Applied Network Science

RESEARCH

Open Access

Online news ecosystem dynamics: supply, demand, diffusion, and the role of disinformation

Pietro Gravino[1,3*], Giulio Prevedello[1,3] and Emanuele Brugnoli[2,3]

*Correspondence:
pietro.gravino@sony.com

[1] Sony Computer Science Laboratories - Paris, 6, Rue Amyot, Paris 75005, France
[2] Sony Computer Science Laboratories - Rome, Joint Initiative CREF-SONY, Centro Ricerche Enrico Fermi, Via Panisperna 89/A, Rome 00184, Italy
[3] Enrico Fermi's Research Center, Via Panisperna 89/A, Rome 00184, Italy

Abstract

The digital age provides new challenges as information travels more quickly in a system of increasing complexity. But it also offers new opportunities, as we can track and study the system more efficiently. Several studies individually addressed different digital tracks, focusing on specific aspects like disinformation production or content-sharing dynamics. In this work, we propose to study the news ecosystem as an information market by analysing three main metrics: Supply, Demand, and Diffusion of information. Working on a dataset relative to Italy from December 2019 to August 2020, we validate the choice of the metrics, proving their static and dynamic relations, and their potential in describing the whole system. We demonstrate that these metrics have specific equilibrium relative levels. We reveal the strategic role of Demand in leading a non-trivial network of causal relations. We show how disinformation news Supply and Diffusion seem to cluster among different social media platforms. Disinformation also appears to be closer to information Demand than the general news Supply and Diffusion, implying a potential danger to the health of the public debate. Finally, we prove that the share of disinformation in the Supply and Diffusion of news has a significant linear relation with the gap between Demand and Supply/Diffusion of news from all sources. This finding allows for a real-time assessment of disinformation share in the system. It also gives a glimpse of the potential future developments in the modelisation of the news ecosystem as an information market studied through its main drivers.

Keywords: News ecosystem, Disinformation, Complex systems, Network science

Introduction

Internet and social media have significantly transformed how people access, share, and consume information. While digital environments have considerably promoted disintermediation, enabling diverse voices to participate in the collective dialogue at the expense of professional information, the role of leader nodes in social networks (i.e., the main influential accounts) remains crucial in determining how information is disseminated and consumed [1-3]. Recent works on the dynamics of information dissemination and consumption have surged interest in the complexity of information ecosystems, particularly focusing on disinformation from its very definition (Kapantai et al. 2021; Lazer

Springer Open

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

et al. 2018) to its spread (Del Vicario et al. 2016) and connection to partisanship (Garrett and Bond 2021; Pennycook and Rand 2019). A significant portion of research has examined the impact of disinformation on human behaviour (Bastick 2021), political elections (Morgan 2018), sustainability (Treen et al. 2020), and health (Sasahara et al. 2021). The term 'Infodemic' (Simon and Camargo 2023), which resurfaced during the Covid-19 pandemic (Cinelli et al. 2020), describes the overwhelming flood of both accurate and false information about the virus, leading to confusion and harmful behaviours that exacerbated the pandemic (Rocha et al. 2023). These investigations have led to questions about identifying statistical indicators in news content and, consequently, effective strategies for preventing the spread of disinformation (Del Vicario et al. 2019; Guay et al. 2023; Pacheco et al. 2020).

The broader information ecosystem, constituted by both communities of news producers (e.g. journalists, editors, etc.) and consumers (readers), has received far less attention than the disinformation phenomenology. Few attempts have been made to study the dynamics of interaction between news producers and consumers (King 1998), and we still lack a fundamental understanding of the system. A previous work Gravino et al. (2022) identified the news supply on the production side and the news demand on the consumption side as the main drivers of the systemic dynamics.

The news supply, referred to as Supply in this work, can be defined as the ensemble of contents produced by news outlets over a certain period (e.g. daily). The news demand, referred to as Demand, can be defined over the same period as the cumulative needs and interests of the consumer community. These two quantities are directly linked in many ways. Consumers are more likely to engage with and demand content that aligns with their interests, preferences, and values (Cinelli et al. 2020). At the same time, the type and nature of news content also impact how widely it is shared and diffused. Engaging content significantly influences diffusion as followers are more likely to share relevant information (Turcotte et al. 2015). Producers are likely to cater to topics and formats that align with the interests and demands of their audience (Thurman et al. 2019). At the same time, the demand for certain types of content influences how widely it is shared. If a particular piece of news resonates strongly with the audience, it is more likely to be widely diffused by followers (Thompson et al. 2020).

From these interactions, a third layer of the dynamics emerges: the diffusion of the news (Diffusion). Supply cannot meet Demand if the news is not shared through some channel. Different media are used, ranging from paper to word-of-mouth, from TV to news websites but, in recent years, social media has been the most prominent. Furthermore, the disintermediation typical of these platforms allows us to study the dynamics in unprecedented ways. We can define Diffusion as the sharing volume of news content over the observation period. This layer of the dynamics is perhaps the most studied in the last years, e.g. for its commercial importance. News discussion can influence future content creation strategies. News producers may observe what types of content are gaining traction through diffusion and adjust their production accordingly (Andrews and Caren 2010). At the same time, the Diffusion affects what other consumers are exposed to and, consequently, what they may demand. Popular content that has been widely diffused may generate increased demand from new audiences (Iyengar et al. 2004). On the other hand, sharing behaviour can sometimes result in the formation of echo chambers,

i.e. user groups that share a common narrative (Brugnoli et al. 2019; Pratelli et al. 2023), in persistent recurring patterns (Desiderio et al. 2023), or in self-organised collective actions (Mancini et al. 2022).

Separately, Supply, Demand and Diffusion have been subjects of several studies (Brugnoli et al. 2023; Patuelli and Saracco 2023; Mattei et al. 2022; Hohenberg 2023), but together, they could provide a deeper and systemic understanding of the news ecosystem. Still, it has to be proved that their mutually influenced interplay underscores the complex dynamics of the news ecosystem.

Our study connects the dots between news Demand, Supply, and Diffusion, analysing the news ecosystem as a single complex system. We aim to prove that the chosen metrics effectively track the system and account for the phenomenology. Finally, we show how this approach helps better understand the system's health status, assessing disinformation production and spreading levels. This work aims to confirm and generalise some of the results that emerged in a previous study (Gravino et al. 2022). The previous observations will be expanded to include more keywords and different social networks, and model-agnostic techniques will be adopted to provide further generalisation. We select the main news outlets in Italy, encompassing a wide range of news media outlets active from December 2019 to August 2020. We monitor their posts' production and users' sharing volumes concerning the most relevant keywords in the observed period. These keywords have been identified by looking at the most important keywords used in the Google Search Engine, which has also been used to track the Demand for information. The dynamics of these interactions can differ among social media platforms because of variations in their business models and content selection algorithms (De Marzo et al. 2023; Tommasel and Menczer 2022). For these reasons, we focused on the two main social media in the considered time frame: Facebook and X. The latter will be referred to as Twitter, as this was still the name at the time of data gathering.

Results and discussion

In this work, the information ecosystem is studied as a market driven by three main metrics: Supply, Demand, and Diffusion of information. We will formally define these quantities and we will show how they are related in terms of scales and dynamics without assuming any specific model. Then, we will show how the relation between these forces can be used to provide useful insights about the health status of the information system, providing an independent assessment of the Non-Trustworthy levels of information supplied and diffused.

The three forces of the news ecosystem

First, we formally introduce the three quantities that are the main subjects of our investigation: Supply, Demand, and Diffusion. In our work, we consider these quantities as the main metrics to identify the collective status of the news ecosystem and its behaviour. We will call these metrics forces as they influence the system's status and each other as we will unravel in this work.

The news Demand represents the aggregated need for information in the community. Let's imagine that, at a given moment, every community member would want to know more about one or more subjects. If we identify every subject by a distinctive, unique

keyword, we could associate to every member a list of keywords corresponding to the individual needs. We could then aggregate all the lists, count all the occurrences of the keywords, and represent the result as a vector:

$$\overline{Dem}(t) = (Dem_{k_1}(t), Dem_{k_2}(t), ..., Dem_{k_N}(t)) \tag{1}$$

where each component $Dem_{k_i}(t)$ corresponds to the count for a given keyword k_i and a given day t. $\overline{Dem}(t)$ represents the collective news Demand at time t. Of course, such information is inaccessible unless we survey the whole population simultaneously. In principle, the time resolution can vary, as we could aggregate the Demand not instantaneously but daily or over different time windows. Also, instead of surveying the population, we can use the keywords input in web search engines. In this work, as a proxy for Demand, we will use the daily time-series of keyword searches on Google that have been collected from the Google Trends platform with a procedure that allows us to elaborate an absolute scale. Such scale allows comparisons between different keywords Demand (see Sect. 3). This proxy is limited to the 14 most important independent keywords (see Sect. 3), but it still grasps the most important aspects of the dynamics, as we will show.

Supply represents the aggregated production of information in the community. While every actor can produce information, we want to focus our analysis on the most strategic and influential actors, i.e. news outlets. Sticking to the daily time resolution, we can define a vector similar to Eq. 1:

$$\overline{Sup}(t) = (Sup_{k_1}(t), Sup_{k_2}(t), ..., Sup_{k_N}(t)) \tag{2}$$

where each component $Sup_{k_i}(t)$ corresponds to a keyword and counts how many pieces of news concerning the subject identified by k_i have been produced during the day t. We filter and count the components of those corresponding to the 14 most searched keywords (see Sect. 3). We elaborate the Supply of news by aggregating the posts' published from an extensive list of Italian news outlets' profiles on the two main social media (Facebook and X/Twitter). Social media posts' production is strictly linked with their overall news production (see Sect. 3) and provides a natural link with the third force.

Diffusion represents the aggregated spreading of the pieces of news in the community. Given our definition of Supply Eq. 2 and its proxy, it follows naturally that the vector:

$$\overline{Dif}(t) = (Dif_{k_1}(t), Dif_{k_2}(t), ..., Dif_{k_N}(t)) \tag{3}$$

represents the Diffusion, with each component $Dif_{k_i}(t)$ corresponding to a keyword and counting how many times pieces of news concerning the subject identified by k_i have been shared during the day t. This operative definition strongly depends on the social media adopted in the study. We calculated the Diffusion by summing the shares of the news posts in the Supply for both monitored social media (see Sect. 3). All forces are monitored for the most prominent keywords in Italy from December 2019 to August 2020. More details are provided in the Sect. 3. We report the cumulative sum of the three forces over the monitored period in Fig. 1. Unless differently specified, the same force on different social media will be treated in the analysis as two different forces for two reasons. First, the two different social media are used by different communities in different ways, and we want to give an account about that. Second, if the relations

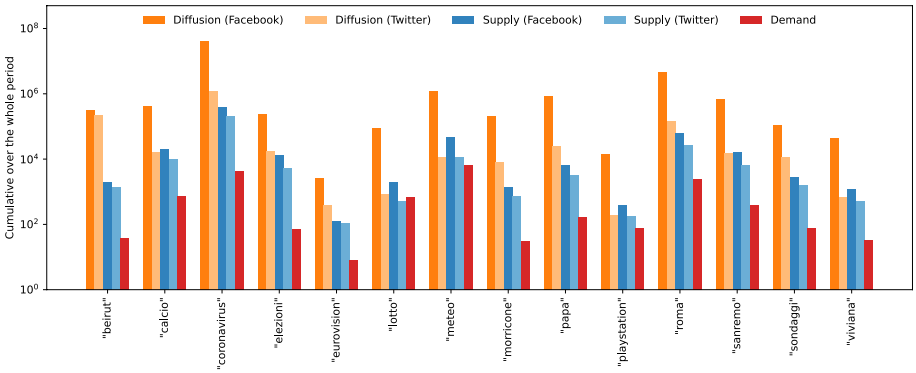

Fig. 1 The cumulative count over the whole period (from December '19 to August '20) of the three forces where Supply and Diffusion are reported for both Facebook (FB) and Twitter(TW) for all the keywords. The order of the forces is almost always the same for all keywords

between the forces are the same and are independent of the platform, we should observe this, so treating them independently will serve as validation of the existence of deeper relations between the forces.

While the magnitude of the cumulative sum varies between the different keywords, the order of the cumulative sum of forces (from the larger to the smaller) for a given keyword is almost always the same. This suggests the existence of relations between the forces.

Correlations and relative scale

To start studying the relations between the forces, we measure their mutual correlation. We look at the logarithm of monthly aggregation of the forces in both social networks for all different keywords because. We report the results in Fig. 2. In all cases, correlation coefficients are high and significant, and the forces can be considered linearly related. The monthly aggregation has been chosen to avoid disturbance from the dynamics in shorter windows that might cause larger fluctuations, but similar results are also observed for weekly and daily aggregation, as reported in SI. Correlation is not causality, so this still does not prove that there are direct relations between the forces.

Still, the intercept of the linear regressions of the logarithms suggests that there are typical relative scales that we can measure directly without implying a linear model. We chose the median force in terms of order of magnitude to be used as an offset (the Supply on Facebook) and normalise by that every monthly aggregate value for each force, for all different keywords. We report the results in Fig. 3.

For comparison, we also reported the histogram of the monthly aggregated values of the forces. We also report the mean and standard values for all distributions. As can be seen by comparing the standard deviations, the scale relations between the forces are much more narrow than the original distributions. This means that the forces are not only correlated but also refer to dynamics with typical scales tied by precise relation relations. In other words, if these forces were at equilibrium, given one of them, we could calculate the value that all other forces should have. Still, we cannot discuss equilibrium relations without proving dynamic relations between the forces, which is the subject of the next section.

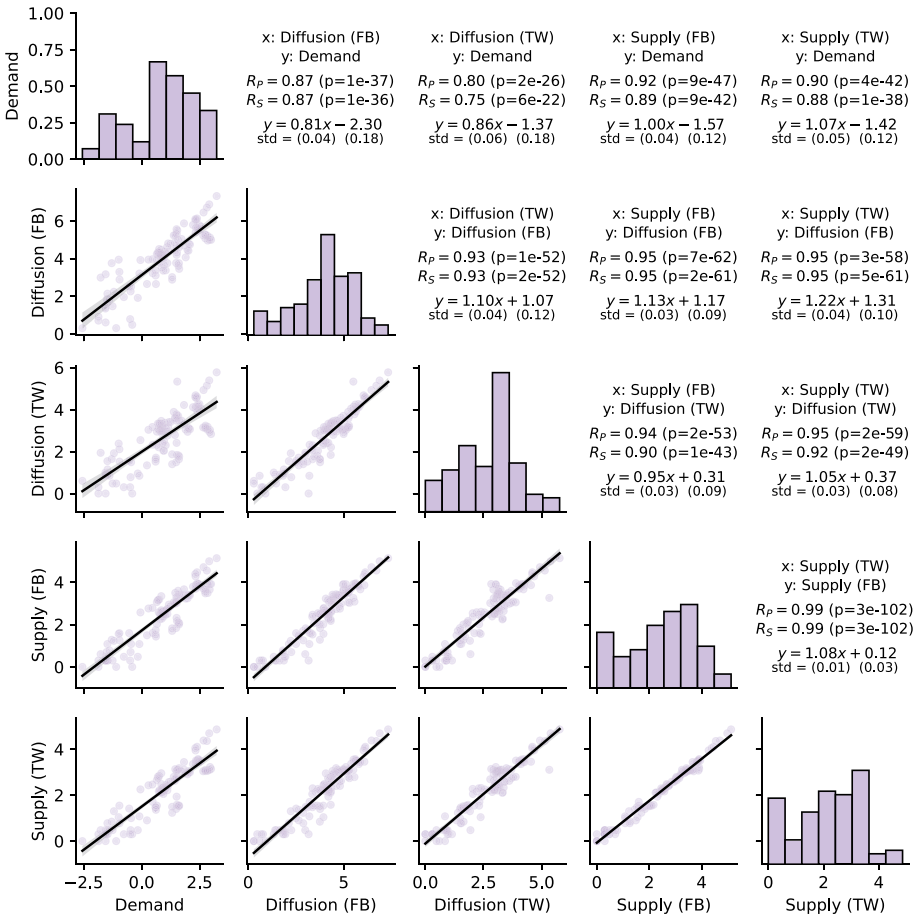

Fig. 2 Correlation (Pearson and Spearman) and linear regressions between the logarithms of the monthly values of the forces for all different keywords. Supply and Diffusion are reported for both Facebook (FB) and Twitter(TW)

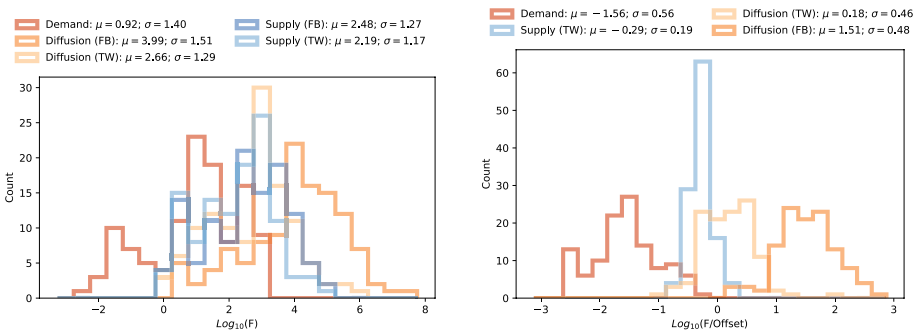

Fig. 3 **Left.** The histogram of the monthly values of the forces for all different keywords. **Right.** The histogram of the monthly values of the forces divided by the Supply (FB) for all different keywords. **Both.** Supply and Diffusion are reported for both Facebook (FB) and Twitter (TW)

Dynamics relations: stationarity and causality

In order to show that the three forces trace the same dynamics, we started measuring their stationarity. By stationarity of a force over a certain period of time T, we

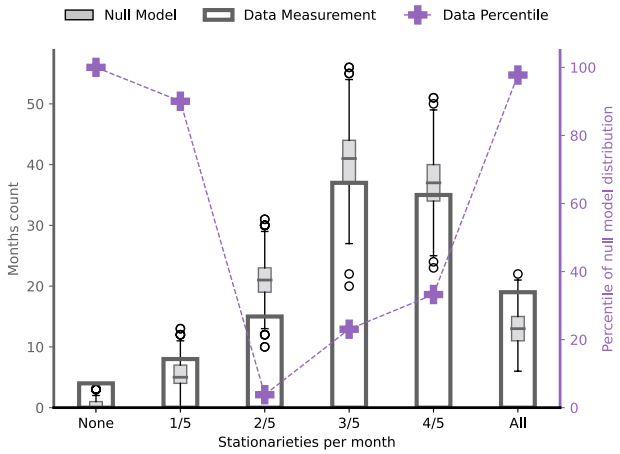

Fig. 4 The histogram of monthly co-occurring stationarieties of the forces on all months and all keywords, together with a null model obtained by 1000 reshuffling of the monthly sationarieties sequence (the boxplot in grey) and the representation of the actual measured values in terms of percentile (the purple crosses). The peaks at the extremes show that the five forces for a given keyword tend to have the same behaviour (stationarity or non-stationarity). If forces would have not been "synchronized", the purple line would simply have floated around the middle of the scale (corresponding to the null model)

mean that the unconditional joint probability distribution of the force values does not change when shifted in time (but still inside T). For example, a force stationary in a given month should have, in that month, the same average value, and the same variance. The basic idea is that if the forces are related to the same dynamics, they should be stationary at the same time or non-stationary at the same time. If the forces are not related to the same dynamics, stationarity (or non-stationarity) should not co-occur among different forces more often than in the random case. We performed the Augmented Dickey-Fuller (Hamilton 1994; MacKinnon 1994; Mushtaq 2011; Perktold et al. 2023) on the daily time-series of every force for every keyword and every month. Then, for every month and every keyword, we measured how many of the five forces were stationary. The results are reported in Fig. 4. The results must be compared to a null model to understand their significance. In fact, if, for example, all forces were stationary for all months and all keywords, stationarieties would obviously always co-occur, but that would not imply any special relation between the forces. The null model can simply be obtained by reshuffling the sequence of the stationarieties for each force. This will show how many co-occurrences are expected if there is no relation between the forces. We performed 1000 reshuffling and reported the result in Fig. 4, where we also show the percentile of the measurement on the actual data in the distributions obtained from the reshuffling.

We observe how the extremes (where none of the five forces or all of them) are much more frequent than in the null model, while the cases in the middle (with around half of the forces stationary) are much less frequent than in the null model. This shows the forces are stationary (or non-stationary) all together, so their months of dynamics seem to be synchronised, as their months of stasis. This implies, for example, that if the mean of a force is constant for a given keyword, the means of the other forces will also be constant. And when the mean of a force is moving, also the

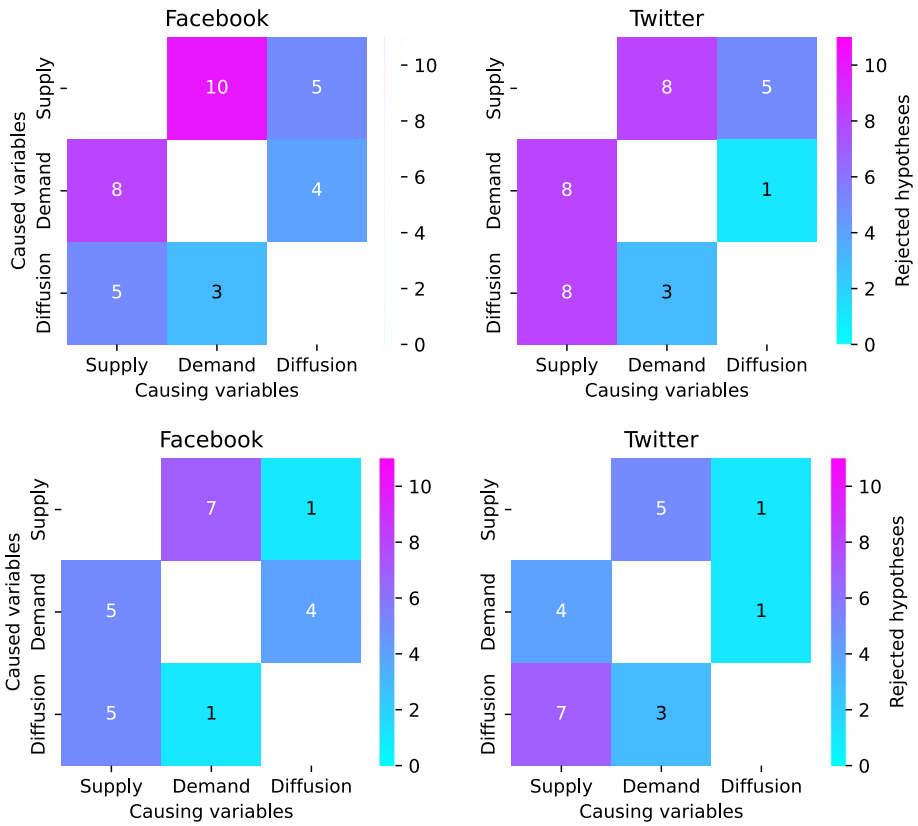

Fig. 5 Summary of causality tests for information dynamics in Facebook (left panels) and in Twitter (right panels). For the top tables, each cell shows the total number of tests (see Sect. 3) rejecting the hypothesis that the one variable (cell's column) is causing another (cell's row) conditioned on a third (off cell's axes). For the bottom tables, each cell indicates the total number of rejections from the same tests as above, but counting only the causal effect that is most significant, in the case both directions are rejected (e.g., if conditionally on Diffusion both null hypotheses "Demand do not causes Supply" and "Supply does not cause Demand" are rejected, only the one with smallest *p*-value contributes to the sum shown in the relative cell)

other will move. We can then argue the forces time-series are telling the story of the same phenomenon from different angles.

To deepen our analysis, we studied the dynamics more in detail focusing on how different forces can cause each other. This part of our study focuses again on the different forces time-series, and assesses the causality relation by measuring the information transfer. This allows to quantify how much information from one force influences the behaviour of another by statistically analysing the distributions of time-series value. An important advantage of this approach is that there is no need for a theoretical model of the dependence between the considered forces. For each keyword, we used a statistical hypothesis test for conditional independence between any two forces time-series conditioning on the third. We used a procedure based on resampling via smooth bootstrap. This allows to assess the significance of information transfer in all six possible directions of causality between the three forces. For this measure, we considered the two social networks separately, and we report the results in Fig. 5, aggregated by causality direction.

Only a minority of keywords (one for Facebook and four for Twitter) showed no signal. The most common direction of causality is from Demand to Supply on Facebook, while

the importance is more distributed on Twitter. Diffusion seems to be the less important in terms of causality, in particular on Twitter. Still, all combinations and all directions occur. Figure 5 also reports the comparison between opposite directions, confirming again the importance of the relation between Supply and Demand on Facebook, while on Twitter, the relation from Supply to Demand seems to be more common. This relation, present also on Facebook, is expected, to some extent, since Diffusion can happen, by construction, only when Supply is present. More generally, this analysis shows that the forces exchange information and can drive, in various ways, the system's dynamics. Now that we have shown that our definitions of the forces are actually tracking the dynamic of the information ecosystem, we speculate that the relative levels shown in Fig. 3 are the equilibrium level of the system. We can now use these forces to assess the status of the health of the system.

Semantics and the role of disinformation

So far we neglected the semantic aspect, aggregating the different keywords. We are now going to define the semantic vectors for each force. E.g. the Demand semantic vector for a given day has the values of the demand for the different keywords for that day as components. The "real" semantic vector would include all possible keywords, while in this work we include only the most important. So, even if the conclusions we draw cannot be considered to be valid for the whole system, we can claim that we are studying the most important part of the public debate. We are also going to analyse disinformation, which in this context we can define for Supply and Diffusion as the production of posts (and their shares) by a subset of sources annotated as "Non-Trustworthy" by professional fact-checkers (see Sect. 3) on both the analysed social media. So, now we can define, for each day, nine semantic vectors: one for Demand, four for Supply (for both social media, for all sources and for the Non-Trustworthy subset), and four for Diffusion. In SI we report a UMAP embedding of all nine vectors for each day. We synthesize a more agile representation by calculating the mutual similarities between the vectors for each day. We chose the Pearsons'correlation similarity between the semantic vectors of the considered couple of forces for a given day. Then, we took the median of the distances in the daily distribution for each couple of vectors. All the medians were significant, positive, and high (the minimum observed is 0.66, between Demand and Diffusion on all sources of Twitter). Then we arranged the node according to the Fruchterman-Reingold force-directed algorithm [41-43], and reported the result in Fig. 6.

In this representation, we observe two main features. First, he two sets of sources (all sources and Non-Trustworthy sources) cluster separately, pointing out that the difference of platform is less important than the type of source. E.g., Non-Trustworthy Supply on Facebook is closer to Non-Trustworthy Supply on Twitter than on Supply from all sources on Facebook. This suggests a coherence in the Non-Trustworthy news that transcends the boundaries of the platforms. Second, perhaps more importantly, the Non-Trustworthy cluster, particularly the Non-Trustworthy supplies, is closer to Demand than the other cluster. This is a potential danger to social discussion since it suggests that Non-Trustworthy news production and spreading are closer to the community's needs for information than general news. To conclude the analysis, we also probed the relation between the semantic similarities of general news and the share of Non-Trustworthy

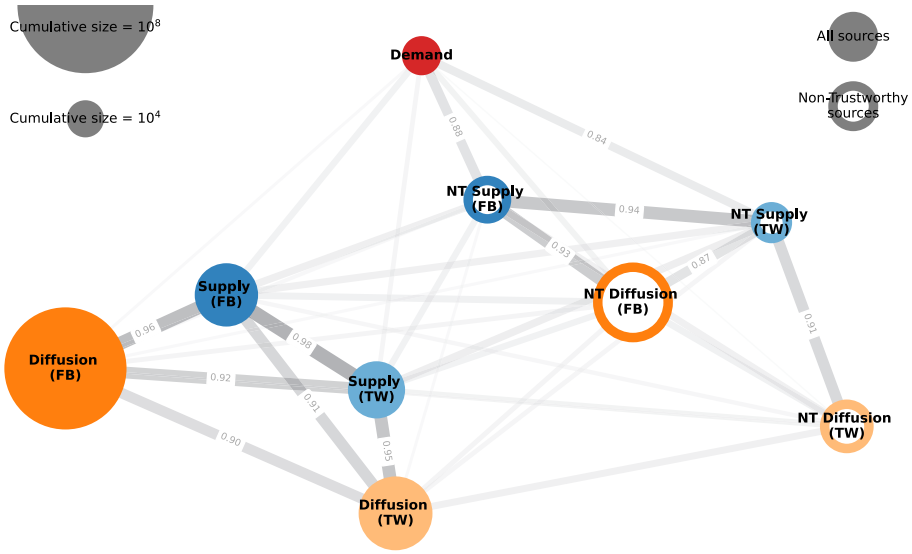

Fig. 6 The graph of the forces differentiated for the type of sources (all sources and Non-Trustworthy, for both social media) where the links are the median of the correlations between the daily semantic vectors of the forces. The darker and thicker the link, the higher the value of similarity. Labels are reported for the highest 33% values. The size of the nodes represents the order of magnitude of the cumulative value over the observed period. If two dots are closer, it means the corresponding forces for that source type are semantically closer. E.g. the closest nodes are the Supplies from all sources for FB and TW. The farthest nodes are the Diffusion from all source for FB and the Demand

Table 1 The Spearman correlation coefficients of general forces similarities versus the share of Non-Trustworthy supply and diffusion in the two social networks

Spearman correlations	**NT supply share**		**NT diffusion share**	
(* $\Rightarrow p - value > .05$**)**	**Facebook**	**Twitter**	**Facebook**	**Twitter**
$R_S(Demand, Supply)$	− 0.32	− 0.57	− 0.48	− 0.44
$R_S(Diffusion, Supply)$	0.01*	− 0.31	0.08*	− 0.22
$R_S(Demand, Diffusion)$	− 0.41	− 0.57	− 0.41	− 0.42

* indicates p-value is greater than 5%

news in the overall system. We measured Spearman's correlation coefficient between the forces similarities in the general news production and the fraction of Non-Trustworthy news Supply and Diffusion over the general news Supply and Diffusion. We report the results in Table 1.

All the significant coefficients are negative, meaning that the lower the similarities between the forces, the higher the volume of Non-Trustworthy news Supply and Diffusion. We observe that this effect is limited for the distance between Diffusion and Supply, i.e., it is present only on Twitter. This is unsurprising since we have already observed how Demand has a central role in the dynamics, especially on Facebook. More generally, despite the results being an unsettling signal that disinformation providers seem to take advantage of gaps between the forces in general news, they also provide an assessment strategy of the health status of the information market. In fact, the measurement of the gaps between the forces can show the vulnerability of the information market almost in

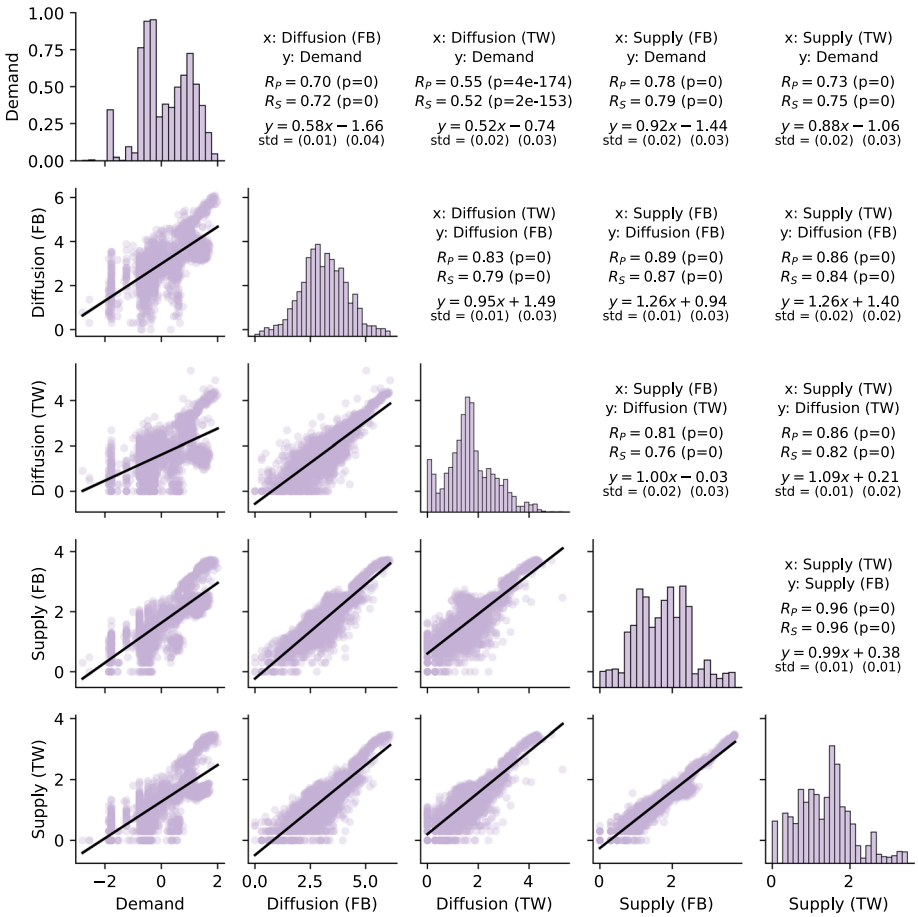

Fig. 7 Correlation (Pearson and Spearman) and linear regressions between the logarithms of the daily values of the forces for all different keywords. Supply and Diffusion are reported for both Facebook (FB) and Twitter(TW)

real-time (daily, at least) to the attempts of escalation in disinformation production and diffusion and help to prevent them. The importance of this result also comes from the fact that the measurement of the gaps between the forces does not depend on the measure of Non-Trustworthy sources' supply. A global monitoring of the forces in the general news ecosystem is sufficient. On the other hand, the current assessment of Non-Trustworthy sources depends on the manual annotation work of professional fact-checkers. While their work is crucial for the health of the news ecosystem, and for research, it makes makes real-time assessment unfeasible due to the limited resources and the overwhelming production of misinformation. Our results suggest a possible strategy that can overcome these limitations (Figs. 7, 8, 9, 10).

Materials and methods

Data collection and pre-elaboration

This study investigates the interplay between the three fundamental forces shaping the news market: Demand, Supply, and Diffusion. Namely, we capture Demand through the main terms used in the Google Search Engine from December 2019 to August 2020, as provided by the Google Trends tool. The tool does not provide an absolute scale but

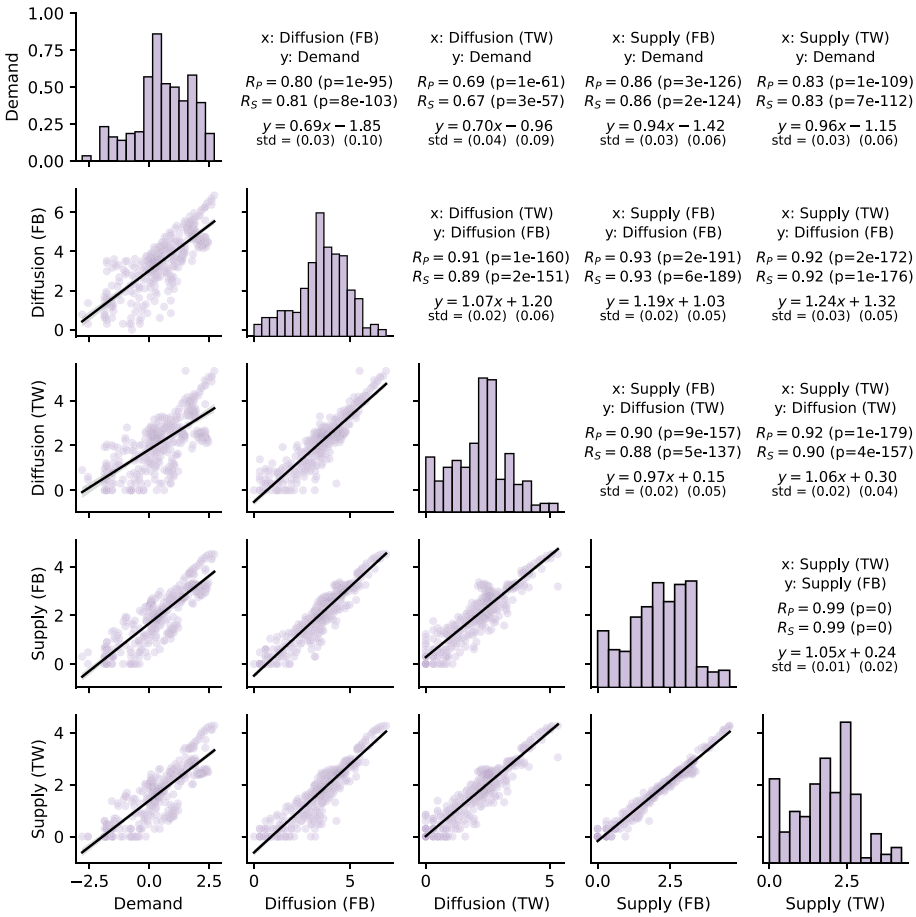

Fig. 8 Correlation (Pearson and Spearman) and linear regressions between the logarithms of the weekly values of the forces for all different keywords. Supply and Diffusion are reported for both Facebook (FB) and Twitter(TW)

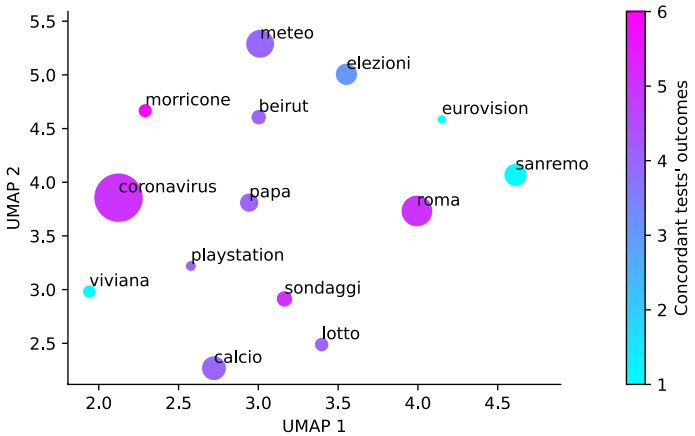

Fig. 9 Embedding of keywords' test significance. For every keywords, the vector of the *p*-values from the twelve tests for causal analysis, from Facebook and Twitter dynamics, is projects in two dimensions via UMAP embedding using the cosine distance. Each dot is coloured by the number of tests (out of six) that are rejected or accepted for both social media. Dots are sizes according the square root of their total supply

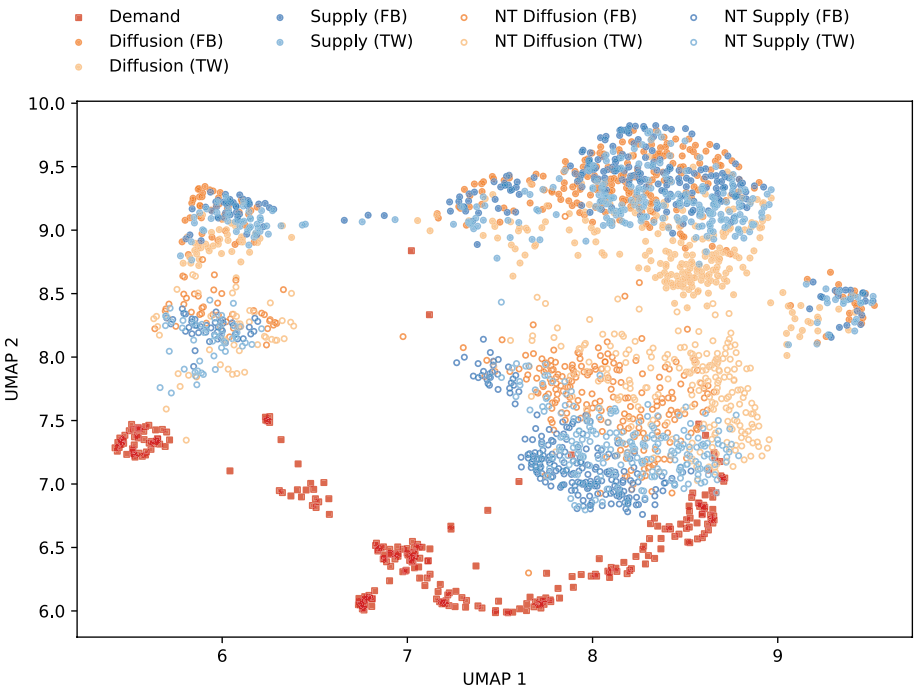

Fig. 10 2-D UMAP embedding using the correlation distance of the forces daily semantic vectors, differentiated for the type of sources: all sources and Non-Trustworthy (NT)

allows to probe multiple terms simultaneously, returning the results in a scale that can be used for comparison. In other words, the Demand metric is missing an unknown factor, but this is irrelevant since the comparisons and considerations we draw in the paper do not rely on knowing the absolute scale. The keywords gathered include:

- *Beirut, Campania, Italia, Lombardia, Milano, Piemonte, Roma, Sicilia, Veneto* (terms related to Geographic locations);
- *bollettino, casi, contagi, coronavirus, dati, decreto, mappa, morti, sintomi* (terms related to the Covid-19 outbreak);
- *calcio, campionato, champions, serie A* (terms related to soccer game);
- *Gioele, Viviana, Viviana Parisi* (terms related to a famous crime news incident in Italy);
- *Eurovision, Morricone, Sanremo* (terms related to music);
- *Papa, Papa Francesco* (Pope Francis);
- *playstation, ps5* (Sony gaming console);
- *elezioni, regionali, sondaggi* (terms related to political elections);
- *lotto, meteo, news* (other general terms).

Many of these terms are semantically overlapping, so we select a shorter set of keywords to account for the most searched topics. We also remove the Italian locations (except *Rome*) and the term *news* because they are too generic and not related to a specific topic. The selected list is the following: *beirut*, *calcio*, *coronavirus*, *elezioni*, *Eurovision*, *lotto*, *meteo*, *Morricone*, *Papa*, *playstation*, *Roma*, *Sanremo*, *sondaggi*, *Viviana*. To analyse the

Table 2 Breakdown of the Supply and Diffusion dataset. Data are divided by source type (Reliable or Non-Trustworthy) and social media platform (Facebook or Twitter)

	Sources	Facebook		Twitter	
		Posts	Shares	Tweets	Retweets
Reliable	330	1, 956, 941	121, 108, 643	1, 302, 008	6, 291, 579
Non-Trustworthy	81	155, 737	11, 225, 099	108, 703	2, 102, 646

Table 3 Correlation (Pearson and Spearman) between the daily direct supply of news from the selected information leaders and the corresponding social media production on Facebook and Twitter, respectively

	Facebook		Twitter	
	Corr	*p*-value	Corr	*p*-value
Pearson R_p	0.791	~ 0	0.861	~ 0
Spearman R_s	0.783	~ 0	0.872	~ 0

news Supply, we rely on a list of news outlets provided by AGCOM, the Italian Authority for Communications Guarantees, which covers the main leaders of information in Italy during the time span under analysis (AGCOM 2018). The list includes traditional newspapers, online-only news outlets, information agencies, TV, radio websites, and scientific sources. Moreover, the data have specific annotations on Non-Trustworthy sources. The source-based methodology is widely recognized and firmly established in the existing literature on disinformation (Grinberg et al. 2019). We adopt the same method, which is especially suitable for examining the conduct of Non-Trustworthy sources, as in the current study. Note that the resulting leader dataset is the same used in Gravino et al. (2022) to which the reader can refer for further details.

Limited to contents containing the selected keywords, we use news posted by the selected news outlets on the two major social media platforms in Italy - Facebook and Twitter - as a proxy for the Supply. We trace the Diffusion of these contents through the corresponding user engagement, represented by the number of shares a post gained on the belonging platform. For gathering data from Facebook, we rely on CrowdTangle (CrowdTangle Team 2023), a Facebook-owned tool that tracks interactions on public content from various social media platforms. For Twitter, we exploit the official API accessed through the academic account before the limitations introduced by the new management.[1]

The final supply dataset consists of 2, 112, 678 Facebook posts and 1, 410, 711 tweets from 411 different news sources, as more clearly detailed in Table 2, which in turn also includes the corresponding Diffusion statistics.

In addition to being functional for a swift identification of Diffusion statistics, using social media production as a proxy for Supply by selected information leaders is also legitimate, given the high and significant correlation it has with the general direct production of news from the same news sources (Gravino et al. 2022), as reported in Table 3.

[1] https://twitter.com/XDevelopers/status/1621026986784337922.

In our work, we performed different kinds of aggregation on different series. The Demand is already provided by the Google Trends platform as a daily time-series. Supply and Diffusion have also been aggregated daily to have a similar format. Then, for the different analyses, different levels of aggregation have been adopted: daily, monthly or over the whole period. The level of aggregation used for every analysis is described in the main text.

Causal analysis

To investigate causal interactions between time-series of Supply, Demand and Diffusion of a given keyword in a given social media (Figs. 5 and 9), we used a statistical hypothesis test for conditional independence between time-series based on resampling via smooth bootstrap (Efron and Tibshirani 1994; Prevedello and Monechi 2024). This testing procedure is a nonparametric counterpart of the Granger causality test that relaxes the requirements imposed by vector autoregression modelling and improves upon other nonparametric techniques based on local permutation resampling (Runge 2018).

Briefly, to test for the null hypothesis that two signals X and Y are independent conditioned on the signal Z, the Transfer Entropy statistic is used Schreiber (2000). To approximate its distribution, the statistic is calculated over data resampled via smooth bootstrap: first, the joint distribution of $X_t, \ldots, X_{t-m}$, $Y_t, \ldots, Y_{t-m}$, $Z_t, \ldots, Z_{t-m}$ is estimated by Kernel Density Estimation (KDE) with Gaussian kernels from the observations $D = (x_t, \ldots, x_{t-m}, y_t, \ldots, y_{t-m}, z_t, \ldots, z_{t-m})_{t \geq m}$, with lag $m = 7$ days, using Scott's bandwidth, and imposing that the covariance between X and Y given Z is null; from this distribution, the dataset D is sampled $B = 1000$ times, each time drawing from the KDE as many samples as observations, thus generating $D_1^*, \ldots, D_B^*$; finally, the statistic S is calculated on every bootstrapped dataset to determine the p-value $p = B^{-1} \sum_{i=1}^{B} \mathbf{1}(\{S(D_i^*) >= S(D)\})$, where $\mathbf{1}(A) = 1$ if A is true and is null otherwise.

For each keyword and social media, six p-values were then obtained, one for every permutation of (Supply, Demand, Diffusion) assigned to (X, Y, Z). Each set of six p-values was adjusted for multiple testing comparisons using the Holm-Bonferroni method.

Conclusions

The digital age provides new challenges as information travels more quickly in a system of increasing complexity. But it also provides new opportunities, as we can more easily track and study digital trails of the system. These trails have often been studied separately focusing on different aspects (like disinformation production or sharing dynamics) individually. In this work, we propose to study the news ecosystem as an information market by analysing three main metrics: Supply, Demand, and Diffusion of information. Working on a dataset relative to Italy from December 2019 to August 2020, we validate the choice of the metrics, proving their static and dynamic relations. We demonstrate that they seem to have specific equilibrium relative levels. We reveal the strategic role of Demand in leading a non-trivial network of causal relations. We show how disinformation news Supply and Diffusion seem to cluster by transcending social media platforms. It also appears to be closer to information Demand than the general news Supply and Diffusion, implying a potential danger to the health of the public debate. Finally, we prove that the share of disinformation in the Supply and Diffusion of news has a

significant linear relation with the gap between Demand and Supply/Diffusion of news from all sources.

This work confirms and expands the result of a previous work Gravino et al. (2022), pointing out the potential of the analysis of the whole news ecosystem through its main drivers. The results proved to be valid for different keywords and on different social media platforms. This is another step toward a potential real-time analysis of the information market to assess and possibly prevent vulnerabilities. Still, the work presents limitations, and there is much more to do. The analyses need to be expanded to different countries and different timeframes. Also, a more comprehensive strategy for keyword selection could be defined, e.g. leveraging topic detection algorithms. On the theoretical side, the challenge is on one side to define and test a model to describe the complex behaviour of the system, replicating all the crucial aspects. Such a model should be able to reproduce, for example, what happens when the equilibrium relative levels are modified by external perturbations. On the other side, besides the disinformation fraction, a wider set of metrics (e.g. polarisation) can be taken into account to assess the status of the system's health. The comparative analysis of the model metrics and the semantic relations could then be crossed with these metrics to be able to perform a more detailed real-time evaluation of the news ecosystem vulnerabilities.

Acknowledgements
We thank M. Delmastro of AGCOM for providing access to the database of Italian news outlets. The database was shared in the framework of the Task Force on 'Digital Platforms and Big Data - Covid-19 Emergency', established by AGCOM to contribute, among other things, to the fight against online disinformation on issues related to the COVID-19 crisis.

Author contributions
Conceptualization: Gravino. Methodology: Gravino, Prevedello. Validation: Gravino, Prevedello. Software: Gravino, Prevedello. Writing-original draft: Gravino, Prevedello, Brugnoli. Visualization: Gravino, Prevedello. Supervision: Gravino. Proofread: Gravino, Prevedello, Brugnoli. All authors read and approved the final manuscript.

Funding
This work has been supported by the Horizon Europe VALAWAI project (grant agreement number 101070930).

Availability of data and materials
Google Search engine data were generated by the Google Trends platform and is publicly available at https://trends.google.com. Derived data for Supply, Demand and Diffusion supporting the findings of this study are available at https://github.com/SonyCSLParis/news_searches.

Code availability
All codes for data analysis are available at https://github.com/SonyCSLParis/news_searches.

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Conflict of interests
Not applicable

Received: 15 January 2024 Accepted: 27 June 2024
Published online: 19 July 2024

References

Acampa S, Crescentini N, Padricelli GM (2022) Is it still disintermediated? the role of the influencer newsmaker in the social platform era. Cult Stud Soc 7:10–30. https://doi.org/10.30958/ajss.10-1-2
AGCOM (2018) News vs. fake in the information system. Technical report, AGCOM

Andrews KT, Caren N (2010) Making the news: movement organizations, media attention, and the public agenda. Am Sociol Rev 75(6):841–866. https://doi.org/10.1177/0003122410386689
Bastick Z (2021) Would you notice if fake news changed your behavior? an experiment on the unconscious effects of disinformation. Comput Hum Behav 116:106633. https://doi.org/10.1016/j.chb.2020.106633
Brugnoli E, Cinelli M, Quattrociocchi W, Scala A (2019) Recursive patterns in online echo chambers. Sci Rep 9(1):20118. https://doi.org/10.1038/s41598-019-56191-7
Brugnoli E, Galletti M, Lo Sardo R, Prevedello G, Di Canio M, Gravino P (2023) Decoding political social media posts. Nat Italy. https://doi.org/10.1038/d43978-023-00026-7
Cinelli M, Quattrociocchi W, Galeazzi A, Valensise CM, Brugnoli E, Schmidt AL, Zola P, Zollo F, Scala A (2020) The covid-19 social media infodemic. Sci Rep 10(1):16598. https://doi.org/10.1038/s41598-020-73510-5
Cinelli M, Brugnoli E, Schmidt AL, Zollo F, Quattrociocchi W, Scala A (2020) Selective exposure shapes the facebook news diet. PLoS ONE 15(3):1–17. https://doi.org/10.1371/journal.pone.0229129
CrowdTangle Team (2023) CrowdTangle. Facebook, Menlo Park, California, United States
Del Vicario M, Bessi A, Zollo F, Petroni F, Scala A, Caldarelli G, Stanley HE, Quattrociocchi W (2016) The spreading of misinformation online. Proc Natl Acad Sci 113(3):554–559. https://doi.org/10.1073/pnas.1517441113
Del Vicario M, Quattrociocchi W, Scala A, Zollo F (2019) Polarization and fake news: early warning of potential misinformation targets. ACM Trans Web 13(2):52. https://doi.org/10.1145/3316809
De Marzo G, Gravino P, Loreto V (2023) Recommender systems may enhance the discovery of novelties. arXiv. https://doi.org/10.48550/ARXIV.2312.08824
Desiderio A, Mancini A, Cimini G, Di Clemente R (2023) Recurring patterns in online social media interactions during highly engaging events. arXiv. arXiv:2306.14735 [physics]. https://doi.org/10.48550/arXiv.2306.14735
Efron B, Tibshirani RJ (1994) An introduction to the bootstrap. CRC Press, Boca Raton
Fruchterman TMJ, Reingold EM (1991) Graph drawing by force-directed placement. Softw Pract Exp 21(11):1129–1164. https://doi.org/10.1002/spe.4380211102
Garrett RK, Bond RM (2021) Conservatives' susceptibility to political misperceptions. Sci Adv 7(23):1234. https://doi.org/10.1126/sciadv.abf1234
Gravino P, Prevedello G, Galletti M, Loreto V (2022) The supply and demand of news during covid-19 and assessment of questionable sources production. Nat Hum Behav 6(8):1069–1078. https://doi.org/10.1038/s41562-022-01353-3
Grinberg N, Joseph K, Friedland L, Swire-Thompson B, Lazer D (2019) Fake news on twitter during the 2016 US presidential election. Science 363(6425):374–378. https://doi.org/10.1126/science.aau2706
Guay B, Berinsky AJ, Pennycook G, Rand D (2023) How to think about whether misinformation interventions work. Nat Hum Behav 7(8):1231–1233. https://doi.org/10.1038/s41562-023-01667-w
Hagberg AA, Schult DA, Swart PJ (2008) Exploring network structure, dynamics, and function using networkx. In Varoquaux G, Vaught T, Millman J (eds.) Proceedings of the 7th Python in Science Conference, Pasadena, CA USA, pp 11–15
Hagberg AA, Schult DA, Swart PJ. spring_layout - NetworkX 3.2.1 documentation. https://networkx.org/documentation/stable/reference/generated/networkx.drawing.layout.spring_layout.html Accessed 2024-01-14
Hamilton JD (1994) Time Series Analysis. Princeton University Press, Princeton, New Jersey, United States. https://doi.org/10.1515/9780691218632
Hohenberg B (2023) Truth and bias, left and right: testing ideological asymmetries with a realistic news supply. Public Opin Q 87(2):267–292. https://doi.org/10.1093/poq/nfad013
Iyengar S, Norpoth H, Hahn KS (2004) Consumer demand for election news: the horserace sells. J Pol 66(1):157–175. https://doi.org/10.1046/j.1468-2508.2004.00146.x
Kapantai E, Christopoulou A, Berberidis C, Peristeras V (2021) A systematic literature review on disinformation: toward a unified taxonomical framework. New Media Soc 23(5):1301–1326. https://doi.org/10.1177/1461444820959296
King E (1998) Redefining relationships: interactivity between news producers and consumers. Convergence 4(4):26–32. https://doi.org/10.1177/135485659800400404
Lazer DMJ, Baum MA, Benkler Y, Berinsky AJ, Greenhill KM, Menczer F, Metzger MJ, Nyhan B, Pennycook G, Rothschild D, Schudson M, Sloman SA, Sunstein CR, Thorson EA, Watts DJ, Zittrain JL (2018) The science of fake news. Science 359(6380):1094–1096. https://doi.org/10.1126/science.aao2998
MacKinnon JG (1994) Approximate asymptotic distribution functions for unit-root and cointegration tests. J Bus Econ Stat 12(2):167–176
Mancini A, Desiderio A, Di Clemente R, Cimini G (2022) Self-induced consensus of Reddit users to characterise the GameStop short squeeze. Sci Rep 12(1):13780
Mattei M, Pratelli M, Caldarelli G, Petrocchi M, Saracco F (2022) Bow-tie structures of twitter discursive communities. Sci Rep 12(1):12944. https://doi.org/10.1038/s41598-022-16603-7
Morgan S (2018) Fake news, disinformation, manipulation and online tactics to undermine democracy. J Cyber Policy 3(1):39–43. https://doi.org/10.1080/23738871.2018.1462395
Mushtaq R (2011) Augmented dickey fuller test. SSRN Electron J. https://doi.org/10.2139/ssrn.1911068
Pacheco D, Flammini A, Menczer F (2020) Unveiling coordinated groups behind white helmets disinformation. In Companion Proceedings of the Web Conference 2020. WWW '20, pp 611–616. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3366424.3385775
Patuelli A, Saracco F (2023) Sustainable development goals as unifying narratives in large UK firms' Twitter discussions. Sci Rep 13(1):7017. https://doi.org/10.1038/s41598-023-34024-y
Pennycook G, Rand DG (2019) Lazy, not biased: susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. Cognition 188:39–50. https://doi.org/10.1016/j.cognition.2018.06.011
Perktold J, Seabold S, Sheppard K, Chad F, Shedden K, Ibrockmendel, j-grana6, Quackenbush Peter, Arel-Bundock Vincent, McKinney Wes, Langmore Ian, Baker Bart, Gommers Ralf, yogabonito, s-scherrer, Zhurko Yauhen, Brett Matthew, Giampieri Enrico, yl565, Millman Jarrod, Hobson Paul, Vincent, Roy Pamphile, Augspurger Tom, tvanzyl, alexbrc, Hartley Tyler, Perez Fernando, Tamiya Yuji, Yaroslav Halchenko (2023) statsmodels/statsmodels: Release 0.14.1. Zenodo. https://doi.org/10.5281/ZENODO.593847

Pratelli M, Saracco F, Petrocchi M (2023) Entropy-based detection of Twitter echo chambers. arXiv. arXiv:2308.01750 [physics]. https://doi.org/10.48550/arXiv.2308.01750
Prevedello G, Monechi B (2024) Estimating causal effects by conditional independent test via smooth bootstrapping. arXiv, arXiv: https://arxiv.org/abs/2312.08824
Rehman AU, Jiang A, Rehman A, Paul A, Din S, Sadiq MT (2023) Identification and role of opinion leaders in information diffusion for online discussion network. J Amb Intell Hum Comput. https://doi.org/10.1007/s12652-019-01623-5
Rocha YM, Moura GA, Desidério GA, Oliveira CH, Lourenço FD, Figueiredo Nicolete LD (2023) The impact of fake news on social media and its influence on health during the covid-19 pandemic: a systematic review. J Public Health 31(7):1007–1016. https://doi.org/10.1007/s10389-021-01658-z
Runge J (2018) Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In International Conference on Artificial Intelligence and Statistics, pp 938–947. PMLR
Sasahara K, Chen W, Peng H, Ciampaglia GL, Flammini A, Menczer F (2021) Social influence and unfollowing accelerate the emergence of echo chambers. J Comput Soc Sci 4(1):381–402. https://doi.org/10.1007/s42001-020-00084-7
Schreiber T (2000) Measuring information transfer. Phys Rev Lett 85(2):461
Simon FM, Camargo CQ (2023) Autopsy of a metaphor: the origins, use and blind spots of the 'infodemic'. New Media Soc 25(8):2219–2240. https://doi.org/10.1177/14614448211031908
Thompson N, Wang X, Daya P (2020) Determinants of news sharing behavior on social media. J Comput Inf Syst 60(6):593–601. https://doi.org/10.1080/08874417.2019.1566803
Thurman N, Moeller J, Helberger N, Trilling D (2019) My friends, editors, algorithms, and i. Dig J 7(4):447–469. https://doi.org/10.1080/21670811.2018.1493936
Tommasel A, Menczer F (2022) Do recommender systems make social media more susceptible to misinformation spreaders? In: Proceedings of the 16th ACM Conference on Recommender Systems. RecSys '22, pp. 550–555. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3523227.3551473
Treen KM, Williams HTP, O'Neill SJ (2020) Online misinformation about climate change. WIREs Clim Change 11(5):665. https://doi.org/10.1002/wcc.665
Turcotte J, York C, Irving J, Scholl RM, Pingree RJ (2015) News recommendations from social media opinion leaders: effects on media trust and information seeking. J Comput-Mediat Commun 20(5):520–535. https://doi.org/10.1111/jcc4.12127
Welbers K, Opgenhaffen M (2018) Social media gatekeeping: an analysis of the gatekeeping influence of newspapers' public facebook pages. New Media Soc 20(12):4728–4747. https://doi.org/10.1177/1461444818784302

Publisher's Note

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