Global misinformation spillovers in the online vaccination debate before and during COVID-19

Jacopo Lenti[1], **Kyriaki Kalimeri**[1], **Andre Panisson**[2], **Daniela Paolotti**[1], **Michele Tizzani**[1], **Yelena Mejova**[1], **Michele Starnini**[1,3]
[1]ISI Foundation, Torino, Italy; [2]Centai, Torino, Italy
[3]Departament de Fisica, Universitat Politecnica de Catalunya, Barcelona, Spain
Keywords: *misinformation, social media, cross-national communication, public health*

Anti-vaccination views pervade online social media, fueling distrust in scientific expertise and increasing vaccine-hesitant individuals. Thus far, the scientific study of the debate around vaccination on online social media (OSM) has focused on specific countries [1, 2, 3, 4] or English-speaking users [5]. Upon arrival of COVID-19, it is now imperative to understand the flows of anti-vaccine — or *no-vax* — information not only nationally but internationally, in order to have a bird-eye view on the topic and inform effective communication campaigns. To address this need, in this work we build and analyze a series of international information flow networks by leveraging 316 million Twitter posts related to vaccines in 18 different languages from a pre-COVID era to April 2021. To this aim, we first investigate (i) how polarized, in terms of echo chambers phenomenon, the vaccination debate is in different countries, over time, to identify users in no-vax communities and (ii) how susceptible, in terms of circulation of information, are these no-vax communities to low quality information. We propose a flexible, language-neutral community detection approach, and combine it with human-in-the-loop expert knowledge to track polarization and echo chambers in different countries across time. Below, we highlight some of the results of this work.

We start by selecting 4 three-months periods before and during the pandemic, which we dubbed as i) pre-COVID, ii) pre-vaccine, iii) vaccine development, and iv) vaccine roll-out periods. We then select 28 countries in Europe, America and Oceania having at least 2000 unique users in each period. We construct the retweet (RT) networks corresponding to each country and time period, detect communities by using hierarchical clustering, and label a sample of tweets from each community to identify clusters of users exposed to no-vax content. We found 52 of these "no-vax" communities. Note that this does not imply all users in these communities hold anti-vaccination opinions, but that they are more likely to be exposed to such material. We find that no-vax communities are generally present in English-speaking countries, with respect to Spanish speaking ones. However, some of the relatively largest country-specific no-vax communities appear in France, Italy, Netherlands, Poland, and the United States.

Turning to potential echo-chambers in these networks, we first quantify the degree of polarization in the vaccination debate by using the Random Walk Controversy (RWC) score [6], which measures how much users in no-vax communities are exposed to information coming from their own side vs. the rest of the network. The RWC score is overall very high, indicating that **the vaccination debate is generally highly polarized**. However, it decreases substantially over time, suggesting that users in novax communities became less isolated in the vaccination discourse during the COVID pandemic. Secondly, we investigate whether the users in the no-vax communities are exposed to information sources different from the rest of users [7]. To this aim, we look at the content shared by the users, constructing a co-sharing (CO) network where users are connected by a link if they share the same URL, and gauge the similarity between the RT and CO networks by computing the Normalized Mutual Information (NMI) between their community structures. On average, the NMI of the networks with a novax community is higher than the others (0.27 vs 0.22, $p < 0.05$), indicating that users **in no-vax communities tend to have common information sources**. Some countries, such as the U.S. and Brazil, show an especially high NMI, indicating that the polarization in the retweet network is reflected in the different content shared.

Considering the behavior of **users in no-vax communities, we find that they are more likely to retweet, share URLs, and especially URLs to YouTube than other users**. Furthermore, the URLs they post are much more likely to be from low-credible domains (identified using lists of such domains in 4 languages), compared to those posted in the rest of the networks. The difference is remarkable: **26.0% of domains shared in no-vax communities come from lists of known low-credible domains, versus only 2.4% of those cited by other users** ($p < 0.001$).

Next, we investigate the effects of content moderation by Twitter on the vaccination debate. We find that the average proportion of suspended accounts in no-vax communities is much larger than the rest of users, for each country and period considered (average 13.3% vs 1.8%, $p < 0.001$). **A large portion of suspensions come after the January 2021 U.S. Capitol attack in Washington, D.C.**[1] These findings suggest that political leaning is often associated with strong stances taken in the vaccination debate (in line with previous literature [1, 4]) and that actions taken in the political domain may greatly impact the quality of the public health discourse.

Next, we quantify the information spillover across countries by considering the number of retweets from one country to another, normalized by the total number of retweets produced and received in the two countries (Fig 1a). We find

[1]The suspensions were announced by Twitter https://blog.twitter.com/en$_us/topics/company/2021/protecting--the-conversation-following-the-riots-in-washington--$

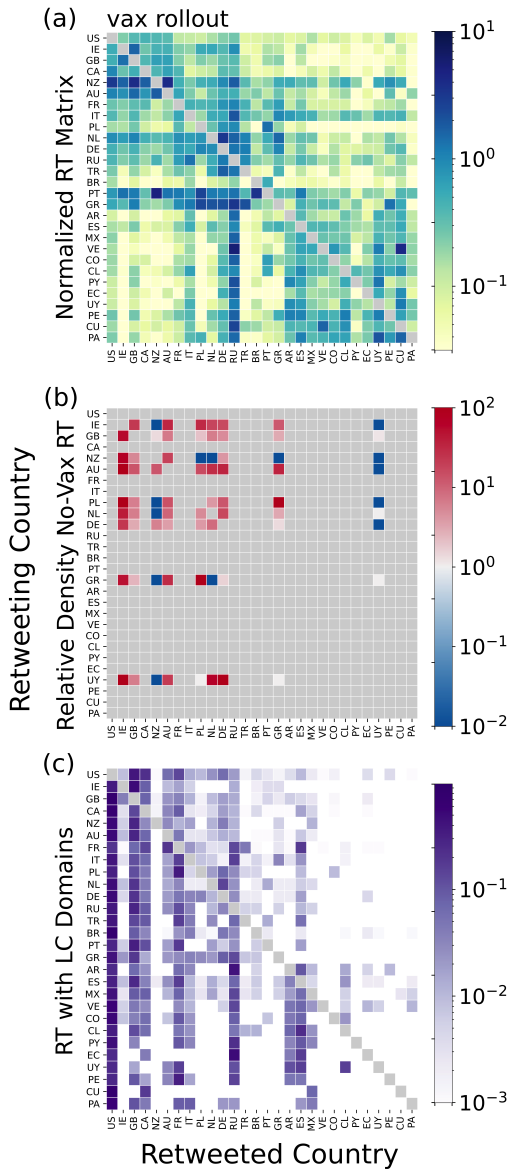

Fig. 1. **Cross-border information flows in the global vaccination debate for last period – vaccination rollout.** (a) Normalized number of retweets (excluding diagonal elements from the plot, colored in grey), (b) Probability of interaction between users in no-vax communities from one country to another, with respect to the interactions between other users from the same pair of countries (see Methods). Darker red (blue) elements of the matrices represent higher (lower) tendency of cross-border interactions between users in no-vax communities with respect to other users (countries without no-vax communities colored in grey), (c) Proportion of URLs that come from Retweeted Country among the low-credible domains imported by Retweeting Country (countries importing less than 10 low-credible URLs are coloured in grey). Element a_{ij} of each matrix represents information flow from country j to country i.

that the cross-border interaction matrices are not symmetric: information generally flows with a preferred direction. For instance, Spanish-speaking countries retweet English-speaking ones much more than the opposite. **The United States is central in the global information flow (despite flows being normalized)**, being a net exporter of information to the rest of the world. Interestingly, from pre-vax period, Russia is also a net exporter, especially to South American countries: some of the most used hashtags in pre-vaccine and vax development periods are #sputnikesesperanza and #sputnikparaelpueblo.

Next, we quantify the strength of cross-border interactions between users in no-vax communities with respect to the rest of users (Fig 1b). We find that cross-border interactions between users in no-vax communities are generally much stronger, sometimes by orders of magnitude, than interaction from the rest of users, creating a **tightly-knit global no-vax network**. In particular, users in no-vax communities of English-speaking countries, Germany, and the Netherlands are tightly connected in all periods. Conversely, users in no-vax communities from Cuba and Russia are quite isolated.

Finally, we focus on the misinformation flows across countries by considering the fractions of low-credible domains imported per country (Fig 1c). As in the previous case, the matrices show a clear asymmetry. **U.S. users act as global misinformation superspreaders to the rest of the world**: 68% of all low-credible URLs retweeted worldwide come from U.S. (average over the four periods), a proportion much higher than the total volume (42%) retweeted from U.S.. Interestingly, the fraction of low-credible URLs coming from U.S. dropped from 74% in the vax development period to 55% in the vax rollout. This large decrease can be directly ascribed to Twitter's moderation policy: 46% of cross-border retweets of U.S. users linking to low-credible websites in the vax development period came from accounts that have been suspended following the U.S. Capitol attack. Finally, despite not having a list of low-credible domains in Russian, **Russia is central in exporting misinformation in the vax rollout period, especially to Latin American countries**. In these countries, the proportion of low-credible URLs coming from Russia increased from 1% in vax development to 18% in vax rollout periods.

In conclusion, despite the platform's tweet flagging and removal policies around COVID-19, it is the bout of account suspensions around the Washington riots that made the most impact on the national and international spread of vaccine-related misinformation, suggesting that the political concerns elicit much stronger curbing of the freedom of speech than the health one. As interaction with vaccine hesitant social media content has been related to an increased delay of vaccination [8], the lack of action in the first three periods of study may have contributed to the unnecessary deaths of unvaccinated individuals (estimated to be in hundreds of thousands in the U.S. alone [9]). In ongoing work, we are developing quantitative tools in order to gauge to what extent anti-vaccination sentiments expressed on Twitter relate to politicization of the topic by the political actors within each country.

[1] Alessandro Cossard, Gianmarco De Francisci Morales, Kyriaki Kalimeri, Yelena Mejova, Daniela Paolotti, and Michele Starnini. Falling into the echo chamber: the italian vaccination debate on twitter. In *Proceedings of the International AAAI conference on web and social media*, volume 14, pages 130–140, 2020.

[2] Giuseppe Crupi, Yelena Mejova, Michele Tizzani, Daniela Paolotti, and André Panisson. Echoes through time: Evolution of the italian covid-19 vaccination debate. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 102–113, 2022.

[3] Mauro Faccin, Floriana Gargiulo, Laëtitia Atlani-Duault, and Jeremy K Ward. Assessing the influence of french vaccine critics during the two first years of the covid-19 pandemic. *arXiv preprint arXiv:2202.10952*, 2022.

[4] Matt Motta, Dominik Stecula, and Christina Farhart. How right-leaning media coverage of covid-19 facilitated the spread of misinformation in the early stages of the pandemic in the us. *Canadian Journal of Political Science/Revue canadienne de science politique*, 53(2):335–342, 2020.

[5] Ana Lucia Schmidt, Fabiana Zollo, Antonio Scala, Cornelia Betsch, and Walter Quattrociocchi. Polarization of the vaccination debate on Facebook. *Vaccine*, 36(25):3606–3612, 2018.

[6] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Quantifying Controversy in Social Media. In *WSDM '16: 9th ACM International Conference on Web Search and Data Mining*, pages 33–42, 2016.

[7] Bjarke Mønsted and Sune Lehmann. Characterizing polarization in online vaccine discourse—a large-scale study. *PloS one*, 17(2):e0263746, 2022.

[8] Sahil Loomba, Alexandre de Figueiredo, Simon J Piatek, Kristen de Graaf, and Heidi J Larson. Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa. *Nature human behaviour*, 5(3):337–348, 2021.

[9] Krutika Amin, Jared Ortaliza, Cynthia Cox, Joshua Michaud, and Jennifer Kates. Covid-19 mortality preventable by vaccines. *Health System Tracker*, 2022.