Social Science & Medicine 285 (2021) 114215

Contents lists available at ScienceDirect

Social Science & Medicine

journal homepage: www.elsevier.com/locate/socscimed

Emergence of knowledge communities and information centralization during the COVID-19 pandemic

Pier Luigi Sacco [a], Riccardo Gallotti [b], Federico Pilati [a], Nicola Castaldo [b], Manlio De Domenico [b,*]

[a] IULM, Via Carlo Bo 1, 20143, Milan, Italy
[b] Comune Lab, Fondazione Bruno Kessler, Via Sommarive 18, 38123, Povo, Italy

ARTICLE INFO

Keywords:
Global Communication
Infodemic
Public Health

ABSTRACT

Background: As COVID-19 spreads worldwide, an infodemic – i.e., an over-abundance of information, reliable or not – spreads across the physical and the digital worlds, triggering behavioral responses which cause public health concern.

Methods: We study 200 million interactions captured from Twitter during the early stage of the pandemic, from January to April 2020, to understand its socio-informational structure on a global scale.

Findings: The COVID-19 global communication network is characterized by knowledge groups, hierarchically organized in sub-groups with well-defined geo-political and ideological characteristics. Communication is mostly segregated within groups and driven by a small number of subjects: 0.1% of users account for up to 45% and 10% of activities and news shared, respectively, centralizing the information flow.

Interpretation: Contradicting the idea that digital social media favor active participation and co-creation of online content, our results imply that public health policy strategies to counter the effects of the infodemic must not only focus on information content, but also on the social articulation of its diffusion mechanisms, as a given community tends to be relatively impermeable to news generated by non-aligned sources.

1. Introduction

One of the most challenging aspects of the current COVID-19 pandemic crisis (World Health Organizationet al., 2020) is its infodemic dimension, that is, an overabundance of COVID-related information that makes it difficult for the majority of the public opinion to distinguish between reliable and unreliable sources (Tangcharoensathien et al., 2020; Zarocostas, 2020). Indeed, although the information disorder (Wardle and Derakhshan, 2017) related to COVID-19 has often been described as pertaining only to the production and circulation of fake news (Orso et al., 2020), the infodemic phenomenon presents two distinct but closely related dimensions: on the one hand, an overabundance of information makes it difficult for users to find the right answers to their questions, and may lead them to settle for the first inputs at hand, being thus guided by selection bias (Prior, 2005); on the other hand, a large volume of misleading content generates a pollution problem in the online environment in which they look for information (Cinelli et al., 2020). The combination of the two effects becomes particularly threatening as users must, at the same time, filter a very large amount of content and discern its informational quality in a highly noisy, emotionally charged setting, while relying upon limited personal competence on the matter; furthermore, in a situation where mistakes can be very costly in terms of health consequences. This difficulty in distinguishing between reliable and unreliable sources causes the spreading of confusion and anxiety in the population and favors the emergence of incorrect and socially detrimental beliefs, as well as of misinterpretation or refusal of prescribed behavioral guidelines (Romer and Jamieson, 2020), to the extent of threatening the efficacy of public health measures (Editorial, 2020).

The scale and complexity of this phenomenon is clearly exacerbated by the recent transition from traditional media into an increasingly globalized and digital knowledge ecosystem (Holliman, 2011; Chadwick, 2017). The internet functions as an online commons (Milberry and Anderson, 2009) where news consumption becomes spatially and temporally ubiquitous (Struckmann and Karnowski, 2016). Here, new forms of opinion leadership filter, contextualize and interpret available

* Corresponding author.
E-mail address: mdedomenico@fbk.eu (M. De Domenico).

https://doi.org/10.1016/j.socscimed.2021.114215
Received 14 November 2020; Received in revised form 29 March 2021; Accepted 6 July 2021
Available online 31 July 2021

information at a variety of social scales, from national audiences to personal friendship networks (Bergstrom and Jervelycke Belfrage, 2018). Individuals make an active use of social media to access news as part of their personal strategy of social capital acquisition, civic engagement and political participation (Gil de Zu'ñiga et al., 2012). Therefore, the nature of the information that is circulated and validated within an individual's digital relational sphere has a considerable impact on her/his orientations and conduct in a variety of highly sensitive matters, including public health. For instance, an increasing number of rapidly spreading conspiracy theories have been recorded throughout the pandemic (Enders et al., 2020). As previously observed in vaccination-related information campaigns, such digital niches, in which false medical information freely circulated and was largely acknowledged as 'evidence', had a significant impact on the behavior of users who subscribed to such conspiracy theories (Jolley and Douglas, 2014). Conspiracy mindsets (Sutton and Douglas, 2020) tend to be associated with extreme political positions, both right-wing and left-wing, although comparatively more strongly at the right extreme of the spectrum, and tend to be closely related to lack of compliance with socially responsible prevention behaviors, turning the social interaction networks of conspiracy theory endorsers into dangerous channels of contagion transmission for the individuals involved, and for society as a whole (Douglas, 2021). A generalized, unrestricted access to processes of information collection, curation and dissemination may be welcomed in principle as a basis for an inclusive, digital citizenship (Choi et al., 2017) founded on participatory readiness (Allen, 2016). However, socially beneficial digital participation requires a rich and diverse set of capabilities (Shelley et al., 2004), enabling for instance individuals to refer to collective intelligence to discern the quality and reliability of news sources (Pennycook and Rand, 2019), to learn to play by the rules of public deliberation and debate (Parkins and Mitchell, 2005), and to constructively engage in discussions with people with different opinions and ideological orientations (Glück, 2019). In this regard, the COVID-19 infodemic can be seen as a very significant acid test of where we stand in the transition toward a mature form of digital democracy and knowledge society. The evidence that can be gathered from the natural living lab of the COVID-19 infodemic makes for a sobering call. The production and large-scale diffusion of misleading and fake news has been massive since the early stages of the pandemic crisis, and escalating ever since (O'Connor and Murphy, 2020; Orso et al., 2020), so that infodemics are now considered a major threat to public health (Zarocostas, 2020).

To understand communication phenomena on a global scale, such as the infodemic related to COVID-19, the best approach is to analyze their digital traces produced through social media (Lazer et al., 2009). Studies conducted in the past that have focused on Twitter as a global communication platform have improved our understanding of the patterns of information dissemination on the world wide web (Bakshy et al., Adamic), and explained how groups of users are segmented into echo chambers (Colleoni et al., 2014). However, the unique features of the COVID-19 pandemic imply that the analysis of the COVID-related infodemic phenomenon provides us with an unprecedented opportunity to directly observe the action of the micro-processes that shape the network structures through which the digital dissemination of information takes place at a global scale in the middle of the most serious social crisis in decades.

In this paper, we develop a computational approach to gain insight on a key social feature of infodemics, that is, the structure of the knowledge communities that are endogenously formed in the process of creation, filtering and dissemination of COVID-19-related information. We map the global communication ecosystem from more than 200 million interactions on Twitter, and show that the COVID-19 infodemic presents a highly characteristic community structure, shaped by ideological orientation, typology of fake news, and geographical areas of reference, that reflects complex geo-political patterns, and presents very specific features that are not found in the previous literature on digital communities. Compared to previous studies that stressed an adherence of Twitter's digital communities to mere geographic location or membership in fandoms and groups with common (e.g. professional) interests, the emerging knowledge communities reflect the geopolitical structure of cultural exchanges and influences that distinguish the current world order (Leetaru et al., 2013). For example, it is possible to note how in Africa or India several top influencers are British mainstream media, despite the fact that these are not the main information outlets for the audiences of such geographical areas. Another peculiarity of our results concerns the high level of politicization of the public debate around COVID-19. As a rule, health debates should not be mostly driven by ideological positions (Barber'a et al., 2015), but we find that in the case of COVID-19, in line with what is emerging from other studies (Jiang et al., 2020; de Bruin et al., 2020), political polarization makes a clear difference as to pandemic-related beliefs and attitudes. Finally, we find that the conversation within each community is basically shaped by an unexpectedly small number of actors who generate most of the content and are responsible for most of the circulated news sources. Moreover, the identified community structure suggests that online discussion is driven by a social hierarchy of influencers (Goel et al., 2016). By analyzing how online Twitter users spontaneously cluster into strongly characterized knowledge communities, whose conversations reflect a varying incidence of mainstream and verified news sources versus unreliable and fake news ones, we find that the incidence of fake news in some of the community conversations is so high that misinformation becomes a key epistemic community trait. By considering the interactions between the various knowledge communities, we are able to map what we could call the "cosmic web of COVID-19 infodemics", where each community can be regarded as a "galaxy" being part of a "local group" of related communities, and where each community is in turn modularly structured into sub- and sub-sub-communities, etc. Our analogy is not meant as a scientific statement of structural isomorphism, but rather as a cue for a quick intuitive understanding that builds upon the curious similarity between the community organization we find, and the way extragalactic matter in the outer Universe organizes to build galaxies and superclusters, i.e., groups of galaxy clusters, separated by vast spaces of empty regions, i.e., voids.

2. Methods

Mapping of an online global communication network. We have collected about 200 million social interactions related to the spread of COVID-19 on Twitter (Gallotti et al., 2020), between 22 January and April 16, 2020 (see Materials and Methods for details about data collection and completeness), covering different stages of the COVID-19 pandemic, including sub-national or national pre-lockdown and lockdown periods of at least 77 countries worldwide. The aim of the present work is to identify the network structure that drove global communication related to COVID-19 during the early stage of the pandemic. Indeed, it is precisely at the beginning of unexpected events that the emergence of online communities defines hierarchies in the flow of information also for times to follow (Goel et al., 2016). We focused on Twitter because of its relevance in the hybrid media system. Indeed, compared to other social media, Twitter has acquired a central role in the distribution and research of information regarding breaking news and journalistic investigations (Kwak et al., Moon). Another reason is strictly methodological: Twitter provides access to publicly available messages upon specific requests through their application programming interface (API). Therefore, we identified a set of hashtags and keywords that gained collective attention since the first recorded cases of COVID-19: #coronavirus, #ncov, #Wuhan, #covid 19, #covid-19, #sarscov2 and #covid. This set includes the official names of the virus and of the disease, including the early tentative ones, as well as the name of the city where the first cases of COVID-19 were recorded. More details about methodological choices, such as the use of Twitter as the sole data source or the selection of terms, are available in Gallotti et al. (2020). Pairwise interactions (retweets, replies and mentions) are used to build an undirected weighted communication network where nodes are the user's accounts and weighted links represent the number of active (e.g., who endorses whom, who replies to whom) or passive (e.g., who is endorsed by whom, who receives a reply by whom) interactions between two users in the time frame considered (De Domenico et al., 2013). We kept links only if there were at least 10 social interactions, in order to guarantee that only highly interacting accounts are considered, whereas sporadic interactions are neglected. Our methodology implicitly focuses on the points of concentration of the information flow and the choice to cut at accounts with at least 10 interactions has allowed us to keep the computational burden manageable by pruning the most marginal nodes whose influence of community structure is relatively minor. The actual cutoff value has been chosen on the basis of trial experiments. We focused on the largest connected component – i.e., the largest cluster connecting accounts through their social dynamics – allowing us to build a complex network representation of our data set consisting of about 0.4 million users connected by 1.1 million social activities. We unravel the mesoscale organization of this communication network by means of the Louvain method (Blondel et al., 2008), one of the most adopted algorithms for detecting groups in massive complex networks. The result shows that the network is highly modular, with an estimated modularity of 0.82, highlighting the existence of a well-defined group structure consisting of 44 communities of significant size; far from being compatible with the hypothesis of random interactions. We show in Fig. 1 the map of the resulting online communication network (please see the Supplementary Materials for a High-Resolution Version of Fig. 1). Fig. 2 shows, using circular visualization (Abel and Sander, 2014), the network of groups where individual accounts within the same community are merged together to form a supernode and the overall number of interactions across groups encodes inter-community communication exchange.

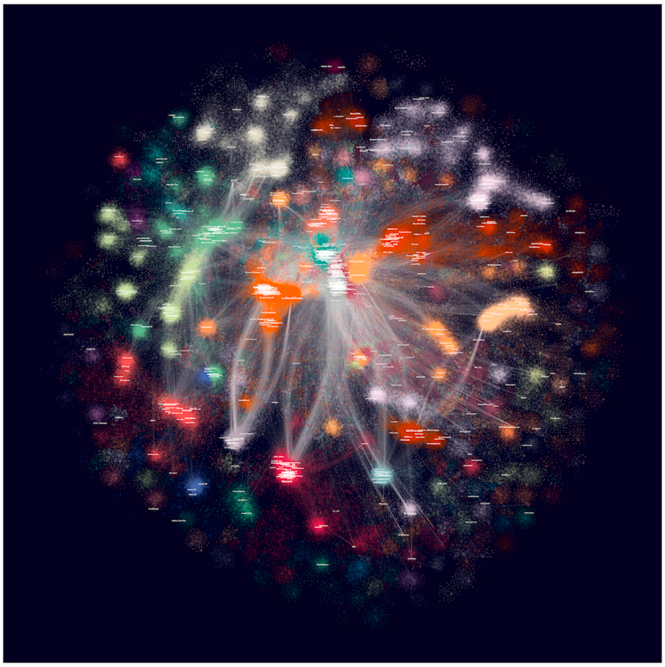

Fig. 1. Map of the online global COVID-19 communication network on Twitter. We display a web of about 1.1 million social interactions about COVID-19 observed worldwide between 22 January and April 16, 2020. Nodes represent about 0.4 million user accounts and links encode their social interactions aggregated over the observational period. Nodes are colored according to their social group inferred using the Louvain method. Only nodes belonging to groups which are at least 0.1% of system size are colored, i.e., the smallest colored group consists of about 400 accounts. A label with the account name is shown for extremely active users, the one with an overall social activity of at least 3500 interactions (either active or passive). The network exhibits a highly heterogeneous, modular and hierarchical organization, confirmed by our quantitative measurements. See text for details.

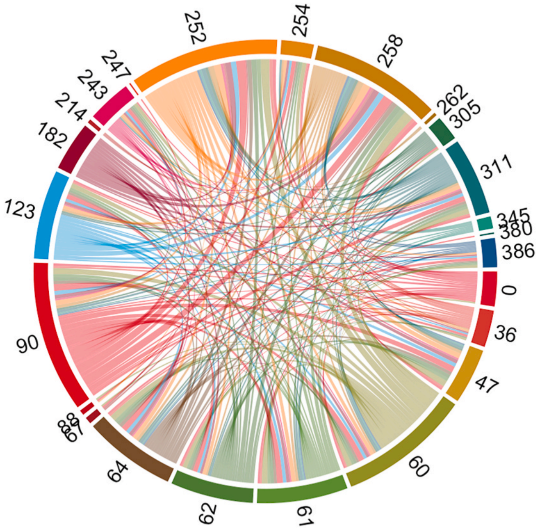

Fig. 2. Mesoscale organization of the COVID-19 communication network. Accounts shown in Fig. 1 are clustered together into a supernode, denoting their group, and interactions across groups are aggregated to build a map of inter-community communications. Sectors are labeled by the corresponding group identifier (see Table 1) and colored accordingly. Note that 0 here encodes the group of all accounts belonging to groups smaller than 0.1% of system size, i.e., with less than 400 users, and we are showing only the 10% strongest interactions for sake of clarity. The network of groups exhibits a non-trivial connectivity pattern, typical of communication systems.

3. Results

Functional organization of the communication network. We have enriched the network map with information obtained from the analysis of the contents shared during the observational period. Specifically, we focused on the URLs appended to messages, and we have labeled each URL according to the political leaning (left, left-center, neutral, right-center and right) of its media source and the type of source (political, satire, mainstream media, science, conspiracy/junk science, clickbait, fake/hoax) as manually classified by external experts. We refer to Materials and Methods for details. For each group separately, we quantify the distribution of media sources by political leaning and type, showing the results in Fig. 3. Strikingly, this analysis highlights the existence of large differences across groups. Communities are characterized by very different mixes of typologies of news sources and by very different prevailing ideological orientations, and in all communities one or more forms of political bias prevail in the composition with respect to neutral orientation. The most socially threatening forms of misinformation from a public health perspective, namely conspiracy/junk science and fake/hoax, are highly represented in certain communities and practically absent from others. One community is for instance largely characterized by sharing clickbait URLs, whereas a few communities focus on the circulation of conspiracy and fake/hoax media sources while mostly disseminating content from right-center-wing and right-wing sources. Communities where a prevalence of neutral, left-center-wing and left-wing media sources is observed, are more orientated to disseminate mainstream media and scientific information. However, it is also worth pointing out that in practically all communities, circulation of information directly coming from certified scientific sources is quite marginal, and that the main carriers of reliable information are still the mainstream media, which not incidentally are a majoritarian typology of content source in most communities. Despite this, the fact that such relatively more reliable information is seamlessly mixed in most

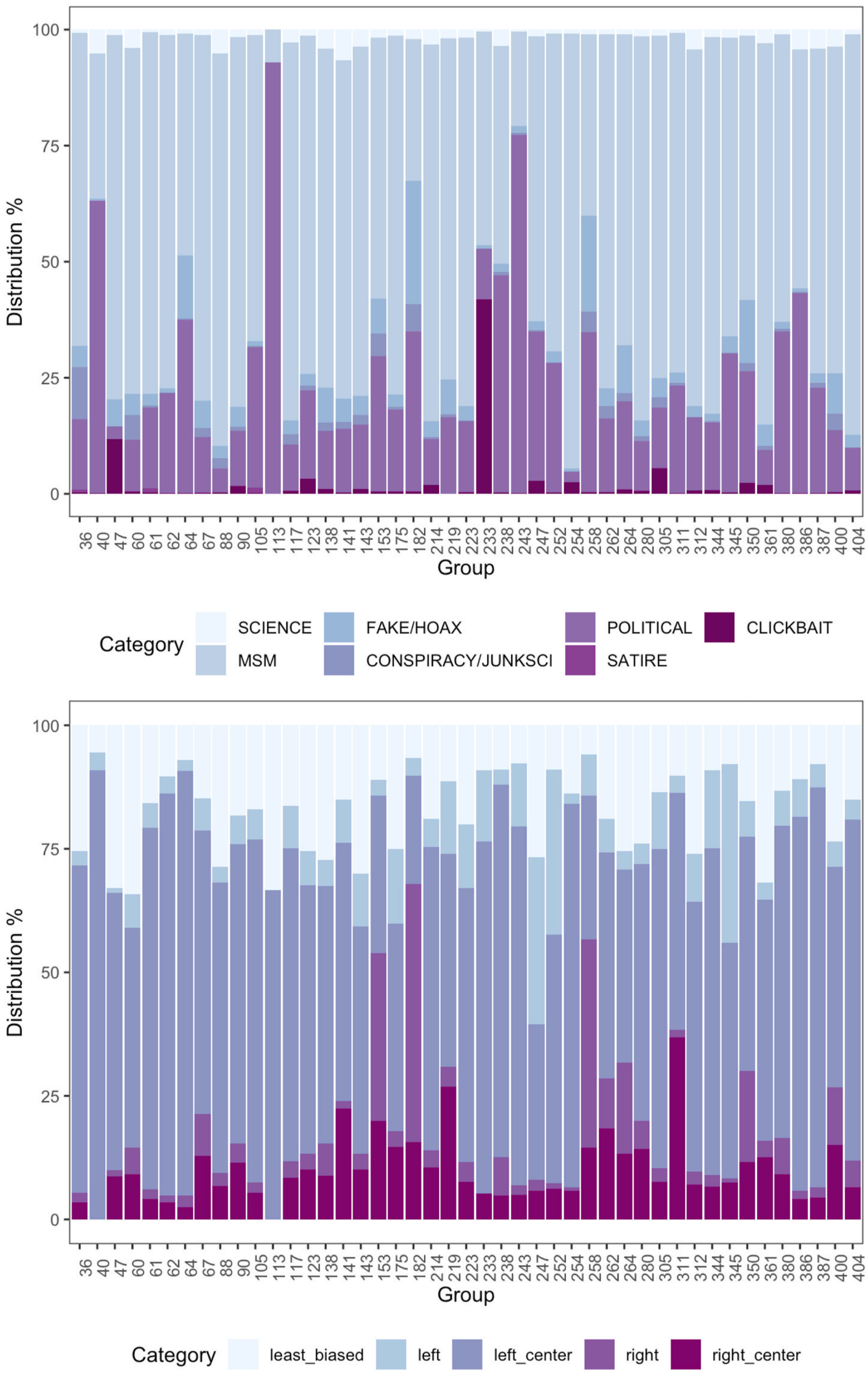

Fig. 3. Functional organization of the COVID-19 communication network. User accounts are enriched by the type of information they share, quantified from the classification of URLs appended to their messages during the observational period. We have two types of information: one, shown in the top panel, related to the media source classification provided by human experts in terms of: mainstream media, science, political, fake/hoax, conspiracy/junk science, clickbait and satire; another one, shown in the bottom panel, providing a finer classification of political media sources in terms of ideological orientation (left, left-center, neutral (least biased), right and right-center).

communities to partisan or intentionally misleading information may nevertheless contribute to the confusion and disorientation of the public opinion (Vraga et al., 2011). The above reported information confirms that knowledge communities tend to be characterized by political polarization, and that such polarization reflects into varying attitudes toward typologies of news sources with different reliability, thus potentially generating online echo chambers (Flaxman et al., 2016). However, the extent to which information circulation within the community is widely distributed vs. traceable to a relatively small number of very active subjects makes a big difference in terms of the group dynamics of polarization and of the prevailing informational biases. We therefore move on to examine this key feature of the communities' structure.

Social activities and information sharing are highly centralized. To shed light on information centralization patterns, we quantify how many users are responsible for how many social media activities and news shared during the observational period. The results of our analysis, shown in Fig. 4, highlight an impressive localization around a few accounts, quantified by an inequality index (measured by Gini coefficient) of 0.72 in the case of the interactions and 0.65 in the case of shared news. More specifically, 0.1% of online users in each group account for 5%–45% of all the interactions and for 1%–10% of news shared, whereas

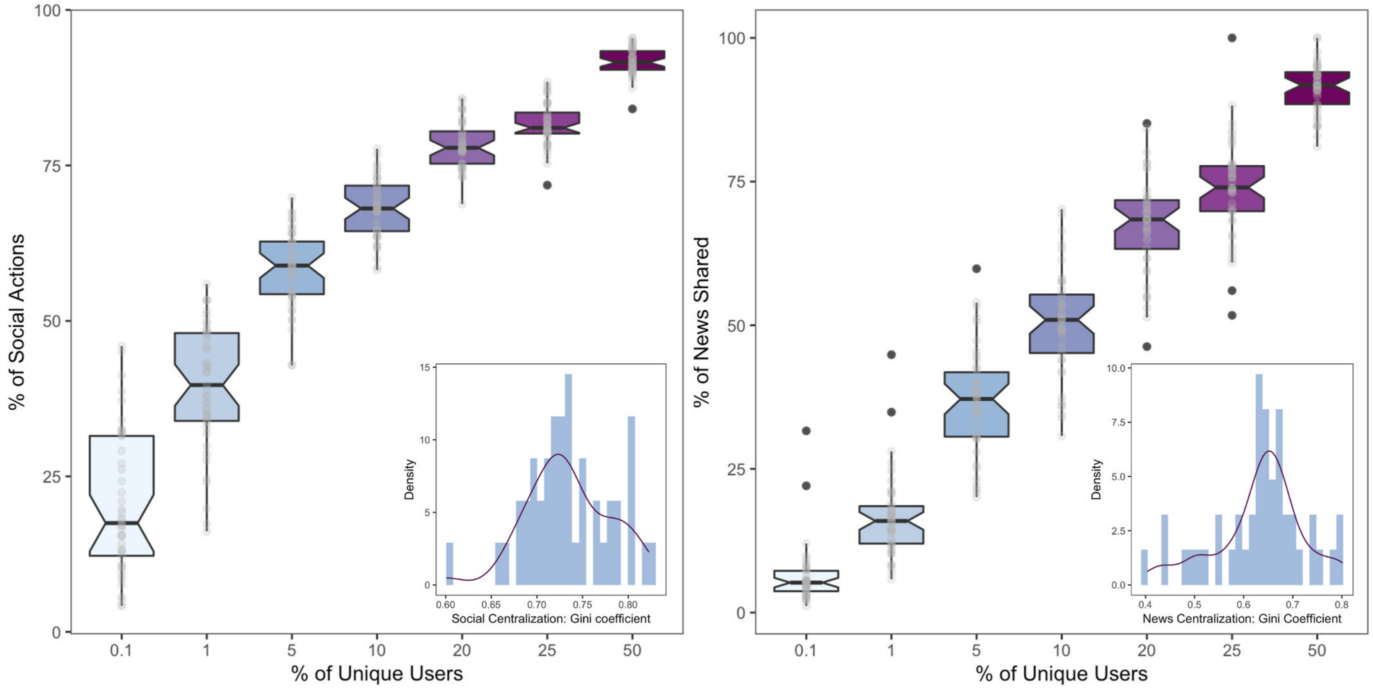

Fig. 4. Social activities and information sharing is highly centralized rather than decentralized. Left: The fraction of social activities versus the fraction of unique users they involve is shown by box plots encoding the distribution across distinct groups and transparent points encoding the empirical values. The analysis reveals that a median of about 20% of activities involves only 0.1% of accounts, whereas up to 50% of activities are accounted for by 1% of users, denoting a striking centralization of actions. The inset shows the distribution of the Gini coefficient, an independent measure to quantify distribution inequality, with an average of 0.72 which is dramatically high. Right: as for the left panel but considering the fraction of news shared instead of social actions. Again, a high centralization is observed, with 1% of users accounting for up to 25% of circulating news (median of about 20%), and an average Gini coefficient of 0.65, still very high.

a 10% of online users in each group account for 55%–75% of the internal interactions and for 30%–70% of news shared. This large difference between users can also be appreciated by observing directly the distributions of the total number of social activities of users in the same community (see Fig. S2), which appear to be fat tailed independently of the communities' size. For many communities, the distribution can be compared to a power-law fit. The extremely skewed shape of these distributions is what yields the observed high values of inequality in the centralization of information, ultimately captured by the Gini coefficient, similarly to how a Pareto distribution of income in a country reflects high inequality. It is remarkable that, despite a difference in centralization between social activities and shared news, in both cases we find a surprisingly high level of inequality. This suggests that news filtering and curation is as an essential aspect of opinion leadership in knowledge communities as opinion making itself. These levels of inequality imply in particular that the actual contribution of most members is fattening the community's critical mass, acting as amplifiers of the opinion leaders' messages. This suggests in turn that we are still very far from the ideal of an inclusive, participatory digital ecosystem, and that at the moment the potential of digital participation is mainly being deployed, as far as COVID-related information is concerned, for the purpose of ideological mobilization. In this kind of social environment, large-scale circulation of reliable information as a basis for effective public health interventions may be problematic. Certain communities may become relatively impermeable to such information if it is de-legitimized by the opinion leaders' judgment or simply filtered off by their news curation. But is such opinion leadership prevalence typical of the upper structural layers of the knowledge community, or are also smaller subgroups typically characterized by similarly localized forms of information centralization? This is the question we must address now.

Non-trivial architecture of online communication. In order to validate our results, we artificially built possible simulations of the identified network structure. The simulations are based on random variables automatically generated by the computer. Comparing these simulations with the model built on the basis of the analysis of the dataset, it has been possible to certify that the latter had an organizational specificity that cannot be ascribed to random interaction effects. Indeed, one can wonder whether the observed patterns are trivially related to the connectivity of each group. To verify this hypothesis, for each group separately we build 20 independent realizations of a null model which preserve the connectivity distribution while destroying existing topological correlations. We consider four distinct measures of topological correlations: average local and global transitivity (quantifying the tendency of accounts to local triadic closure), assortative mixing (quantifying the tendency of accounts to connect with accounts with similar number of connections), and modularity (quantifying the organization of accounts in groups within the group) (Newman, 2003; Boccaletti et al., 2006; Castellano et al., 2009). For each observed group and its random realizations, we measure these four descriptors and show the results in Fig. 5. Our analysis reveals that the organization within each group is far from trivial: the under-abundance of triadic closure indicates that information flow and discussions are mostly pairwise with respect to random expectation; the high disassortative mixing, quantified by more negative assortative mixing than random expectation, denotes that social activities involve accounts with rather different number of connections, e.g. between "hubs/authorities'' and more peripheral users; the higher than expected-by chance modularity denotes the existence of sub-groups within each group, a non-trivial organization within the mesoscale organization of the whole communication system. Taken together with the high value of modularity observed across the whole network, these results indicate the presence of a hierarchical system highly segregated at both macro- and meso-scales. The opinion leadership scheme therefore structures the community at all scales, down to the smaller subgroups, so that the whole architecture of the community can be described as a social hierarchy of influencers. This kind of structure essentially collides with the idea of stimulating a public

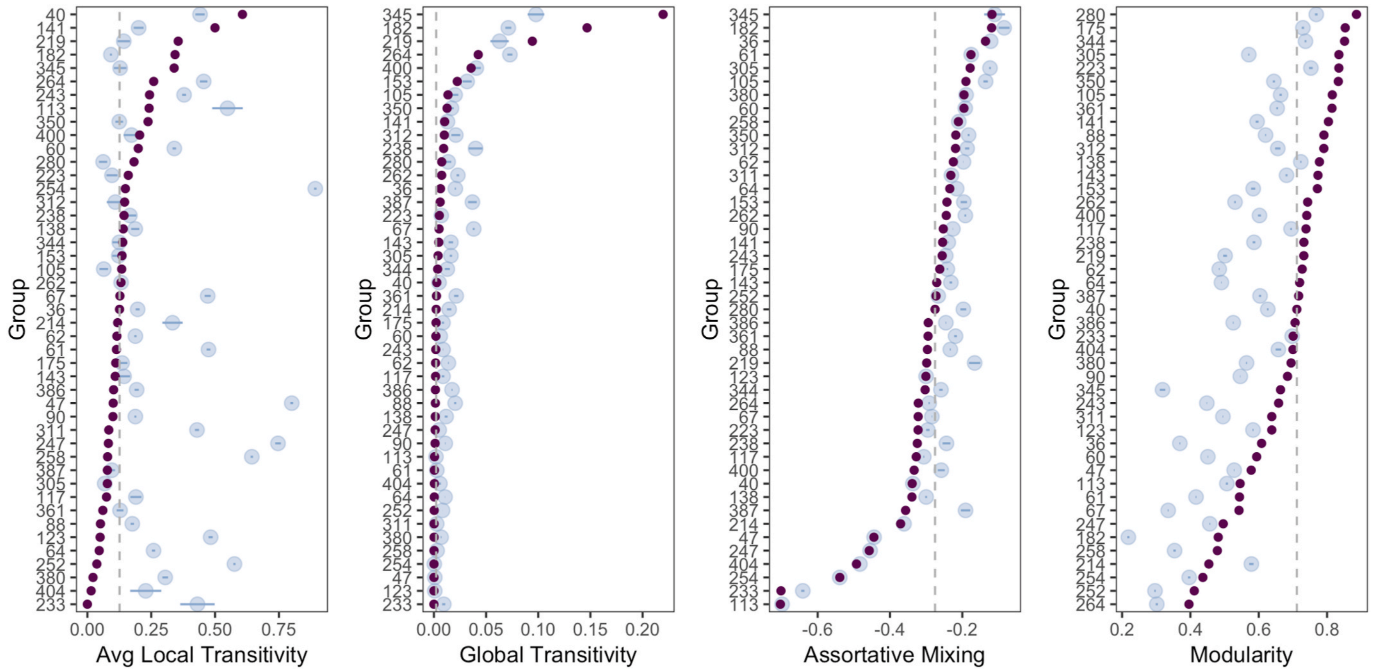

Fig. 5. Connectivity patterns of the communication network are not trivial. The structure of the communication network is tested against its configuration model, preserving its connectivity distribution while washing out topological correlations. For each group separately, we measure average local and global transitivity (quantifying the tendency of accounts to local triadic closure); assortative mixing (quantifying the tendency of accounts to connect to accounts with similar number of connections); and modularity (quantifying the organization of accounts in groups within the group). Values estimated for the observed groups are encoded by solid dark points, whereas values obtained from null models (a baseline equivalent to the system studied where most correlations are destroyed) and averaged across 20 independent realizations are shown with lighter markers and segment denoting the 95% variation around the expected values. The vertical dashed lines encode median values across groups. Overall, the results indicate that some measured features are not observed by chance: most of the groups are characterized by a lack of triadic relationships and a stronger organization into sub-groups, a hallmark of hierarchical organization.

debate and disseminating information around sensitive issues on the basis of a model of evidence-based argumentation, debunking and fact-checking. The hierarchy of influencers is the gateway through which the community receives and sends information, and insofar as the community's social organization is shaped in this way, there is little possibility that information not socially validated by influencers has a chance to have a real impact on its members. In turn, the hierarchical structure enables higher-level influencers to leverage upon powerful "multiplier" effects, suitably orientating lower-level influencers, so as to reach even the most marginal community members at a considerable level of granularity.

Human coding the network map. So far, our unsupervised analyses revealed interesting structural and semantic features which are not compatible with random expectation. To better characterize the identified groups, we have human coded them by manually inspecting the accounts, and their publicly available discussion, for the top 20 users of each community, for a total of 900 human-coded users (see Materials and Methods for the adopted protocol). A statistical summary of manual user type and role labels is found in Fig. 6. There, we can see that the largest fraction of most prominent users have a mediatic role (News Media, News Blog, Blogger, Journalist, Public Communication summing up to more than 60% of opinion leaders). However, the second most frequent role (16%) is that of Troll and, even considering our selection keywords focusing on the medical aspect of COVID-19 during the months of its global expansion, only a small fraction of opinion leaders belongs to the medical or academic roles (Physician, Professor, Scientist summing up to about 3%) or other types of professional (Entrepreneur, Attorney, Civil servant sum < 1%). Contrary to what one might imagine, only a minority (32%) of opinion leaders are individuals sharing an independent opinion, while 41% are institutional accounts and 27% have been identified as social bots. The communities also present differences as to the geographical provenience of their leaders. Many of them focus upon a single country, whereas others mix influencers of different origins. In Fig. 7 we display the geographical provenience of opinion leaders for 20 of the largest groups. We find, as expected, multinational newsgroups (groups 90, 117, 141), but the majority of groups is tightly knit around leaders from a specific country (groups 47, 60, 153, 214, 238, 254, 311, 350, 400, plus other groups not displayed in the figure, see Supplementary Tab. S1). In between, there are a number of international groups (64, 67, 123, 233, 243, 280, 305) which are centered around a country-specific range of topics that attract international attention. Using all the information about role, type, and nationality of community members, as well as group-level statistics, we have manually classified each group as reported in Table 1. As discussed above, a large number of groups match the communication sub-network of a single country. In some cases, most notably USA and Venezuela, the country-level discourse appears to be fragmented into multiple groups, based upon the political orientation of their leading influencers. One would expect that COVID-related knowledge communities would be mostly international in scope and tightly connected, in order to quickly and effectively share valuable information and best practices to facilitate a globally coordinated policy response to the pandemic crisis. However, what we find instead is that they tend to be country-specific and often ideologically focused. For certain countries, the national conversations are literally broken down into 'parallel realities': that is, different communities with opposing ideological orientations whose reciprocal interaction is very limited. The picture that emerges from these results is that the cosmic web of COVID-19 poorly functions as a global commons for the timely circulation of knowledge and for effective, collaborative problem solving of COVID-related issues. Rather, it works as an ideological arena where COVID-related communication is embedded in a wider geopolitical discourse that reflects the evolution of the multipolar world order, and where knowledge communities are largely impermeable to each other and show little concern for mutual cooperation in the

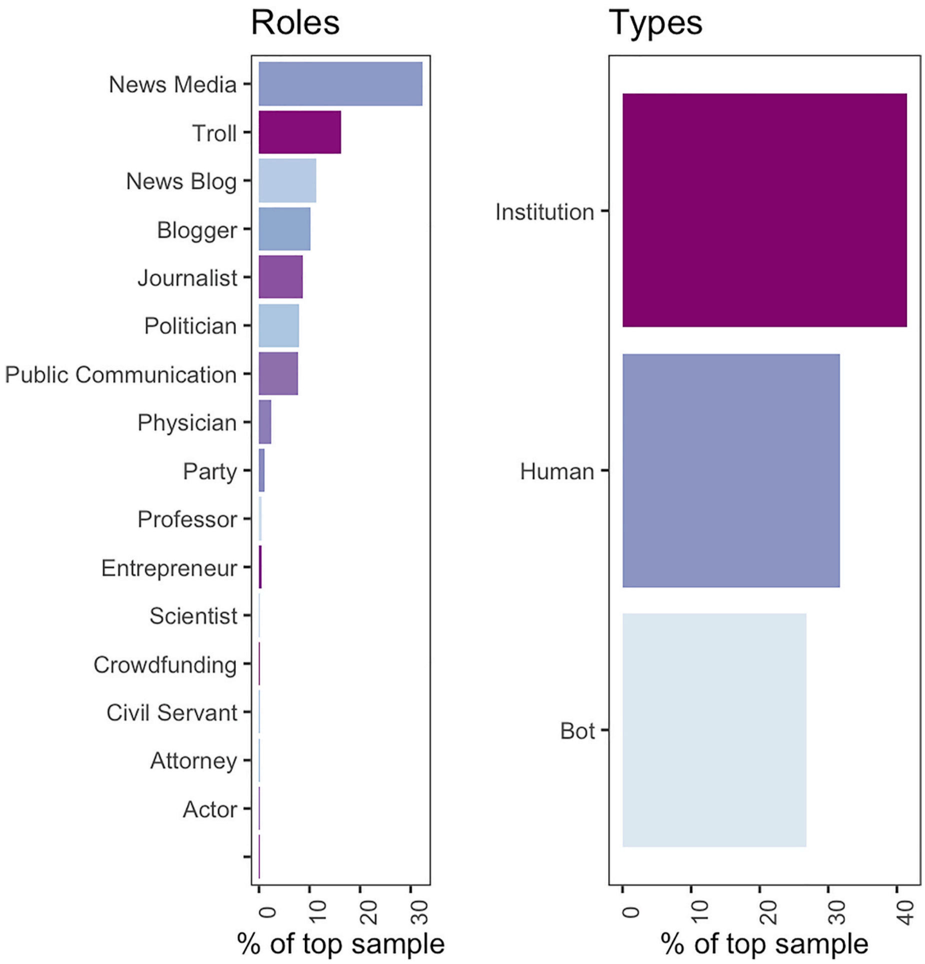

Fig. 6. Aggregated statistics of the roles and types of the top 20 most active accounts for each group. Thanks to our annotation (see Methods) we gain insights on the most active users leading the Twitter conversation about COVID-19 in different communities. Users are labeled according to the different role played (left panel) and their types (right panel): i) institutional (e.g., public communication, news media agencies); ii) human (e.g., bloggers, journalists, politicians, physicians and trolls); and iii) bots functional (news blogs and automated trolls). Official accounts of institutional news media are the most present, followed by trolls (about 90% are bots) and accounts handled by communication professionals (automated news blogs, human bloggers, journalists, politicians and public communication agencies). A very minor role is played by health and science experts (physicians, professors and scientists). Note that the last row in the left panel refers to non-annotated users.

public interest.

4. Discussion

To tackle infodemics as a major public health issue and to build an evidence-based, scientific approach to policy design, we believe that the knowledge community mapping approach developed here is an indispensable informational base. By better understanding the nature of knowledge communities in terms of their characteristic news sources, ideological orientations, and geographical areas of reference, we can develop and test more effective countering and mitigation strategies, such as targeted information inoculation interventions, and prospectively aim towards an integrated approach to public health where the epidemic and infodemic dimensions are dealt with as complementary aspects of an overarching socio-medical challenge. In this regard, we consider our study as a plausible methodological building block of a new, highly cross-disciplinary science of infodemiology (Tangcharoensathien et al., 2020).

Our results show that the structure of the web of COVID-19-related knowledge communities is highly hierarchical and modular. This seems to suggest that, from an infodemic perspective, the global digital knowledge ecosystem is much less inclusive than what is commonly thought. On the other hand, it is clear that the transition from pre-digital, vertical media ecologies to highly pluralistic digital ones has been very recent, and that even in best case scenarios, adapting to an entirely different mode of massively horizontal knowledge and information creation and dissemination, as potentially enabled by digital technologies and media, is a large-scale regime change that takes time (Sacco et al., 2018). Our results are not particularly surprising as to the fact that there is strong ideological polarization across different knowledge communities and that many of them are centered upon specific geographical areas. However, ours is, to our knowledge, the first large-scale study that identifies the structure of the main knowledge communities of the COVID-19 related Twittersphere. The fact that such communities confirm expectations of political polarization and geographical focus is a meaningful result in itself, and has policy implications of special interest in view of the scale and public health consequences of the phenomenon.

In addition to the fact that influencers matter in knowledge communities, our finding that, more specifically, the whole organization can be basically characterized in terms of a modular hierarchy of influencers at all scales is far from obvious. This has important policy implications in view of the controlling and filtering roles of community gatekeepers in facilitating vs. blocking the access of community members to certain sources of information through various means such as attentional cues, strategic legitimization vs. discrediting of sources, stigmatization of diverging opinions, and so on. In this regard, what is really surprising about our results is, in other words, not the fact that the organizational architecture is hierarchical in itself, but the actual size of the effect. Also, the impressive levels of informational centralization we find are not only limited to opinions, where they can be expected to some extent, but also apply to the predominance of news media sources across communities.

It should also be added that the aggregate weight of public figures such as journalists and public communicators is equally remarkable, and can be considered as a further reinforcement of the preponderance of traditional news sources. For these reasons, one possible strategy for countering the negative spillovers of infodemics phenomena in the digital environment is paradoxically to support the social salience of

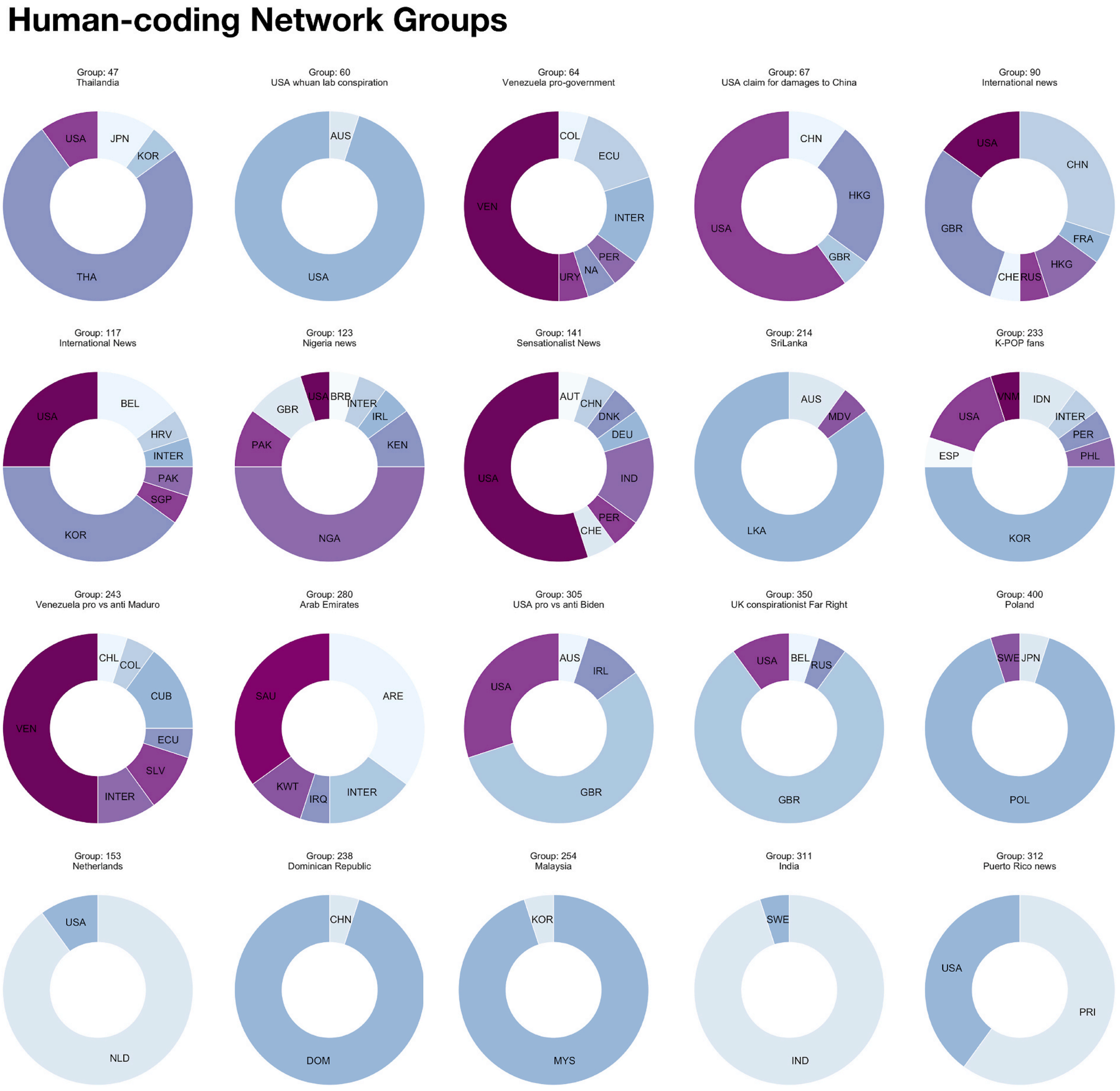

Fig. 7. Statistics of the geographical origin of the top 20 most active accounts for a selection of groups. The 20 groups, whose most active influencers belong to more than a single country, are shown. For graphical reasons we excluded the group 105 "Germany", with 19 influencers from Germany and one from Mongolia, while all other 23 groups have only a single country represented (see Supplementary Materials). The donut charts show the distribution of accounts' nationalities for each group, where INTER stands for *international*. In some cases the distribution is heterogeneous, like for groups discussing international news. Discussion around Venezuela politics also involves various influencers from Latin America. The peculiar community of K-Pop fans, while centered around a large fraction of South Korean influencers, breaks cultural and linguistic borders being, together with the "Nigeria news" group, the community with the most diverse influencer origins. Notice how the most heterogeneous groups often present highly peculiar mixes of countries that seem to reflect complex geopolitical patterns.

mainstream media. In particular, providing economic incentives for quality investigative journalism to disengage mainstream media products from the logic of the attention economy could be an important step in building a resilient media ecosystem (Kaye and Quinn, 2010). However, looking at other actors involved in the information cascades, the aggregate weight of news blogs and individual bloggers, though not matching that of news media alone, comes relatively close to it, suggesting that the new forms of decentralized news and opinion making also play a significant role in the whole ecosystem. In this sense, it is difficult to tell whether the current gap will close or further widen in the future, and it is clear that the current infodemic is shaped by a very complex interaction of news sources whose relative salience differs across knowledge communities.

Another finding is probably less expected though, and causes concern: if we are not aggregating across affine role types, the second most prominent role is played by trolls – a clear sign that the generation of noise and misinformation and the disruption of constructive debate and argumentation is a widespread feature of the digital ecosystem, and a particularly dangerous one from a public health perspective. It is also worth noticing that social bots account for more than one quarter of all manually classified agents. From an infodemic perspective, this is probably the most critical aspect, and the one that calls for immediate action. Currently, the data at our disposal show how fortunately Twitter itself has promptly intervened in blocking the most problematic users. For example, in the case of Indonesia it was possible to notice that most of the trolls and bots were effectively expelled from the platform in the months following the outbreak of the pandemic.

Despite the effort made by Twitter to contrast media manipulation activities, scientific communication is still an extremely marginal fringe. Our results point out the need for scientists and experts to become more

Table 1
List and description of group IDs used in Figs. 2, 3 and 5.

Group ID	Human Coding	Group ID	Human Coding
36	Italy	233	K-POP fans
40	Ecuador Institution	238	Dominican Republic
47	Thailand	243	Venezuela pro vs anti Maduro
60	USA whuan lab conspiration	247	Philippines
61	France	252	USA anti Trump
62	Spain	254	Malaysia
64	Venezuela pro-government	258	Pro Trump
67	USA claim for damages to China	262	Canada
88	Japan nationalist news	264	Iran Political Dissidents
90	International news	280	Arab Emirates
105	Germany	305	USA pro vs anti Biden
113	Venezuela local news	311	India
117	International News	312	Puerto Rico news
123	Nigeria news	344	USA sport
138	Turkey	345	USA left
141	Sensationalist News	350	UK conspirationist Far Right
153	Netherlands	361	Indonesia
175	USA	380	Brazil
182	USA Far Right	386	Mexico
214	Sri Lanka	387	Chile
219	Pakistan	400	Poland
223	USA NY news	404	USA florida news

visible and influential in the structure of knowledge communities, and calls for a bigger emphasis on the public intellectual side of scientific activity as a necessity to increase the efficacy of public health measures. Some interesting experiments in this regard directly involve Twitter, as in the case of Tweetorials: attempts at anti-misinformation inoculation under the form of COVID-19 themed threads posted by physicians and epidemiologists in the context of #MedTwitter and #EPITwitter communities (Graham, 2021). Given the self-defensive nature of conspiratorial knowledge communities, it is not granted that, as already noted, such information may directly reach its members, being preliminarily filtered out by the community influencers. However, there is increasing scientific consensus over the viability of targeted information inoculation strategies as a means for creating large scale psychological resistance against fake news (van der Linden et al., 2020) as a form of 'pre-bunking' rather than debunking. Insofar as hard-to-dismantle, conspiratorial knowledge communities remain relatively circumscribed, such strategies, rather than trying to penetrate closely guarded groups, aim at blocking their expansion by making the public opinion more refractory to the acceptance or mere consideration of fake news. Preliminary evidence seems to show that such strategies may be effective in countering dangerous social attitudes such as vaccine hesitancy, enabling people to better discern accurate news from fake content, and to make them less willing to lend attention to sensationalist, likely unreliable claims (van der Linden et al., 2021). Given their already emphasized prominence in the current Twittersphere, mainstream media outlets may play an important role by coordinating with scientific and policy experts in the design and implementation of societal campaigns of information inoculation (Limaye et al., 2020).

Looking in more detail at the characterization of the single communities, we notice that several of them have a clear geographical focus and ideological orientation. However, certain groups are also representative of complex emergent global cultural phenomena, such as the community of K-pop fans, with its mixed international composition, which has recently caught the attention of international news media for its rapid and unexpected mobilization potential (Li and Jung, 2018). This map reflects a complex geo-political pattern, which seems to be shaped by the current fluid phase of the new multipolar world order (Diesen, 2019). The thematic characterization of the knowledge communities shows however that political orientations play a much stronger structuring role on community identity than opinions and positions about the pandemic itself, and this introduces a further element of complication in the strategic design of infodemic countering or mitigation, in that ideological positions, which reflect in turn political agendas at various geographical scales, mainly shape the informational architecture of COVID-related online conversations.

A key direction for future research on the basis of our results is better understanding the relationship between the infodemic as an overabundance of information, and the centralization of information that reflects into the shaping of a segregated and hierarchical web of knowledge communities. There are several effects, already noted in the literature, that might concur to this phenomenon, and which need to be investigated much more closely to understand their relative importance. The first is the well-known effect of information overabundance on the reduction of people's choice menus, which in the sphere of news consumption leads to characteristic news consumption patterns (Pentina and Tarafdar, 2014). The second is the effect of preferential attention toward perceived influencers caused by information deluge in the light of collective attention mechanisms (De Domenico and Altmann, 2020). The third is the role of prestige and content biases in cultural transmission processes in the selection of relevant information in a context of overabundance (Berl et al., 2020). A thorough analysis of these effects to elucidate the micro-physics of information centralization and community formation in the context of the COVID-19 pandemic clearly calls for a very large and diversified spectrum of competences, and gives a clear idea of the challenge ahead in building a new, trans-disciplinary science of infodemiology.

Finally, we highlight the main limitations of our study. As it is well known, the demographics of Twitter users are biased toward well-educated males (65 percent of Twitter users) between the ages of 18 and 34 (58 per cent of Twitter users, according to Statista GmbH). Our results have to therefore be interpreted keeping such demographic limitations in mind. However, it is important to consider that to the current state of knowledge there is no way to build a potentially unbiased, representative sample of the public opinion at the regional, national or global level, and to track its time evolution for relatively long periods. It will however be important to expand these methods to cover several social media at once, whose combined demographics allow the coverage of different portions of the public opinion. Another important limitation is the necessarily limited choice of hashtags which, although carefully designed, inevitably miss those parts of the social media flow that are not tagged according to the most common signifiers. Finally, the social media discourse on COVID-19 related infodemics may be highly dependent on the local cultural, social and political context, and that its more subtle nuances may only be captured through an expert analysis of posts in the local languages. Both such limitations call for a huge effort in designing and implementing a global infodemic project that makes use of the experience and expertise of a number of local specialists to build a protocol that can effectively adapt to local contexts while maintaining an overall methodological consistency.

Contributors

MDD designed research. NC and MDD collected the data.NC, FP, RG and MDD elaborated the data. PS, RG and MDD analyzed and interpreted the data. FP, PS, RG and MDD wrote the manuscript.

Data sharing

Data describing the network of interaction studied in this paper and the code for reproducing all results is available upon request.

Declaration of competing interest

The authors declare no competing financial interests and no conflict of interests.

Acknowledgments

PS and MDD acknowledge the support of the European Union's Horizon 2020 Project inDICEs (Grant No 870792). The funding source had no role in: designing the study; data collection, analysis or interpretation; writing the manuscript; decide where to submit it for publication.The authors thank the WHO EPI-WIN and the WHO Infodemic Working Group useful discussions. MDD also acknowledges the Universitat Rovira i Virgili for granting access to their high-performance computing facilities.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.socscimed.2021.114215.

References

Abel, G.J., Sander, N., 2014. Quantifying global international migration flows. Science 343, 1520–1522.

Allen, D., 2016. Education and Equality. University of Chicago Press.

E. Bakshy, B. Karrer, L. A. Adamic, Social influence and the diffusion of usercreated content, in: Proceedings of the 10th ACM Conference on Electronic Commerce, pp. 325–334.

Barber'a, P., Jost, J.T., Nagler, J., Tucker, J.A., Bonneau, R., 2015. Tweeting from left to right: is online political communication more than an echo chamber? Psychol. Sci. 26, 1531–1542.

Bergstrom, A., Jervelycke Belfrage, M., 2018. News in social media: incidental consumption and the role of opinion leaders. Digital Journalism 6, 583–598.

Berl, R.E., Samarasinghe, A.N., Roberts, S., Jordan, F., Gavin, M.C., 2020. Prestige and Content Biases Together Shape the Cultural Transmission of Narratives.

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

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.-U., 2006. Complex networks: structure and dynamics. Phys. Rep. 424, 175–308.

de Bruin, W.B., Saw, H.-W., Goldman, D.P., 2020. Political polarization in us residents' covid-19 risk perceptions, policy preferences, and protective behaviors. J. Risk Uncertain. 61, 177–194.

Castellano, C., Fortunato, S., Loreto, V., 2009. Statistical physics of social dynamics. Rev. Mod. Phys. 81, 591–646.

Chadwick, A., 2017. The Hybrid Media System: Politics and Power. Oxford University Press.

Choi, M., Glassman, M., Cristol, D., 2017. What it means to be a citizen in the internet age: development of a reliable and valid digital citizenship scale. Comput. Educ. 107, 100–112.

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C.M., Brugnoli, E., Schmidt, A.L., Zola, P., Zollo, F., Scala, A., 2020. The covid-19 social media infodemic. Sci. Rep. 10, 1–10.

Colleoni, E., Rozza, A., Arvidsson, A., 2014. Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. J. Commun. 64, 317–332.

Diesen, G., 2019. The disorderly transition to a multipolar world. New Perspect. 27, 125–129.

De Domenico, M., Altmann, E.G., 2020. Unraveling the origin of social bursts in collective attention. Sci. Rep. 10, 1–9.

De Domenico, M., Lima, A., Mougel, P., Musolesi, M., 2013. The anatomy of a scientific rumor. Sci. Rep. 3, 2980.

Douglas, K.M., 2021. Covid-19 conspiracy theories. Group Process. Intergr. Relat. 24, 270–275.

Editorial, 2020. The truth is out there, somewhere. Lancet (London, England) 396, 291.

Enders, A.M., Uscinski, J.E., Klofstad, C., Stoler, J., 2020. The different forms of covid-19 misinformation and their consequences. The Harvard Kennedy School Misinformation Review. https://misinforeview.hks.harvard.edu/article/the-different-forms-of-covid-19-misinformation-and-their-consequences/.

Flaxman, S., Goel, S., Rao, J.M., 2016. Filter bubbles, echo chambers, and online news consumption. Publ. Opin. Q. 80, 298–320.

Gallotti, R., Valle, F., Castaldo, N., Sacco, P., De Domenico, M., 2020. Assessing the risks of 'infodemics' in response to covid-19 epidemics. Nature Human Behaviour 4, 1285–1293.

Gil de Zu'ñiga, H., Jung, N., Valenzuela, S., 2012. Social media use for news and individuals' social capital, civic engagement and political participation. J. Computer-Mediated Commun. 17, 319–336.

Glück, J., 2019. Wisdom vs. populism and polarization: learning to regulate our evolved intuitions. Applying Wisdom to Contemporary World Problems. Springer, pp. 81–110.

Goel, S., Anderson, A., Hofman, J., Watts, D.J., 2016. The structural virality of online diffusion. Manag. Sci. 62, 180–196.

Graham, S.S., 2021. Misinformation inoculation and literacy support tweetorials on covid-19. J. Bus. Tech. Commun. 35, 7–14.

Holliman, R., 2011. Telling science stories in an evolving digital media ecosystem: from communication to conversation and confrontation. J. Sci. Commun. 10, C04.

Jiang, J., Chen, E., Yan, S., Lerman, K., Ferrara, E., 2020. Political polarization drives online conversations about covid-19 in the United States. Human Behavior and Emerging Technologies 2, 200–211.

Jolley, D., Douglas, K.M., 2014. The effects of anti-vaccine conspiracy theories on vaccination intentions. PloS One 9, e89177.

Kaye, J., Quinn, S., 2010. Funding Journalism in the Digital Age: Business Models, Strategies, Issues and Trends. Peter Lang.

H. Kwak, C. Lee, H. Park, S. Moon, What is twitter, a social network or a news media?, in: Proceedings of the 19th International Conference on World Wide Web, pp. 591–600.

Lazer, D., Pentland, A.S., Adamic, L., Aral, S., Barabasi, A.L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., et al., 2009. Life in the network: the coming age of computational social science. Science (New York, NY) 323, 721.

Leetaru, K., Wang, S., Cao, G., Padmanabhan, A., Shook, E., 2013. Mapping the Global Twitter Heartbeat: the Geography of Twitter. First Monday.

Li, H., Jung, S., 2018. Networked audiences and cultural globalization. Sociology Compass 12, e12570.

Limaye, R.J., Sauer, M., Ali, J., Bernstein, J., Wahl, B., Barnhill, A., Labrique, A., 2020. Building trust while influencing online covid-19 content in the social media world. The Lancet Digital Health 2, e277–e278.

van der Linden, S., Roozenbeek, J., Compton, J., 2020. Inoculating against fake news about covid-19. Front. Psychol. 11, 2928.

van der Linden, S., Dixon, G., Clarke, C., Cook, J., 2021. Inoculating against covid-19 vaccine misinformation. E-Clinical Medicine 33.

Milberry, K., Anderson, S., 2009. Open sourcing our way to an online commons: contesting corporate impermeability in the new media ecology. J. Commun. Inq. 33, 393–412.

Newman, M.E., 2003. The structure and function of complex networks. SIAM Rev. 45, 167–256.

Orso, D., Federici, N., Copetti, R., Vetrugno, L., Bove, T., 2020. Infodemic and the spread of fake news in the covid-19-era. Eur. J. Emerg. Med.: Off. J. Eur. Soc. Emerg. Med. 27 (5), 327–328. https://europepmc.org/article/med/32332201.

O'Connor, C., Murphy, M., 2020. Going viral: doctors must tackle fake news in the covid-19 pandemic. BMJ 24, m1587.

Parkins, J.R., Mitchell, R.E., 2005. Public participation as public debate: a deliberative turn in natural resource management. Soc. Nat. Resour. 18, 529–540.

Pennycook, G., Rand, D.G., 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. Proc. Natl. Acad. Sci. Unit. States Am. 116, 2521–2526.

Pentina, I., Tarafdar, M., 2014. From "information" to "knowing": exploring the role of social media in contemporary news consumption. Comput. Hum. Behav. 35, 211–223.

Prior, M., 2005. News vs. entertainment: how increasing media choice widens gaps in political knowledge and turnout. Am. J. Polit. Sci. 49, 577–592.

Romer, D., Jamieson, K.H., 2020. Conspiracy theories as barriers to controlling the spread of covid-19 in the us. Soc. Sci. Med. 263, 113356.

Sacco, P.L., Ferilli, G., Tavano Blessi, G., 2018. From culture 1.0 to culture 3.0: three socio-technical regimes of social and economic value creation through culture, and their impact on european cohesion policies. Sustainability 10, 3923.

Shelley, M., Thrane, L., Shulman, S., Lang, E., Beisser, S., Larson, T., Mutiti, J., 2004. Digital citizenship: parameters of the digital divide. Soc. Sci. Comput. Rev. 22, 256–269.

Struckmann, S., Karnowski, V., 2016. News consumption in a changing media ecology: an mesm-study on mobile news. Telematics Inf. 33, 309–319.

Sutton, R.M., Douglas, K.M., 2020. Conspiracy theories and the conspiracy mindset: implications for political ideology. Current Opinion in Behavioral Sciences 34, 118–122.

Tangcharoensathien, V., Calleja, N., Nguyen, T., Purnat, T., D'Agostino, M., Garcia-Saiso, S., Landry, M., Rashidian, A., Hamilton, C., AbdAllah, A., et al., 2020. Framework for managing the covid-19 infodemic: methods and results of an online, crowdsourced who technical consultation. J. Med. Internet Res. 22, e19659.

Vraga, E.K., Edgerly, S., Wang, B.M., Shah, D.V., 2011. Who taught me that? repurposed news, blog structure, and source identification. J. Commun. 61, 795–815.

Wardle, C., Derakhshan, H., 2017. Information disorder: toward an interdisciplinary framework for research and policy making. Council of Europe Report 27, 1–107.

World Health Organization, et al., 2020. Novel coronavirus (2019-ncov). Situation Report 13.

Zarocostas, J., 2020. How to fight an infodemic. Lancet 395, 676.