,363	1,465
'campania'	45,762	1,661
'veneto'	108,043	2,005
'mappa'	23,334	407
'piemonte'	48,883	1,313
'milano'	199,063	3,274
'decreto'	141,274	2,776

The most inclusive metric (in the table indicated as '#occ. ≥1') was used for the COVID-19-related keywords for both N_{AS} data and N_{QS} data.

Time series analysis. Time series of Searches and News from All Sources (from Supplementary Fig. 1) were investigated with the VAR model[62], using Python's 'statsmodels' package for time-series analysis[72]. Data were regularized via $x \mapsto \log(1+x)$ transformation before fitting. For the VAR modelling, the number of lags d was determined as the parameter that minimized the Akaike information criterion[63], with d ranging from 1–14. This modelling strategy was chosen to ensure the interpretability of the fitted model and its regression coefficients.

From the VAR model, we computed Granger-causality[62] to test whether the queries' values provided meaningful information to the prediction of news shares and vice versa. Since two tests were performed on the same data from a given subject (for the null hypotheses, 'S does not Granger-cause N_{AS}' and 'N_{AS} does not Granger-cause S'), resulting P values were corrected by the Holm-Bonferroni method[73]. Thus, pairs of P values in Supplementary Fig. 2 were multiplied by 2 to control for family-wise error rate and to maintain comparability.

In Fig. 1 and Supplementary Fig. 3, the improved models for regression of the News from All Sources were derived by adjusting the VAR models to include Searches at time t. Lags were re-elaborated through the Akaike criterion as before, with similar results. These models were then compared against a null model that forecasts one day proportionally to the value of the day before to benchmark how beneficial the addition of regressing variables was to N_{AS} prediction (Table 1).

To assess the semantic misalignment between News from All Sources and Searches from Supplementary Fig. 4 over the related queries associated to 'coronavirus' at a given time t, the cosine distance was calculated as $d(S(t), N(t)) = 1 - S(t) \cdot N(t)/|S(t)|\,|N(t)|$, on the vectors $S(t) = (S_1(t), \ldots, S_k(t))$, $N(t) = (N_1(t), \ldots, N_k(t))$, where $S_i(t)$ and $N_i(t)$ represented the searches and news, respectively, at time t for the i-th keyword associated to 'coronavirus', with $\cdot$ being the dot product and $|\cdot|$ being the Euclidean norm. Cosine distance was suitable for comparing high-dimensional vectors at different scales, and returned values in (0, 1) for vectors with non-negative entries such as $S(t)$ and $N(t)$.

Comparison of improved models' coefficients. To assess the differences in the coefficients from the improved models of News from All Sources or News from Online Sources, and News from Questionable Sources, we performed a statistical hypothesis test on the basis of bootstrap. First, we created a bootstrap version of the daily values by sampling with replacement from the pool of news from those days that were unlabelled or labelled as questionable. Repeating this procedure many times for every day, we recreated 10^5 bootstrapped versions of the time series for News from All Sources (or News from Online Sources) and News from Questionable Sources. For each of them, we fitted the improved model and calculated Δ as the difference between the parameters of the model for the News from All Sources (or News from Online Sources) and the same parameters of the model for the News from Questionable Sources. As an example, we report in Supplementary Fig. 6 the distribution of the difference in the inertial term α_1 of News from All Sources and News from Questionable Sources, that is, the bootstrap distribution. To challenge the null hypothesis $\Delta \le 0$ against the alternative $\Delta > 0$, we calculated the test's P value as $P = |\Delta \le 0|/(N+1)$, where $|\Delta \le 0|$ is the number of bootstrap repeats where the hypothesis is true and $N = 10^5$ is the total number of repeats (note that 1 is added to the denominator to account for the $\Delta > 0$ that is actually observed). The example above describes the left-tailed version of the testing procedure. Following the same rationale, the P value for the right-tailed test is $P = |\Delta \ge 0|/(N+1)$. The results for this analysis are reported in Table 4.

Combined index validation. To define and validate the combined index from Fig. 3, we split the daily data from News from Questionable Sources concentration on 'coronavirus' into a training set (from 29 January 2020 to 20 March 2020) and a validation set (from 21 March 2020 onwards).

Thus, we defined the combined index as a linear combination of the two starting indices that best fitted the concentration of News from Questionable Sources, using a linear model with Gaussian noise on the training data. The ordinary least squares estimate $\hat{\sigma}$ for the variance of the Gaussian noise was then calculated as the mean squared error (MSE) divided by the statistical degrees of freedom k (that is, the number of observations minus 2, which is the number of parameters in the model).

To assess the predictive potential of the combined index, we adopted the trained model to forecast the concentration of News from Questionable Sources in the validation set. The goodness of fit of this prediction was tested through the reduced chi-squared statistic, which is calculated as the MSE on the validation set divided by $\hat{\sigma}$. This statistic is approximately distributed as a χ^2 with as many degrees of freedom as the size of the validation set (that is, 51), leading to a P value of about 0.945. As such, the null hypothesis that the concentration of News from Questionable Sources for the keyword 'coronavirus' is distributed in agreement with the trained model cannot be rejected.

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

Data availability

Google Search engine data were generated by the Google Trends platform and is publicly available at https://trends.google.com. The news raw data are not publicly available due to copyright restrictions. Derived data about news and searches supporting the findings of this study are available at https://github.com/SonyCSLParis/news_searches.

Code availability

All codes for data analysis and for data gathering searches and preparation are available at https://github.com/SonyCSLParis/news_searches.

Received: 31 May 2021; Accepted: 14 April 2022;
Published online: 23 May 2022

References

1. Funke, D. & Flamini, D. A guide to anti-misinformation actions around the world. *Poynter* https://www.poynter.org/news/guide-anti-misinformation-actions-around-world (2018).
2. Lazer, D. et al. The science of fake news. *Science* **359**, 1094–1096 (2018).
3. Tandoc Jr, E. The facts of fake news: a research review. *Sociol. Compass* **13**, e12724 (2019).
4. Fallis, D. What is disinformation? *Libr. Trends* **63**, 401–426 (2015).
5. Bakir, V. & McStay, A. Fake news and the economy of emotions: problems, causes, solutions. *Digit. Journal.* **6**, 154–175 (2018).
6. Pennycook, G. & Rand, D. G. The psychology of fake news. *Trends Cogn. Sci.* **25**, 388–402 (2021).
7. Cinelli, M. et al. The COVID-19 social media infodemic. *Sci. Rep.* **10**, 1–10 (2020).
8. Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B. & Lazer, D. Fake news on Twitter during the 2016 US presidential election. *Science* **363**, 374–378 (2019).
9. Shu, K., Sliva, A., Wang, S., Tang, J. & Liu, H. Fake news detection on social media: a data mining perspective. *SIGKDD Explor.* **19**, 22–36 (2017).
10. Conroy, N., Rubin, V. & Chen, Y. Automatic deception detection: methods for finding fake news. *Proc. Assoc. Inf. Sci. Technol.* **52**, 1–4 (2015).
11. Vicario, M., Quattrociocchi, W., Scala, A. & Zollo, F. Polarization and fake news: early warning of potential misinformation targets. *ACM Trans. Web* **13**, 1–22 (2019).
12. Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W. & Starnini, M. The echo chamber effect on social media. *Proc. Natl Acad. Sci. USA* https://www.pnas.org/content/118/9/e2023301118 (2021).
13. Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A., Eckles, D. & Rand, D. Shifting attention to accuracy can reduce misinformation online. *Nature* **592**, 590–595 (2021).

14. Visentin, M., Pizzi, G. & Pichierri, M. Fake news, real problems for brands: the impact of content truthfulness and source credibility on consumers' behavioral intentions toward the advertised brands. *J. Interact. Mark.* **45**, 99–112 (2019).
15. Allcott, H. & Gentzkow, M. Social media and fake news in the 2016 election. *J. Econ. Perspect.* **31**, 211–36 (2017).
16. Treen, K., Williams, H. & O'Neill, S. Online misinformation about climate change. *Wiley Interdiscip. Rev. Clim. Change* **11**, e665 (2020).
17. Kata, A. A postmodern Pandora's box: anti-vaccination misinformation on the Internet. *Vaccine* **28**, 1709–1716 (2010).
18. Rothkopf, D. When the buzz bites back. *Washington Post* **11**, B1–B5 (2003).
19. Islam, M. et al. COVID-19-related infodemic and its impact on public health: a global social media analysis. *Am. J. Trop. Med. Hyg.* **103**, 1621–1629 (2020).
20. Dafouz-Milne, E. The pragmatic role of textual and interpersonal metadiscourse markers in the construction and attainment of persuasion: a cross-linguistic study of newspaper discourse. *J. Pragmat.* **40**, 95–113 (2008).
21. Catenaccio, P. et al. Towards a linguistics of news production. *J. Pragmat.* **43**, 1843–1852 (2011).
22. Ryfe, D. The economics of news and the practice of news production. *Journal. Stud.* **22**, 60–76 (2021).
23. Schmidt, A., Ivanova, A. & Schäfer, M. Media attention for climate change around the world: a comparative analysis of newspaper coverage in 27 countries. *Global Environ. Change* **23**, 1233–1248 (2013).
24. Sznitman, S. & Lewis, N. Is cannabis an illicit drug or a medicine? A quantitative framing analysis of Israeli newspaper coverage. *Int. J. Drug Policy* **26**, 446–452 (2015).
25. Wirz, C. et al. Media systems and attention cycles: volume and topics of news coverage on COVID-19 in the United States and China. *Journal. Mass Commun. Q.* https://doi.org/10.1177%2F10776990211049455 (2021).
26. Korobchinsky, M., Chyrun, L., Chyrun, L. & Vysotska, V. Peculiarities of content forming and analysis in internet newspaper covering music news. 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT) 52–57 (IEEE, 2017).
27. Costera Meijer, I. & Bijleveld, H. Valuable journalism: measuring news quality from a user's perspective. *Journal. Stud.* **17**, 827–839 (2016).
28. Cushion, S., Morani, M., Kyriakidou, M. & Soo, N. (Mis) understanding the coronavirus and how it was handled in the UK: An analysis of public knowledge and the information environment. *Journal. Stud.* **23**, 703–721 (2022).
29. Kennedy, P. & Prat, A. 'Where Do People Get Their News?'. London, Centre for Economic Policy Research. https://cepr.org/active/publications/discussion_papers/dp.php?dpno=12426 (2017).
30. Tewksbury, D. What do Americans really want to know? Tracking the behavior of news readers on the Internet. *J. Commun.* **53**, 694–710 (2003).
31. Engle, R. & Ng, V. Measuring and testing the impact of news on volatility. *J Finance* **48**, 1749–1778 (1993).
32. Harteveld, E., Schaper, J., De Lange, S. & Van Der Brug, W. Blaming Brussels? The impact of (news about) the refugee crisis on attitudes towards the EU and national politics. *J. Common Mark. Stud.* **56**, 157–177 (2018).
33. Tewksbury, D. The seeds of audience fragmentation: specialization in the use of online news sites. *J. Broadcast. Electron. Media* **49**, 332–348 (2005).
34. Costera Meijer, I. The paradox of popularity: how young people experience the news. *Journal. Stud.* **8**, 96–116 (2007).
35. Groot Kormelink, T. & Costera Meijer, I. Tailor-made news: meeting the demands of news users on mobile and social media. *Journal. Stud.* **15**, 632–641 (2014).
36. Lee, A. & Chyi, H. When newsworthy is not noteworthy: examining the value of news from the audience's perspective. *Journal. Stud.* **15**, 807–820 (2014).
37. Althaus, S., Cizmar, A. & James, G. Gimpel media supply, audience demand, and the geography of news consumption in the United States. *Polit. Commun.* **26**, 249–277 (2009).
38. Peters, C. Journalism to go: the changing spaces of news consumption. *Journal. Stud.* **13**, 695–705 (2012).
39. Sheller, M. News now: interface, ambience, flow, and the disruptive spatio-temporalities of mobile news media. *Journal. Stud.* **16**, 12–26 (2015).
40. Webster, J.G. & Ksiazek, T. B. The dynamics of audience fragmentation: public attention in an age of digital media. *J. Commun.* **62**, 39–56 (2012).
41. Lee, A. News audiences revisited: theorizing the link between audience motivations and news consumption. *J. Broadcast. Electron. Media* **57**, 300–317 (2013).
42. Acerbi, A. Cognitive attraction and online misinformation. *Palgrave Commun.* **5**, 1–7 (2019).
43. Trussler, M. & Soroka, S. Consumer demand for cynical and negative news frames. *Int. J. Press Polit.* **19**, 360–379 (2014).
44. Iyengar, S., Norpoth, H. & Hahn, K. Consumer demand for election news: the horserace sells. *J. Polit.* **66**, 157–175 (2004).
45. Boczkowski, P. & Peer, L. The choice gap: the divergent online news preferences of journalists and consumers. *J. Commun.* **61**, 857–876 (2011).
46. Jun, S., Yoo, H. & Choi, S. Ten years of research change using Google Trends: from the perspective of big data utilizations and applications. *Technol. Forecast. Soc. Change* **130**, 69–87 (2018).
47. Sullivan, D. How Google delivers reliable information in Search. *Google. The Keyword* https://blog.google/products/search/how-google-delivers-reliable-information-search/2020 (accessed 26 November 2021).
48. Da, Z., Engelberg, J. & Gao, P. The sum of all FEARS investor sentiment and asset prices. *Rev. Financ. Stud.* **28**, 1–32 (2015).
49. Preis, T., Moat, H. & Stanley, H. Quantifying trading behavior in financial markets using Google Trends. *Sci. Rep.* **3**, 1684 (2013).
50. Lampos, V., Miller, A. C., Crossan, S. & Stefansen, C. Advances in nowcasting influenza-like illness rates using search query logs. *Sci. Rep.* **5**, 1–10 (2015).
51. Dugas, A. et al. Influenza forecasting with Google flu trends. *PloS ONE* **8**, e56176 (2013).
52. Strzelecki, A. The second worldwide wave of interest in coronavirus since the COVID-19 outbreaks in South Korea, Italy and Iran: a Google Trends study. *Brain Behav. Immun.* **88**, 950–951 (2020).
53. Choi, H. & Varian, H. Predicting the present with Google Trends. *Econ. Rec.* **88**, 2–9 (2012).
54. Borup, D. & Schütte, E. In Search of a Job: Forecasting Employment Growth Using Google Trends, Journal of Business & Economic Statistics, 40:1, 186-200, https://doi.org/10.1080/07350015.2020.1791133 (2022).
55. Nghiem, L. T. P., Papworth, S. K., Lim, F. K. S., & Carrasco, L. R., Analysis of the capacity of Google Trends to measure interest in conservation topics and the role of online news. *PLoS ONE* https://doi.org/10.1371/journal.pone.0152802 (2016).
56. Broder, A. A taxonomy of web search. *ACM Sigir Forum* **36**, 3–10 (2002).
57. Waller, V. Not just information: who searches for what on the search engine Google? *J. Am. Soc. Inf. Sci. Technol.* **62**, 761–775 (2011).
58. Chevalier, A., Dommes, A. & Marquié, J. Strategy and accuracy during information search on the Web: effects of age and complexity of the search questions. *Comput. Human Behav.* **53**, 305–315 (2015).
59. Grimmelmann, J. The google dilemma. *NY Law Sch. Law Rev.* **53**, 939 (2008).
60. Weeks, B. & Southwell, B. The symbiosis of news coverage and aggregate online search behavior: Obama, rumors, and presidential politics. *Mass Commun. Soc.* **13**, 341–360 (2010).
61. Trielli, D. & Diakopoulos, N. Search as news curator: the role of Google in shaping attention to news information. In *Proc. 2019 CHI Conference on Human Factors in Computing Systems* 1–15 (ACM, 2019).
62. Hamilton, J. *Time Series Analysis* (Princeton Univ. Press, 1994).
63. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* **19**, 716–723 (1974).
64. *News VS Fake Nel Sistema Dell'informazione* (AGCOM, 2018); https://www.agcom.it/documents/10179/12791486/Pubblicazione+23-11-2018/93869b4f-0a8d-4380-aad2-c10a0e426d83,2018,11
65. Fletcher, R., Cornia, A., Graves, L. & Nielsen, R. Measuring the reach of "fake news" and online disinformation in Europe. *Australasian Policing* **10**(2), (2018).
66. Ciampaglia, G., Mantzarlis, A., Maus, G. & Menczer, F. Research challenges of digital misinformation: toward a trustworthy web. *AI Mag.* **39**, 65–74 (2018).
67. Caldarelli, G., De Nicola, R., Del Vigna, F., Petrocchi, M. & Saracco, F. The role of bot squads in the political propaganda on Twitter. *Commun. Phys.* **3**, 81 (2020).
68. Cinelli, M. et al. Dynamics of online hate and misinformation. *Sci. Rep.* **11**, 22083 (2021).
69. Vosoughi, S., Roy, D. & Aral, S. The spread of true and false news online. *Science* **359**, 1146–1151 (2018).
70. Bovet, A. & Makse, H. A. Influence of fake news in Twitter during the 2016 US presidential election. *Nat. Commun.* **10**, 1–14 (2019).
71. Pierri, F., Piccardi, C. & Ceri, S. Topology comparison of Twitter diffusion networks effectively reveals misleading information. *Sci. Rep.* **10**, 1372 (2020).
72. Seabold, S. & Perktold, J. statsmodels: econometric and statistical modeling with Python. In *Proc. 9th Python in Science Conference* (statsmodels, 2010).
73. Lehmann, E. & Romano, J. *Testing Statistical Hypotheses* (Springer-Verlag, 2006).

Acknowledgements

We thank M. Delmastro of AGCOM for insightful discussions about the Italian news ecosystem as well as for providing access to the database of Italian news. The database was shared in the framework of the Task Force on 'Digital Platforms and Big Data - Covid-19 Emergency', established by the AGCOM to contribute, among other things, to the fight against online disinformation on issues related to the COVID-19 crisis. The authors received no specific funding for this work.

Author contributions

P.G. directed the project. P.G., G.P. and M.G. wrote the code and processed the data. P.G. and G.P. designed the statistical analysis. P.G., G.P., M.G. and V.L. contributed to the design and implementation of the research, the analysis of the results and the writing of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-022-01353-3.

Correspondence and requests for materials should be addressed to Pietro Gravino.

Peer review information *Nature Human Behaviour* thanks Vian Bakir and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

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.

nature portfolio

Corresponding author(s): Pietro Gravino

Last updated by author(s): Apr 4, 2022

Reporting Summary

Nature Portfolio 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 Portfolio 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	Custom Python codes for data gathering and preparation is available at https://github.com/SonyCSLParis/news_searches Versions: Python 3.8.10 unidecode 1.3.4 pytrends 4.8.0 pandas 1.3.5 numpy 1.21.0 matplotlib 3.5.1
Data analysis	Custom Python codes for data analysis is available at https://github.com/SonyCSLParis/news_searches Versions: Python 3.9.7 re 2.2.1 numpy 1.20.3 pymysql 1.0.2 matplotlib 3.4.3 json 2.0.9 csv 1.0 pandas 1.3.4 seaborn 0.11.2 scipy 1.7.1 statsmodels.api 0.12.2 matplotlib 3.4.3

scipy 1.7.1

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 Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Google search engine data was generated by the Google Trends platform and is publicly available at https://trends.google.com. The raw data are not publicly available due to copyright restrictions. Derived data about news and about searches supporting the findings of this study is available at https://github.com/SonyCSLParis/news_searches

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

Behavioural & social sciences study design

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

Study description	The supply and demand of information is studied by comparing the daily time series for the numbers of news items produced in Italy and searches from Google Search engine. News and searches were filtered based on selected tokens/keywords to represent information supply and demand for specific subjects.
Research sample	News data covered most of the media production in Italy. Searches data were proportional to the totality of a query's volume in Italy.
Sampling strategy	No sampling strategy was designed. No sample-size calculation was performed, as data size was large.
Data collection	Italian news articles production was provided by AGCOM, the Italian Authority for Communications Guarantees. Google's searches data were collected using Google Trends platform's API.
Timing	Start: 6/10/2019. End: 31/08/2020.
Data exclusions	We pre-processed the data for duplicates and incomplete logs elimination. Particularly, we excluded items coming from Facebook and Twitter sources, since our purpose is to monitor direct production of news and usually social media copy contents created elsewhere. Also, an outlier was found in the pieces of news coming from a source called ``Non siamo soli'', that were reported just for few days, and were therefore excluded.
Non-participation	No information are available about dropping out of news sources (for news data) or Google Search Engine users (for searches data) but we believe this effect to be negligible in the observed time-frame over the size of the observed communities (basically, the ensemble of news outlets and Google users).
Randomization	Participants were not allocated into groups.

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	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
☒	□ Human research participants
☒	□ Clinical data
☒	□ Dual use research of concern

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging