CESifo WORKING PAPERS

10184
2022

Original Version: December 2022
This Version: March 2023

Vax Populi: The Social Costs of Online Vaccine Skepticism

Matilde Giaccherini, Joanna Kopinska, Gabriele Rovigatti

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 10184*

Vax Populi: The Social Costs of Online Vaccine Skepticism

Abstract

This paper quantifies the impact of online vaccine skepticism on pediatric vaccine uptake and health outcomes. We propose a novel methodology that combines Natural Language Processing and an instrumental variable strategy that leverages the intransitivity of the social network's connections. By matching the universe of Italian vaccine-related tweets for 2013-2018 with vaccine coverage and preventable hospitalizations at the municipality level, we find that a 10pp increase in anti-vaccine sentiment causes i) a 0.43pp decrease in coverage of the Measles-Mumps-Rubella vaccine, ii) additional 2.1 hospitalizations among vulnerable populations per 100,000 residents, and iii) an 11% increase in the relevant healthcare expenses, equivalent to 7,311 euros. Drawing on the results of a simulated model, we further show the importance of targeted interventions to counter misinformation and improve vaccine uptake.

JEL-Codes: I180, D850, L820, C810.

Keywords: social network, Twitter, vaccines, controversialness, polarization, text analysis.

*Matilde Giaccherini**
University of Rome 'Tor Vergata' / Italy
matilde.giaccherini@uniroma2.it

Joanna Kopinska
University of Rome 'La Sapienza' / Italy
joanna.kopinska@uniroma1.it

Gabriele Rovigatti
Bank of Italy / Rome
gabriele.rovigatti@bancaditalia.it

*corresponding author

March 10, 2023
We would like to thank Sofia Amaral-Garcia, Tiziano Arduini, Federico Belotti, Vincenzo Carrieri, Francesco Decarolis, Ruben Durante, Ludovica Gazze, Leonardo Giuffrida, Sauro Mocetti, Tommaso Orlando, Francesco Pierri, Monou Prem, Francesco Sobbrio, Lucia Rizzica, Lorenzo Rovigatti, Gianluca Russo, for useful suggestions and the participants of the ASSA Annual Meeting 2023, CESifo Area Conference in Economics of Digitization 2022, EuHEA Seminar 2022, SEHO Conference 2022. Special thanks go to Fabian Baumann, Philipp Lorenz-Spreen, Igor M. Sokolov, and Michele Starnini for sharing their code before its publication. The views expressed in this paper are those of the authors and do not involve the responsibility of the Bank of Italy. The usual disclaimers apply. The authors acknowledge the financial support of the Einaudi Institute for Economics and Finance (EIEF) and the computational support of ISCRA/CINECA.

1 Introduction

The phenomenon of misinformation is deeply ingrained in contemporary society, impacting political, economic, and social well-being (Vosoughi et al., 2018). Before the COVID-19 era, a supposed link between pediatric vaccines and autism was one of the most propagated fake news, stemming from A. Wakefield's 1998 Lancet article on the trivalent Measles-Mumps-Rubella (MMR) vaccination (Jolley and Douglas, 2014, Leask et al., 2006, Opel et al., 2011). Although the article has been retracted, and despite overwhelming evidence supporting the safety and efficacy of vaccines, this piece of disinformation remains widespread (see among others Allcott et al., 2019, Chiou and Tucker, 2018).

The diffusion of internet and, more recently, the rise of social media have provided an unparalleled platform for the dissemination of similar takes on vaccines (Burki, 2019),[1] as they have granted virtually unlimited access to information not subject to fact-checking or editorial judgment. As a result, the ability of consumers to discriminate between true and "fake" (or unsubstantiated) news has decreased. Additionally, social networks dynamics usually foster the creation of ideological "echo chambers" (Cinelli et al., 2021) that fuel polarization (Azzimonti and Fernandes, 2022, Flaxman et al., 2016, Sunstein, 2001, 2017, 2018), ideological self-segregation (Berinsky, 2017, Gentzkow and Shapiro, 2011, Mullainathan and Shleifer, 2005) and misinformation diffusion (Allcott and Gentzkow, 2017).

Due to the spread of false claims about the vaccine-autism correlation, an increasing number of parents choose not to vaccinate their children, exploiting the benefits of the immunity granted by vaccinated peers (Esposito et al., 2014, Smith et al., 2017). In Italy, as in other countries, decreasing vaccination rates have led to outbreaks of previously-controlled diseases such as measles. This has sparked a policy debate and prompted the implementation of legal measures that impose costs on those who choose not to vaccinate. Although vaccine mandates curtailing individual freedom of choice have always been controversial, the Italian healthcare department has argued that falling uptake poses a risk not only to the eligible but also to vulnerable individuals who cannot be vaccinated.[2]

In this ongoing conflict between personal interests and public health endeavors, one of the most important pieces of the puzzle relates to the effects of online vaccine skepticism on vaccination rates and vaccine-preventable diseases. If skepticism spread through social media has a sizeable impact on vaccine hesitancy, addressing it could help individuals make better decisions in their own best interest. Furthermore, since communicable diseases can create significant externalities - including health risks, higher hospitalization rates and

[1] In addition, the fact-checking standards on social media are often lax, and the emotional appeal of such messages can make them particularly effective in spreading quickly (Zhuravskaya et al., 2020).

[2] These include infants aged 0-12 months, pregnant women, and immunosuppressed patients.

costs for individuals not targeted by vaccination campaigns - any comprehensive analysis should consider these additional costs imposed on the society.

We empirically quantify these effects on local public health outcomes such as vaccination rates, vaccine-preventable hospitalizations, and their relative costs. We focus on Italy in the 2013-2018 period, just around a major reform (2017) that expanded the pediatric vaccine mandate to include MMR.[3] The new law was triggered by years of declining vaccine coverage and followed months of heated debates in offline and online communities. We use Twitter data to analyze the spread of vaccine skepticism and construct a measure of vaccine-related attitudes at the local level. The data include all vaccine-related tweets in Italian. To determine their stance on vaccines, we develop an anti-vaccine (hereafter *anti-vax*) classifier using a Natural Language Processing (NLP) algorithm as in Polignano et al. (2019). According to Kim (2022), Twitter data can be used to accurately track public attitudes towards policy-relevant topics over time and across different locations (see also Grossman et al., 2020, Jin et al., 2021). Hence, we use the average anti-vaccine sentiment expressed in geolocated tweets as a stand-in for the anti-vaccine movement's relevance and spread across Italian municipalities.

In estimating the causal relationship between exposure to anti-vaccine views on social media and the health outcomes of vaccine hesitancy we face two issues: (i) the endogeneity that characterizes such relationship, and (ii) the lack of data on individual vaccine hesitancy - these are typically unavailable and cannot be linked to social media accounts.

To address the first point and formalize the sources of endogeneity, we analyze the evolution of individual vaccine stances on Twitter through the lens of a model of opinion dynamics in social networks, borrowed from Baumann et al. (2020). We show that even moderate degrees of homophily, when combined with strongly controversial topics, endogenously result in the formation of echo chambers and opinion polarization. Our dynamic model highlights two complementary effects: firstly, the "*link formation effect*" - i.e., users are more inclined to share content and establish connections with individuals showing similar beliefs - and secondly, the "*exposure effect*" - i.e. users' stances are swayed by the opinions they are exposed to, in particular, by those at the extremes of the spectrum. The former channel implies that the amount of exposure to anti-vaccine content is endogenously determined by the users' own stances; the latter is the focus of our analysis. To estimate it, we employ an Instrumental Variables approach that leverages the intransitivity of network connections, as described in Bramoullé et al. (2009).[4] More specifically, we take as a reference every user's network at the end

[3]Until 2017, Italy required only four vaccines (polio, diphtheria, tetanus, and hepatitis B, often combined with haemophilus influenzae type-b and whooping cough), but the mandate was rarely enforced. Vaccines for MMR, chickenpox, meningococcal, and pneumococcal were strongly recommended, so their use was at the discretion of parents. Only in late 2017 the scope of mandatory pediatric vaccines became legally enforceable upon school enrollment, with a one-year transitional period allowing parents to comply with the new rules.

[4]The relation between A and B, and between B and C is *transitive* if it implies a relation between A and C. Social networks (either offline or online) are *intransitive* as long as being linked to a user does not necessarily imply being directly connected to that users' friends.

of the sample period (i.e., when all link formation effects have taken place), and instrument the individuals' stances with the average value of the contents that they have been exposed to through their indirect connections - the friends of their own friends. Under standard assumptions, such "*friends of friends*" exposure constitutes an exogenous source of variation.

To overcome the lack of individual data on vaccine hesitancy and its outcomes, we pair aggregated vaccine-related tweet stances with disease-specific vaccine coverage rates, vaccine-preventable hospitalizations, and relative costs at municipal level. We are able to exploit both the power of the individual-level Twitter data and the highly detailed municipal data on vaccinations, hospitalizations and health-related costs through a Mixed Two-Stage Least Squares (M2SLS) approach (Dhrymes and Lleras-Muney, 2006). Building on the individual-level first-stage regression, we aggregate the instrumented variable at municipal level to obtain a valid causal estimate of the exposure effect.

Our estimates show that exposure to online vaccine skepticism causes a significant reduction in vaccination rates for the MMR, particularly targeted by anti-vax misinformation. However, we find no impact on vaccines not affected by fake news (Hexavalent, Meningococcal, Pneumococcal). A 10 pp increase in average vaccine skepticism at the municipality level leads to a 0.43 pp decrease in vaccination coverage (mean value 89.50). Furthermore, vaccine skepticism leads to higher rates of hospitalization for vaccine-preventable diseases, and increased costs. Specifically, a 10 pp increase in the average stance leads to 2.1 additional hospitalizations per 100,000 residents (mean value 22) and an excess expenditure of 7,311 euros, or an 11% increase in the relevant healthcare expenses. We perform several robustness checks to control for the impact of Twitter algorithm changes,[5] local vaccine campaigns, and the impact of populist votes, finding virtually unchanged results. In addition to the baseline analysis, we propose an alternative estimation strategy that addresses potential concerns about the exogeneity of our preferred instrument. The results are comparable to the baseline, both in magnitude and statistical significance.

Finally, we examine the implications of our findings for policymakers and public health organizations in terms of actions to be taken on social media to successfully engage with the public. We first investigate the potential non-linearity in the exposure effects on individual user stances. Specifically, we examine whether pro- or anti-vaccine individuals react differently to the exposure to friends of friends' content - i.e., are more or less receptive to external stances. We find a stronger "persuasion" effect on pro-vaccine users compared to anti-vaccine users, implying that the most effective interventions should be directed toward retaining doubtful individuals, rather than aimed at convincing anti-vax supporters. Second, in the spirit of Athey et al. (2022), we exploit the exogenous timing of events like epidemics, scientific breakthroughs, court rulings, legislation,

[5]Twitter introduced an "ampliphication algorithm" in 2016. As argued by Acemoglu et al. (2021), such algorithms are aimed at maximizing engagement and tend to create more homophilic communication patterns, or "filter bubbles."

and news to test whether their "type"[6] reduce or strengthens the exposure effect. Our results show that political events and news on vaccines coming from national or international institutions (*trustworthy sources*) support pro-vax stances - and lessen the exposure effect of anti-vax content. In addition to these empirical results, we conduct Monte Carlo counterfactual analyses by simulating two alternative scenarios of our dynamic model, either implementing a *Censorship* policy for anti-vax content, or running vaccine *Informative Campaigns*. We find that the latter is the most efficient approach to counteract the effects of online misinformation and to reduce the polarization. These findings imply that social media vaccine awareness campaigns may be a practical and scalable intervention to increase public understanding of public health issues and contain the spread of misinformation.

While a growing body of literature examines the effects of fake news on vaccine hesitancy (Carrieri et al., 2019, Chiou and Tucker, 2018), anti-vaccine beliefs and behavior (Allam et al., 2014), and improving immunization (Alatas et al., 2019), to the best of our knowledge, this is the first paper that jointly (i) uses detailed data at a fine-grained geographical level on vaccination rates and hospitalizations; (ii) provides a data-driven approach to proxy users' stance toward vaccine-related topics; (iii) implements a causal identification strategy at the user level; and most importantly, (iv) quantifies the monetary costs of online vaccine skepticism, distinguishing between the target population and the externalities for the vulnerable individuals not subject to the vaccination campaigns.

This work also contributes to the literature on the effects of vaccine mandates. Previous research has shown that mandates can significantly impact vaccination uptake and decrease the incidence of infectious diseases, such as pertussis, smallpox, chickenpox, and hepatitis A, with large long-term effects on affected individuals (Abrevaya and Mulligan, 2011, Carpenter and Lawler, 2019, Holtkamp et al., 2021, Lawler, 2017). Our results suggest that counteracting the spread of pediatric vaccine skepticism can have a significant impact on immunization. Forced medical interventions are often seen as curtailments of individual freedom, which can lead to controversy and unintended consequences. Athey et al. (2022) have recently shown that social media had a significant impact on self-reported beliefs and knowledge about COVID-19 vaccines through public health organization campaigns on Facebook and Instagram. The results of the study conducted by Larsen et al. (2022) showed that using a counterstereotypical messenger on social media[7] can be a powerful catalyst in encouraging COVID-19 vaccine uptake among the hesitant. Additionally, Breza et al. (2021) found that mobility and COVID-19 infection rates decreased as a result of randomly assigned exposure to Facebook messages encouraging preventive health behaviors. Bailey et al. (2020) also showed that Facebook users with friends exposed

[6] We classify them into four broad categories: vaccine efficacy, statements from trustful institutions, politics and mandates, and allegations that vaccines are unsafe.

[7] They used a Youtube "public service" announcement featuring Donald Trump encouraging his supporters to get vaccinated.

to COVID-19 were more likely to support social distancing and other public health behavior measures. Our findings provide direct evidence of the potential benefits of policies aimed at raising awareness of the risks of communicable diseases and promoting preventive immunization to combat the effects of vaccine skepticism on public health.

2 Institutional background

The advances in vaccine technology have been major contributors to the increases in life expectancy that characterized the 19^{th} and 20^{th} centuries. Paradoxically, due to the past success of collective vaccination efforts, individuals tend to underestimate the value of immunization and are more willing to risk being unprotected. Additionally, the "self-eroding" nature of vaccination can lead to fluctuations in vaccine coverage for newborns, which in turn affect the level of protection for the entire community whenever the so-called herd immunity is not achieved (Siegal et al., 2009). In this sense, the vaccine uptake can be seen as an example of a free-riding problem, where individuals may prioritize their own interests over those of the community when deciding whether or not to get vaccinated. In turn, this can lead to cycles of suboptimal participation in vaccination campaigns.

One of the turning points in the history of Italian vaccine campaigns was the eradication of smallpox between 1978 and 1998, followed by the introduction of hepatitis B and anti-pertussis vaccines. In the early 2000s, the first national vaccination plans were introduced within the context of the National Plan of Vaccine Prevention (PNPV). The PNPV establishes a vaccine calendar and offers eligible individuals free vaccines at Local Health Authorities (LHAs).[8]

Until 2017, the mandate for pediatric vaccines included four shots: polio, diphtheria, tetanus, and hepatitis B, which are frequently combined with haemophilus influenzae type b and whooping cough within the so-called hexavalent or 6-in-1 vaccine. Vaccines for the trivalent MMR, chickenpox, meningo- and pneumococcal diseases were only strongly recommended. In 2012, a local court in Rimini issued a sentence against the Health Ministry based on the (false) claim of a causal link between the MMR vaccine and autism. As a result, the immunization rates started to decrease, reaching the minimum historical coverage rate in 2015 - the year in which the local court of Bologna reversed the controversial Rimini sentence (Carrieri et al., 2019). In response to the falling immunization rates, and a sharp increase in measles cases in Italy, a strong political commitment against anti-vax movements led to the approval of a new PNPV in 2017,[9] this extended the scope of mandatory

[8]The regional authorities implement public health policies through their health departments, while health protection and promotion fall under the responsibility of the Departments of Prevention within the 101 LHAs. LHAs cover on average 590,000 individuals each, and are divided into 711 districts with an average population of 84,000. LHAs manage and deliver vaccinations free of charge to the eligible (pediatric population, the elderly, and other protected categories.

[9]The *Lorenzini*'s Decree.

pediatric vaccines by enforcing them upon school enrollment and introduced harsher sanctions against anti-vax doctors.

Under the 2017 PNPV, the number of mandatory vaccinations increased from four to ten (adding whooping cough, Haemophilus influenzae type b, measles, mumps, rubella, and chickenpox). Although vaccine mandates curtailing individual freedom have always been disputed, the PNPV proposers argued that the falling coverage rate - driven by anti-vax sentiment - created sizable negative externalities, increasing the risk of infection not only for the eligible but mostly for vulnerable individuals not targeted by the vaccination.

3 Data

We collect the data on vaccine skepticism and related discussions from Twitter. Specifically, we used the Twitter Application Programming Interface (API) to retrieve all publicly available tweets, written in Italian, containing vaccine-related keywords and a wide range of information on users for the period 2013-2018. In addition, we hand-collected news-related data from newspapers and official sources of information on topics related to vaccines, including vaccine-preventable disease outbreaks, judicial cases, court rulings, and local or nationwide regulatory interventions.

On the health side, we use two primary data sources. The first contains yearly information on disease-specific vaccination rates provided by the LHAs, aggregated at the municipal level for the period 2013-2018. The second is an administrative dataset on the universe of Italian hospital admissions provided by the Italian Ministry of Health, allowing us to focus on vaccine-preventable conditions in both the target population and the population excluded from the vaccination plan, such as infants aged 0-12 months, pregnant women, and immunosuppressed patients, aggregated at the municipality/year level for the period 2013 to 2016.

3.1 Twitter data

Twitter is the fourth most-used social media platform in Italy, after Facebook, Instagram, and LinkedIn, with 8 million unique users in 2018. Along with TikTok, it has the fastest growing user base. Twitter users tend to be older, and 39% of the users is female. Twitter is also the most populated by news outlets, TV channels, and blogs, as it primarily focuses on spreading information. In addition to its actual users, Twitter content is also spread across other social media platforms, with 84% of the users also using Facebook, 80% also using YouTube, and 88% also using Instagram.[10]

Given the role of Twitter data for spread of information, we exploit the *Academic Research product track* to access the full archive of (as-yet-undeleted) tweets. In addition to the text of the tweets, the API provides

[10]Data from the *Authority for Communications Guarantees* (AgCom).

information about both the tweet and the related user. Net of the text analysis of the tweets, we focus on mapping users in terms of their geolocation and online network. Geolocation data can help us better understand the environments in which target populations live (Martinez et al., 2018). We also use the API to retrieve the complete list of users that each user in our vaccine sample follows and is followed by, which allows us to build user-specific online networks.[11] This allows us to study the interplay between users' conversations on Twitter and their local environments.

Download and filtering. We collect all tweets containing the Italian correspondents of any of the following keywords: "vaccine(s)", "vaccination", "vaccinating", "anti-vax", "vax" for a total of 2,031,448 observations.[12] The current version of the dataset was downloaded on April 23^{rd}, 2021.

Each retrieved object contains i) the plain text of the tweet; ii) the unique tweet ID, the creation date, the count of the associated replies, likes, mentions, retweets, hashtags, and multimedia contents, as well as the tweet-specific location, when available; iii) the user contents: ID, Twitter handle, display name, short bio and a few metrics - the number of friends, followers, and tweets posted - a verified status of the account, date when Twitter was joined, and the location, when available (see Figure 1).[13]

We also collect information on on the *followers* and *followings* (hereafter *friends*). *Followers* are Twitter users who follow a specific user, while *friends* are the Twitter users that a specific user follows and whose content she is directly exposed to. We discuss the specific aspects of the latter group in more detail in subsection 5.1.

Data cleaning. We extracte relevant content from tweets, excluding hashtags, special characters, emojis, and multimedia items. We omit *ex-post* all tweets containing only links or mentions[14] and those produced by accounts that are temporarily unavailable due to violation of the Twitter media policy.[15] We also disregard all tweets referring to pets' vaccinations, those where the string "vax" is only retrieved in a URL contained in the tweet, and those written in other languages. In total we excluded 13.909 tweets.

Within the Twitter sample, we geocode the tweets in three consecutive steps: first, we use the tweet-specific geo-tag information ("Place fields" in Figure 1); second, for the remaining tweets, we rely on the geo-tag information of the users ("location" within the "User Fields" in Figure 1); finally, we exploit Twitter

[11] To date, Twitter API v.2 allows us to retrieve the following/follower structure at the date of the download - which in our case, is the period between May and September 2021. In this sense, we build the network-related variables using the "equilibrium" network, which results from all the (endogenous) interactions across users during the 2013-2018 analysis period.

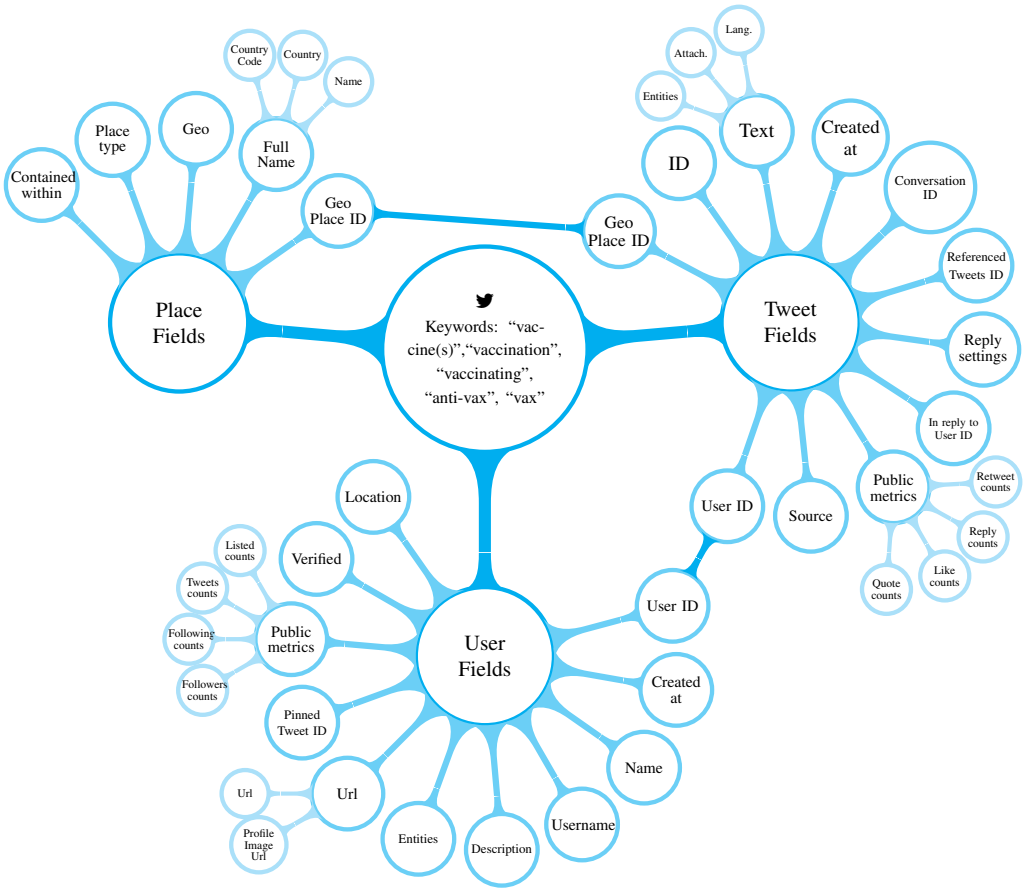
[12] Specifically, we exclude tweets (mainly ads) referring to cow milk ("latte vaccino" in Italian). The specific query reads "(vaccino OR vaccini OR vaccinazione OR vax OR novax OR vaccinarsi OR vaccinato OR vaccinati) -mozzarella -latte lang:it".

[13] No personally identifiable information is included in this study.

[14] A tweet containing another user's username, preceded by "@".

[15] Since 2021, Twitter has applied labels to tweets that may contain misleading information about COVID-19 vaccines and removed the most harmful misleading information from the service.

Figure 1: Twitter objects



Notes: Structure of Twitter objects returned by the API. The structure includes: i) Place fields, ii) Tweet fields and iii) User fields. The former two are matched through the Geo place ID, the latter two are univocally connected through the User ID.

users' profile information with place-name-dictionaries (e.g. "live in Rome"). We map the geocoded tweets to the Italian municipalities based on the latitude and longitude through geospatial shapefiles.[16] Figure A1 in Appendix A shows tweets' distribution across municipalities over time.

We distinguish between original tweets, retweets, and mentions - i.e., the first time an original content appears on the social network, the "plain" copy, and a copy with a comment.[17].

Descriptive Statistics. Out of roughly 2.03 million tweets, the initial cleaning process leaves us with a sample of 2,017,539 tweets related to 227,182 unique users. The geolocalization limits the sample to 830,253 tweets written by 80,471 unique users, distributed across 4,220 municipalities between January 2013 and December 2018. This longitudinal user-specific sample is strongly unbalanced, with only 4.04% of unique users present

[16]Roughly 5% of tweets or users location falling outside the Italian territory is excluded.

[17]We screen tweets' contents for prefixes "RT @", indicating reposting of an original tweet. We identify Twitter handles of the original tweets' creators by extracting the content following "@" and before the main text. Through this procedure, we also identify replies and mentions to original and retweeted versions of the contents

in the whole 6-year period, 7.13% in 5 years, 9.56 % in 4 years, 15.38% in 3 years, 25.35 % in 2 years, and 38.54% in 1 year only. Table 1 reports the main characteristics of the users (panel a), tweets (b), and activity (c) in our sample. On average, users opened their accounts in 2012 and tweeted about vaccines ten times; only 0.7% of them have a verified account.[18]

Table 1: Summary statistics: Twitter data

	median	mean	sd	min	max
Panel a: User characteristics					
Tweets about vaccine	1.00	6.24	32.82	1.00	3,720
Total *tweets*	5,586.00	19,793.54	50,699.13	1.00	1,825,203
Total *followers*	335.00	3,692.14	51,951.40	0.00	3,262,940
Total *friends*	462.00	970.31	2,759.93	0.00	189,582
Account's date of creation		2012	2.49	2006	2018
Verified accounts		0.007	0.084	0	1
Panel b: Tweets' characteristics					
Length of the tweet (number of characters)		102.42	42.05	0	306
Number of words		16.13	6.96	0	62
Retweets (%)		0.60	0.49	0	1
Replies (%)		0.10	0.30	0	1
Panel c: Original Tweets' metrics					
Retweet count		2.59	35.85	0.00	6696
Reply count		0.73	7.10	0.00	1106
Quote count		0.06	1.31	0.00	341
Like count		5.71	90.44	0.00	14,188

Notes: *(a)* summary statistics of 80,471 geotagged users tweeting on vaccines (2013-2018); *(b)* summary statistics of 830,253 geotagged tweets cleaned by hashtag, "RT @", "@", url and emoji; *(c)* Tweet-related popularity metrics of 328,879 original tweets.

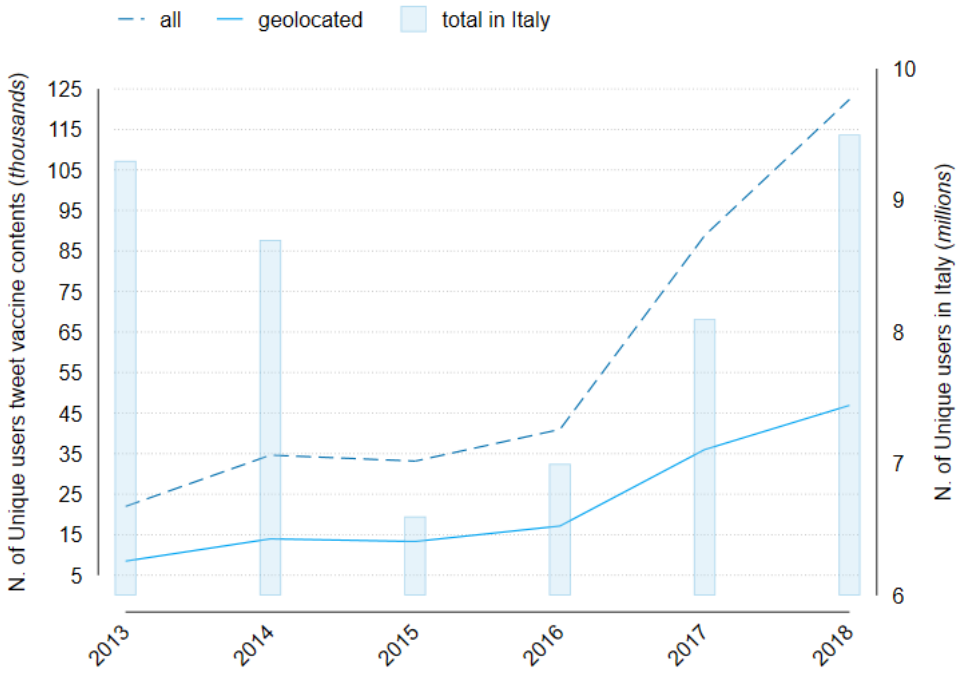
In the sample, 60% of the tweets are retweets or mentions, while 10% are replies. On average, original tweets are retweeted 2.5 times, receive 0.7 replies, 1.6 likes and 0.06 quotes (Table 1).

In Figure 2, we plot the number of unique Twitter users in Italy over time. The bars show the total number of Twitter users, the dashed line shows the number of users who contributed to the Twitter debate on vaccines, and the solid line shows the number of users in the previous group who were geolocalized. The number of users in all three categories shows an increasing trend and peaks at the end of our analysis sample, reflecting the recent growing popularity of Twitter .

Among the geolocalized tweets, 1% has an average of 1 user only tweeting about vaccines in a year. In our analyses, we will disregard this first percentile of municipalities and test the sensitivity of our results to this sample restriction in Table A.9 in Appendix A.

[18] A verified Twitter user in the analysis period was an account of public interest, often belonging to well-known individuals in fields such as music, acting, fashion, politics, religion, news, sports, and business.

Figure 2: Number of unique users



Notes: The figure shows the absolute and geotagged users who tweeted vaccine contents in Italy (left-hand axis) and the total number of Unique users in Italy as reported by AgCom (right-hand axis).

Anti- and pro-vax stances In order to quantify the label vaccine-skeptic tweets, we rely on a Natural Language Processing (NLP) transfer learning model. In particular, we build our model based on BERT, a pre-trained Bidirectional Encoder Representations from Transformers model (Devlin et al., 2018a), trained in Italian (similar in the spirit to the one proposed by Polignano et al., 2019). Specifically, we develop an anti-vax tweet classifier, which we call VaxBERTo, on top of a large pre-trained neural network, providing the very "last mile" of data needed to fine-tune our desider task - i.e., label anti-vax tweets.

First, we construct a training set of tweets pre-labeled as 0/1, with 1 indicating vaccine skeptic content. As in Pierri et al. (2020), our training set is based on tweets from renowed fake news spreaders and vaccine-skeptic users (label 1), pro-vaccine activists and mainstream media outlets (label 0). The training sample consists of 43,472 tweets, split into 20,422 pro-vaccine tweets (46.98%) and 23,050 anti-vaccine tweets (53.02%), relative to a total of 108 unique users. We divide the sample into a training sample consisting of 39,124 tweets ($\approx 90\%$ of the total) and a validation sample of 4,348 tweets to fine-tune the training. Finally, we build a labeled test sample of 4,830 tweets to evaluate the model's performance on a different set of users than those included in the training sample. (See Appendix C for technical details.)

In NLP applications, the performance of a model is directly influenced by the choice of the training sample. In our case, by building the training sample based on pro- and anti-vaccine *users* rather than individual tweets, we are implicitly assuming that i) unlike "fringe" users on Twitter, whose stance on vaccination can change

over time, the users in our training sample consistently express the same stance within their tweets; ii) there are characteristics of language, syntax, and structure that distinguish pro-vaccine and anti-vaccine tweets. As an example, consider the two tweets shown in Figure 3. In panel (a), a popular Italian fake news outlet falsely claims that a baby died as a result of a vaccine. The tweet uses several linguistic constructs commonly found in fake news, such as alluding to conspiracy ("nobody told us about"), attacking mainstream media outlets ("mercenaries," "accomplices"), and expressing doubts and mysteries ("whether or not a link exists"). [19] In contrast, panel (b) shows a tweet from a mainstream media outlet reporting the death of a pediatric leukemia patient from measles contracted from unvaccinated siblings. The language used in this tweet is plain and unemotional, without alluding to conspiracy theories or attacks on mainstream media outlets.

Figure 3: Example of anti-vax (left) and pro-vax (right) tweets used for training

(a) Anti-vax tweet

(b) Pro-vax tweet

Notes: Translation of Italian tweets is provided by Google.

With the trained model we proceeded to classify all the tweets in our sample. Specifically, we generate a label $l_\tau \in 0, 1$ for each tweet τ. Finally, we need to indentify the attitude of users. Following Cinelli et al. (2021), we define the stance of a user according to the average leaning of their tweets. Let i be a user who produces a_i tweets, $C_i = c_1, c_2, \ldots, c_{a_i}$. The activity of user i is given by a_i, and the leaning of each tweet is given by its label l_τ. The individual stance of user i in year t is then their average vaccine stance in that period, which we define as the fraction of tweets with anti-vaccine leaning ($l_\tau = 1$) within their vaccine-related tweets

[19] These linguistic con structures have been e.g. analyzed by Michaels (2008) in the tobacco industry.

in year t. This is given by the following expression:

$$s_{it} \equiv \frac{\sum_{\tau=1}^{a_{it}} c_\tau}{a_{it}} \tag{1}$$

To make the individual stance of a user i more interpretable, we rescale it to a value between 0 and 100 (for example, a user with $s_{it} = 50$ has an equal number of pro- and anti-vaccine tweets, while a user with $s_{it} = 100$ has only anti-vaccine tweets).

3.2 Vaccination data

Through a Freedom of Information Act (FOIA) request, we gathered data on disease-specific vaccination rates at the municipal/year level from every LHAs in Italy.[20] The disease-specific vaccination rates represent the share of the target population that has received the first dose of a vaccine recommended in the national vaccination schedule. The data cover all vaccines included in the Italian routine pediatric immunization schedule: diphtheria*; hepatitis B*, tetanus*, polio*, haemophilus influenzae type B (HIB)**, pertussis** (all included within the hexavalent conjugate vaccine); measles**, mumps**, rubella** (all included within the trivalent conjugate MMR vaccine), meningococcal, and pneumococcal.[21]

Table 2 shows the population-weighted average vaccination rates in the study period, with their median, standard deviation, and minimum and maximum values. As expected, the vaccination rates are strongly correlated across conjugated vaccines (with a pairwise correlation of 0.657 for all hexavalent and MMR individual vaccines), but the levels vary substantially. The hexavalent vaccine shows the highest average vaccination rates (around 94%), likely because it includes four mandatory shots, while the meningococcal vaccine has the lowest vaccination rate (81%).

[20] FOIA grants access to public data regarding data protection regulations.

[21] *: vaccines which were compulsory in Italy between 2013 and 2017.
**: vaccines included in the compulsory list by the "Lorenzin's Law" - Law Decree 73, 2017.
We do not consider the vaccination for chickenpox in our analysis, as a significant portion of the eligible population acquires immunity through natural infection and is exempted from the vaccine mandate.

Table 2: Descriptive statistics of vaccination rates (2013-2018)

		Median	Mean	SD	Min	Max	N
Hexavalent	Diphteria*	94.97	94.29	3.15	54.69	100.00	44,750
	Hephatitis B*	94.80	94.15	3.19	54.69	100.00	44,750
	Polio*	95.00	94.31	3.14	54.69	100.00	44,750
	Tetanus*	95.00	94.38	3.13	54.69	100.00	44,777
	Pertussis**	94.94	94.29	3.14	54.69	100.00	44,750
	HIB**	94.64	94.04	3.17	54.69	100.00	44,749
Hexavalent		94.53	94.09	3.12	54.69	100.00	44,779
MMR	Measles**	91.05	89.52	5.97	10.72	100.00	44,750
	Rubella**	91.00	89.50	5.97	10.72	100.00	44,750
	Mumps**	91.00	89.48	5.96	10.72	100.00	44,750
MMR		91.00	89.55	5.57	10.72	100.00	44,752
Meningococcus		87.40	81.48	15.39	0.17	99.61	43,219
Pneumococcus		91.46	87.26	11.94	.17	100	43,167

Notes: Hexavalent and MMR vaccination rates across 7,929 Italian municipalities for the period 2013-2018. Average values are weighted by the municipality population size. * marks 2013-2017 set of compulsory vaccinations, ** indicates additional mandatory shots introduced by the 2017 Law Decree 73.

3.3 Hospitalization data

The Hospital Discharge Data (SDO), sourced from the Italian Ministry of Health, provides information on the universe of hospitalizations in public and publicly-funded private hospitals for the years 2013-2016. Italy's universal public healthcare system is well-suited to our analysis, as it provides individuals with access to healthcare with minimal or no barriers. In addition, there are no differentials in the expected cost of treatment that could affect vaccine uptake. The records include socio-demographic information (age, gender, nationality, place of birth and residence, educational attainment) as well as clinical data (diagnoses, procedures performed, hospital transfers, discharges) and hospitalization details (hospital type and specialty). Hospital discharge records report information on the primary diagnosis determining each hospitalization, as well as up to five secondary diagnoses.

We focus on the diagnosis of vaccine-preventable diseases in the vaccine-target population and in vulnerable populations that are not targeted by vaccines, such as newborns, pregnant women, and patients with immunosuppressing conditions. These diagnoses are based on the International Statistical Classification of Diseases and Related Health Problems v.9 (ICD-9) codes.[22] Based on the SDO data, we construct municipality-level yearly hospitalization rates and costs per 100,000 residents for both the target and non-target populations.

Table 3 provides a detailed overview of the hospitalization and healthcare costs for different population groups. Figure A2 in appendix A plots monthly trends in hospitalizations among the vaccine-target population and in vulnerable populations that are not targeted by vaccines.

[22]ICD-9 codes for vaccine-preventable diseases are: Rubella 056 and 6475; Measles 055; Diphtheria 032; Pertussis 033 and 4843; Meningococcal 036; Tetanus 037 and 7713; Polio 045–049; Hepatitis B 070[2-3]; Mumps 072; HIB 4822; Pneumococcal 320[1-3] and 481.

Table 3: Descriptive statistics of hospitalizations due to vaccine-preventable diseases (2013-2016)

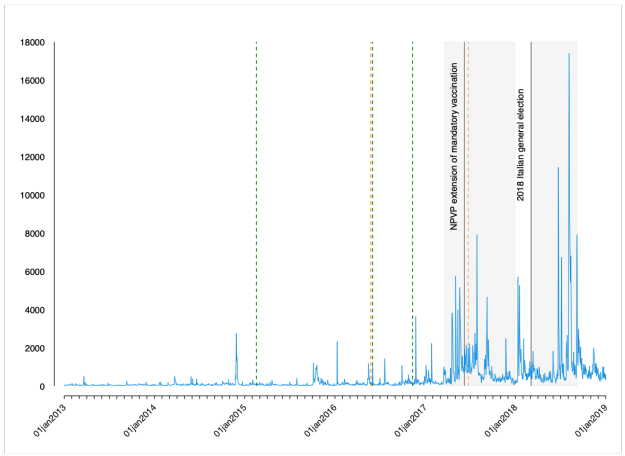
	Median	Mean	sd	Min	Max	N
Panel a: Hopitalizations						
non-target population	14.71	22.21	30.95	0.00	3,202.85	31,760
non-target population (MMR)	0.00	4.99	17.58	0.00	2,846.98	31,760
non-target population (Hexav.)	10.40	16.99	22.02	0.00	355.87	31,760
non-target population (Meningo.)	0.00	0.02	0.26	0.00	29.02	31,760
non-target population (Pneumo.)	0.00	0.88	2.25	0.00	155.04	31,760
Children age 1-10 (MMR)	0.00	2.96	6.87	0.00	1,617.25	31,760
Children age 1-10 (Hexav.)	0.00	1.27	2.70	0.00	152.44	31,760
Children age 1-10 (Meningo.)	0.00	0.04	0.41	0.00	26.21	31,760
Children age 1-10 (Pneumo.)	0.00	0.50	1.76	0.00	132.04	31,760
Panel b: Healthcare costs						
non-target population	38,581.69	66,477.60	116,320.65	0.00	59,880,842.11	31,760
non-target population (MMR)	0.00	15,381.55	96,931.58	0.00	59,880,842.11	31,760
non-target population (Hexav.)	46,275.59	83,151.57	119,925.38	0.00	14,819,697.72	31,760
non-target population (Meningo.)	0.00	150.92	3,976.38	0.00	411,341.22	31,760
non-target population (Pneumo.)	0.00	2,332.30	9,004.03	0.00	1,941,927.83	31,760
Children age 1-10 (MMR)	0.00	4,749.99	25,506.58	0.00	2,274,286.39	31,760
Children age 1-10 (Hexav.)	0.00	2,545.85	9,407.74	0.00	759,286.31	31,760
Children age 1-10 (Meningo.)	0.00	190.58	3,185.72	0.00	409,748.10	31,760
Children age 1-10 (Pneumo.)	0.00	1,255.36	5,365.51	0.00	259,504.65	31,760

Notes: The statistics refer to 7,940 municipalities for the time period between 2013-2016 and are weighted by the municipality population size.

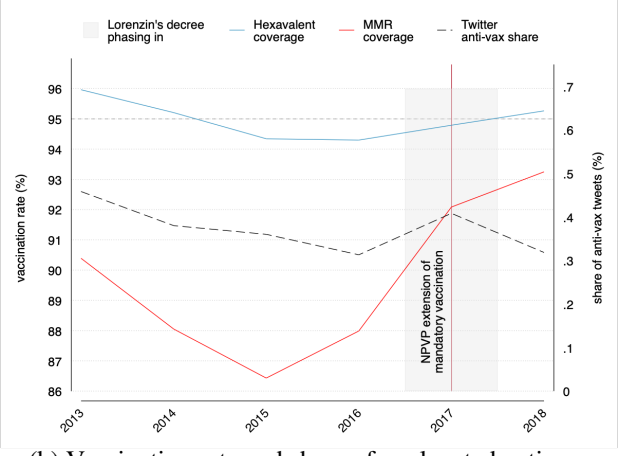
4 Twitter stances and user interactions

Social media platforms actively involve users in the creation of news and place them at the heart of information distribution and discovery. The fundamental component of the social media news distribution system, the user-to-user sharing, results in the dissemination of content online and potentially offline, impacting opinions and real-world behaviors far beyond the platforms' boundaries. Many topics discussed on social media tend to follow typical patterns of attention, in which long-term trends of relatively low interest or *controversialness* are interrupted by sudden spikes of activity. These spikes are often triggered by exogenous shocks (such as unexpected news or events), but on sophisticated platforms they can also be fueled endogenously by algorithms designed to increase users' engagement in the short term (Lorenz-Spreen et al., 2019). This is particularly applicable for Twitter, where algorithmic amplification has been designed to maximize users' exposure to ephemeral, captivating arguments since 2016 (Huszár et al., 2022).

Figure 4: Number of tweets, vaccination rates and anti-vax sentiment in Italy



(a) Number of tweets over time



(b) Vaccination rate and share of geolocated anti-vax tweets

Notes: Panel (a) shows the time series of the number of tweets on vaccinations, 2013-2019. The dashed reference lines report notable (i.e., covered by national media) events regarding vaccination. In particular, they flag i) verdicts (green): the reversal of the Rimini's Court sentence by the Bologna's Appeal Court - February 15^{th}, 2015; the recognition of the inconsistency of the link between the MMR vaccination and autism by the prosecutor of Trani - June 1^{st}, 2016; the dismissal by the court of Milan of the appeal against a sentence establishing the causal link between the vaccine and the severe encephalopathy developed by in an infant - November 10^{th}, 2016; ii) death (orange) of an infant following a mandatory vaccination - May 25^{th}, 2016 and of another infant affected by leukemia of measles contracted from non-vaccinated siblings - June 23^{rd}, 2017. The first grey shaded area marks the period of the debate, which preceded and ensued the approval of "Lorenzin's Law" (June 7^{th}, 2017, solid black line). The second grey area followed the general elections (March 4^{th}, 2018) until the upcoming school starting date - a symbolic moment that created political clashes between the Italian populist parties then ruling the government due to the vaccine mandate's enforcement on school enrollment. Panel (b) reports the yearly average values of hexavalent (solid blue) and MMR (solid red) vaccine coverage rates, as well as the average Twitter anti-vax sentiment (dashed black) as computed in Figure 3.1 recorded between 2013 and 2018.

Based on our data, Figure 4, panel (a) illustrates the dynamics of vaccine-related daily tweets in our sample between 2013 and 2019. The average activity was relatively regular until 2017, when the introduction of the *Lorenzin's law* led to longer and more heated debates, peaking at around 8,000 tweets per day around the approval date. The debate became strongly politicized during the 2018 general election campaign, when populist politicians expressed skepticism about the vaccine mandate.

Figure 4, panel (b), shows the aggregate evolution of coverage rates for both the hexavalent and MMR vaccines, along with the average anti-vax sentiment on Twitter between 2013 and 2018. Since 2012, there has been a progressive decline in the coverage of both vaccinations. Coverage rates began to increase in 2015, when i) the appeal court reversed the sentence which established a vaccine/autism link, and ii) there was a sharp increase in measles cases in Italy. In 2017, the expansion and legal enforcement of mandatory vaccines under the new law led to an increase in MMR coverage.

It is important to note that the dynamics of the average Twitter anti-vax sentiment do not necessarily reflect immediately in lower vaccination rates. Vaccines serve as insurance against diseases, and individuals engage in optimal behavior based on their perception of risk, which can be influenced by cognitive biases and local

epidemiology. As a result, the correlation between coverage rates and anti-vax stances may be distorted due to simultaneity and omitted variables.

Echo chambers formation. We rationalize the evolution of anti-vax views on social media in Italy through the lens of a model of social networks opinion dynamics (Baumann et al. (2020) - the complete representation of our model is described in Appendix B). In the model, the combination of exogenous and endogenous drivers of attention and interactions can lead to the radicalization of opinions among users when the level of controversy around a topic increases. In addition, users with polarized views tend to form links and cluster in echo chambers, which are more responsive to further exogenous shocks and can lead to longer peaks of activity and even more extreme positions. All these elements characterize the endogeneity that affects the relationship between the spread of anti-vax opinions on social media and vaccine hesitancy.

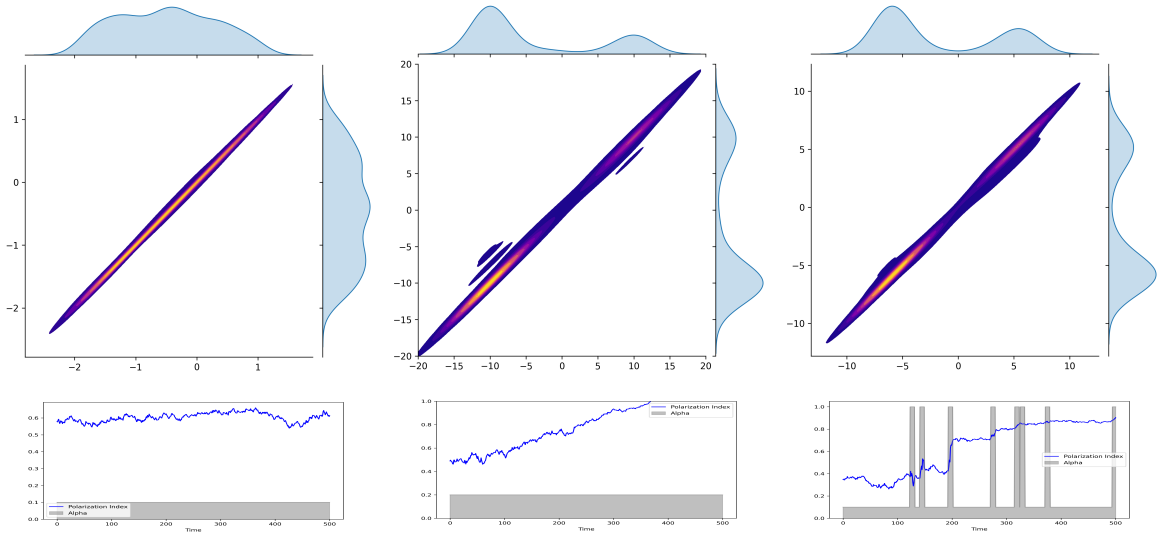
The model highlights two channels through which the users are more likely to influence each other on a controversial topic. On the one hand, exposure to views different from their own sways users' stances, and more "distant" views have a stronger influence (*exposure effect*). On the other hand, the controversial nature of the vaccine-related topic endogenously exacerbates the polarization by influencing the process of network formation (*link formation effect*). Importantly, the former channel represents the impact of anti-vax stances expressed on social media on vaccine hesitancy, which is what we aim to measure. If individual exposure to anti-vax content on Twitter were randomly assigned, the task would be relatively simple. However, due to the *link formation effect*, when a topic becomes controversial users endogenously tend to interact with like-minded peers, which affects the network topology and violates the randomness assumption.

In general, the opinion dynamics within the social network are driven by the interactions among agents, where each agent's i stance (s_i), coupled with the degree of controversialness of the topic, influences the others' stances. Importantly, while the influence of individual stances on other users is tuned by the controversialness (we model it as a hyperbolic function), even moderate opinions can capture the beliefs of their peers. Each agent is characterized by her propensity to interact with a certain number of other agents, and the probability of interaction depends on the degree of homophily, which we model through a decreasing function of the distance between opinions (Bessi et al., 2016). Since we are interested in capturing the possible exchange of opinions between users, we assume that links are the medium through which information flows. For example, if user i is linked to user j, user i is exposed to the content produced by user j, and there is a flow of information from node j to node i in the network. The topology of the network reveals the presence of echo chambers when large shares of users are tied to peers with similar views, and are therefore exposed to similar content with a higher probability. In network terms, this translates into a node i with a given stance s_i being more likely to be

connected with nodes with a stance close to s_i.

The model generates different predictions on the converged state (Figure 5) depending on the level and time-varying behavior of the exogenous topic controversialness. When it is permanently low (panel a), the tendency of users to connect with peers who share similar opinions is counterbalanced by the feeble influence that they exert, which eventually converges to a no-polarization state. When the controversialness is permanently high (panel b), it results in the formation of strongly polarized echo chambers. We obtain a very similar result when the degree of topic controversialness varies over time (panel c). Specifically, as in the observed data, we simulated brief periods of high controversy intertwined with long periods of low controversialness: we find that the former have long-lasting effects on the systemic level of polarization because of the link formation effect (all the model details and the simulation results are reported in Appendix B).

Figure 5: Simulated distribution of stances



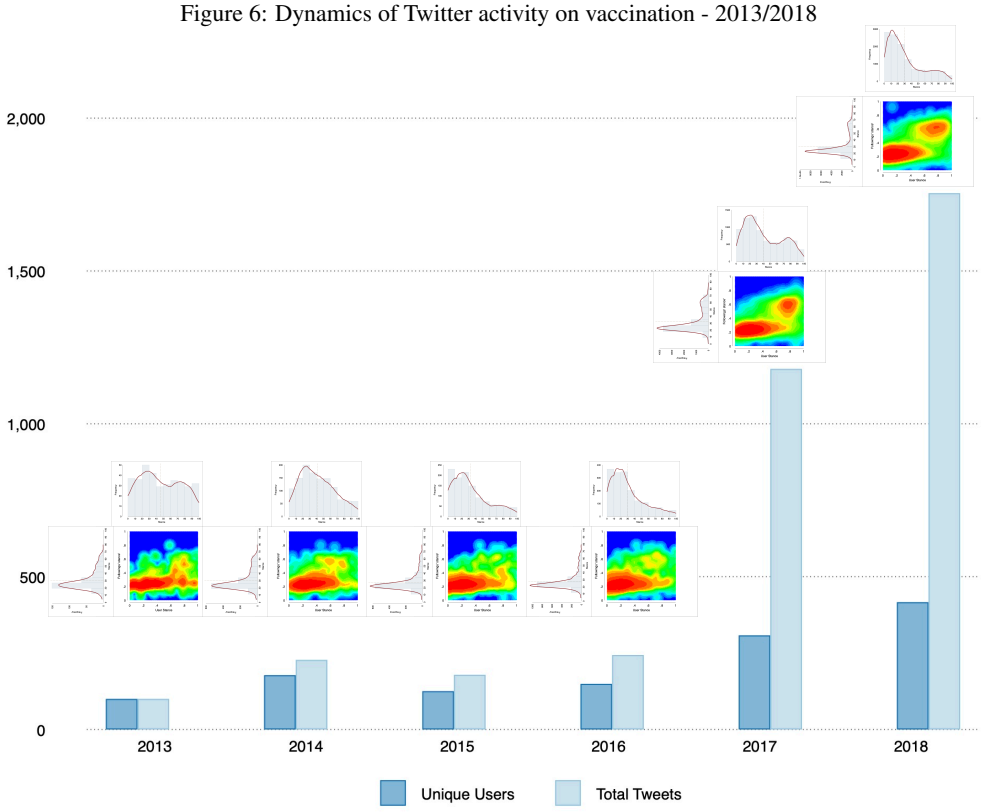
(a) Low controversialness (b) High controversialness (c) Time-varying

Notes: user (x-axis) and average friends' (y-axis) distribution of stances in a simulated model when controversialness is low ($\alpha = .1$ in panel *a*), high ($\alpha = .2$ in panel *b*), and low with short-lived outbursts ($\alpha = 0.1$ and $\alpha = 1$ in panel *c*). In all models, the number of individuals is $N = 500$ and the periods are $T = 5$ - divided in 100 subperiods. Initial values (s_0) are randomly drawn from a gaussian distribution with $\mu = -0.2$ and $\sigma = 0.5$ to match the asymmetry of the initial opinions in the data. The time series report the degree of polarization and the controversialness parameter observed in each subperiod.

In Figure 6, we plot the yearly number of unique users and tweets in our data (2013=100) alongside the joint distribution of users' and their friends' observed stances. In line with Figure 4, between 2013 and 2016 the activity remained steadily low. In 2017, when the vaccination mandate was extended, the number of users and tweets significantly increased. Interestingly, the number of tweets (10×+) increased much more than the corresponding number of users (3×+), suggesting that users already interested in the topic engaged more often in vaccine-related debates (i.e., the topic became more controversial). Accordingly, the heatmaps show the formation of echo chambers. The two opposite clusters suggest an endogenous rise in the radicalization of

opinions among users. In this context, it is likely that the higher controversialness of vaccine-related topics was reinforced by the Twitter amplification algorithm introduced in 2016, which magnifies the exposure to topics that engage users' attention.

Figure 6: Dynamics of Twitter activity on vaccination - 2013/2018



Notes: yearly number of vaccine-related total tweets (grey bars) and unique users (blue bars) between 2013 and 2018 (2013=100). The contour plots report the joint distribution of users' and average friends' stances on vaccination. Colors represent the density of users: the stronger the red hue, the larger the number of agents. The marginal distribution of users' opinions and their friends' are plotted on the x and y-axis, respectively. To construct the figure, we exclude the users with less than 15 friends and 10 tweets/year in the sample to avoid social bots, as their inclusion would artificially generate echo chambers - see, e.g., (Shao et al., 2018).

The evidence of the long-lasting effect of endogenous link formation leading to echo chambers poses a challenge for causal inference. Without adjusting for the systematic tendency towards homophily, naive estimates of the exposure to online anti-vax content on vaccine hesitancy will inevitably be biased. Hence, this set of model predictions thus motivates the use of an IV identification strategy in order to estimate the empirical counterpart of the exposure effect.

5 Empirical strategy

Ideally, we would like to estimate the effect of the exposure to anti-vax content on vaccination rates among children at the individual (parent) level.[23] However, two challenges make this goal difficult to achieve: (i) the endogeneity that characterizes the relationship between exposure and stance, and (ii) the lack of individual data on vaccinations.

First, as discussed in section 4, the homophily and the controversial nature of the vaccine topic lead to the creation of echo chambers. Hence, any correlation in vaccine stances across users may be due to the endogenous choice of friends - for example, user i may form a link with and be exposed to the content produced by user j because they both hold similar views on vaccines. This endogeneity in the formation of social connections poses a challenge to our identification strategy. To address this problem, we use an IV approach. We construct an instrument for the exposure to anti-vax content exploiting the Twitter network structure and the intransitivity of network connections through the local-average model proposed by Bramoullé et al. (2009). Specificaly, user i's "friends of friends" (or *second-degree* network) are not her endogenously-chosen connections, but have an impact on her exposure to vaccine-skeptic content through their interactions with her direct friends - for instance, when a direct connection reacts to a friend-of-friend's post by retweeting or liking it, it will appear in i's feed. To capture this effect in our data, we construct ego networks centered around users who engage in the Twitter vaccine debate. Within these user-specific networks, we measure each user's second-degree exposure to vaccine-skeptic content. To avoid concerns about the potential endogeneity of the influence of indirect friends, we use a rich set of information about the chronology of network creation, which is described in detail in subsection 5.1.

A second challenge is that the data on individual vaccine hesitancy (v_{it}) is typically unavailable and cannot be linked to social media accounts. We therefore use the most granular data currently available on pediatric vaccinations in Italy, which are coverage rates at the municipal/year level. To bridge the mismatch in the level of data aggregation, we link individual Twitter stances on vaccines with municipal-level vaccination rates. To do this, we use a mixed two-stage least squares (M2SLS) strategy, as explained in subsection 5.2. This approach

[23] The ideal model hence would assume the following linear relationship at the individual (parent) level:

$$\boldsymbol{v}_{-it} = \beta s_{it} + X_i + Z_c + \Omega_t + \varepsilon_{it} \tag{2}$$

where $\boldsymbol{v}_{-it}$ reflects vaccine hesitancy of the i^{th} individual's peers at time t, s_{it} is the stance of individual i, X_i are individual characteristics, Z_c are local features and Ω_t is the amount of information available at each point in time, including policy-related interventions (e.g., vaccine mandates), new scientific knowledge, and news related to vaccine-preventable diseases outbreaks. The proposed model assumes a one-to-one mapping between vaccine hesitancy and the observed behavior toward vaccination - i.e., there is a threshold value $v^\star = \mu + \alpha$ above which parents do not vaccinate their children. The parameter of interest β would capture the influence that individual i's stance has on her peers' decision to undertake pediatric vaccinations. In addition, an assumption underlying the above model is that the extent of anti-vaccination persuasion on Twitter is representative of the pressure exerted by vaccine skeptic activists on parents who use other media outlets, both online and offline.

was proposed by Dhrymes and Lleras-Muney (2006) for grouped data.

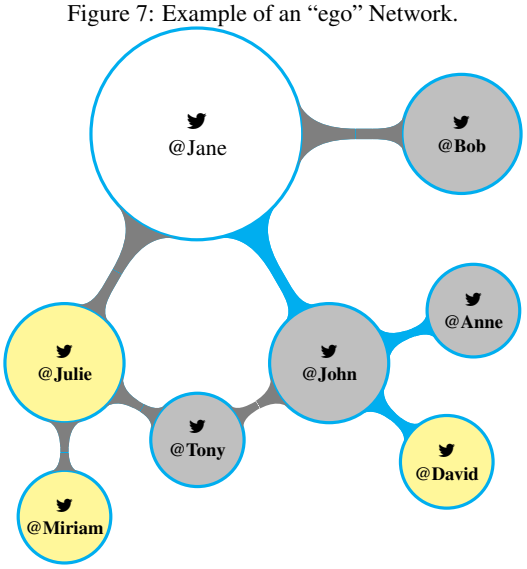
5.1 "Friends of friends" networks

For each user i, we identify two layers of connections: *friends* (lag 1) and *friends-of-friends* (lag 2). The latter constitute incidental connections (i.e., not chosen endogenously) in the directed network that describes each user's social structure. Within the group of friends, we deliberately focus only on those who were already in the direct network of user i before she engaged in any vaccine-related debate on Twitter. This ensures that the link between the user and the friend was established before their involvement in the vaccine debate. In order to further minimize the risk posed by endogenous homophily, among this restricted group of friends we only focus on passive users, i.e. those who do not tweet original content about vaccines, but only react to others' tweets (by liking, retweeting, or replying).

Figure 7 illustrates the selection of friends and friends-of-friends for every "ego" (@Jane in the example). The nodes represent the users: their size depends on the distance from @Jane (lag-1 or lag-2) and the color refers to their engagement level - yellow for active users, gray for passive ones. The edges link each pair of connected users and reflect their endogenous relationships. @Jane lag-1 network includes one active (@Julie) and two passive friends (@John and @Bob). The second-degree network is constituted of four users: one linked to @Julie (@Miriam), two to @John (@David and @Anne), and one to both (@Tony). We build the user-centered network made up of *passive friends* and, through their exclusive connections (e.g., we exclude @Tony because of his connection to an active lag-1 user), we build the measure of indirect and exogenous exposure to anti-vax content of @Jane as described by the blue edges. This instrument reflects the stances of the friends-of-friends of @Jane's passive friends. This way, it overcomes the endogenous link formation effect, provided that the original link between @Jane and e.g. @John was not driven by his friends' stances.

Using our sample of vaccine-related tweets, we identify a total of 65,673,913 friends' nodes. Among these, we identify 8,176,261 unique passive friends. As is typical on social media, we see significant variations in the number of friends and followers: the vast majority of users have only a few friends, while some users are central nodes in the network. The final sample of the second-degree network consists of approximately 2 billion nodes, corresponding to an average of 12,556 friends-of-friends per user. The median user has 469 passive friends and a median of 7,687 friends-of-friends. These connections produced, on average, 142,261 tweets about vaccines (see Table 4).

Figure 7: Example of an "ego" Network.



Notes: The figure plots the architecture of the network on Twitter. The white node (@Jane) is the "ego" user, the gray nodes denote the passive users, in yellow the active ones. The node size depends on the distance from the ego user (lag-1 or lag-2), which is informed by the endogenously-generated links, described by the edges. These are either gray - i.e., passing through an active user in lag-1 or not connected to any lag-2 user - or blue - valid connections to build the measure of indirect exposure.

Table 4: Descriptive statistics of users' networks

	Median	Mean	sd	Min	Max
Friends	469	973.46	2,717.55	1.00	189,433
Friends of friends	7,687	12,556.24	14,078.73	1.00	139,508
Total friends of friends' tweets with vaccine contents	59,535.50	142,261.09	186,460.83	1.00	1,685,355

Notes: The networks refer to 80,471 geotagged unique users who tweeted on vaccines in Italian (2013-2018).

Finally, for each set of friends-of-friends, we compute the average anti-vax stance. This allows us to define each user i's indirect exposure to anti-vax stance in year t as $ffs_{it} = \frac{\sum_{j=1}^{N_{it}} s_{jt}}{N_{it}}$, where N_{it} represents the number of i's friends-of-friends, and the value ranges from 0 to 100. We use this measure to instrument each users' own stance in the IV estimation.

5.2 The Mixed two-stage least squares

In a naive OLS estimation, without taking into account endogeneity, we would measure the impact of online anti-vax skepticism and health outcomes at the municipality level as follows:

$$V_{mt} = \beta \bar{s}_{mt} + T'_{mt}\zeta + C'_{mt}\phi + \gamma_m + \theta_t + \varepsilon_{mt} \tag{3}$$

where V_{mt} is either (one of) the vaccination rates, or the vaccine-preventable hospitalizations/healthcare costs in municipality m in year t; $\bar{s}_{mt}$ is the average vaccine-related stance at municipality/year level; T' represents vectors extracted from the Twitter corpus and the friends' network - i.e. the sum of tweets per municipality/year

and the sum of friends-of-friends tweets per users' municipality/year; C's are socioeconomic characteristics (income per capita at municipality level, birth rate, the share of lower secondary school attainment, the mean age of women at the birth of their first child at province level and health costs per capita at regional level).[24] Additionally, as there might be strong political components to vaccination rates, in C' we include an indicator variable for the rule of *populist* parties at the local level.[25] Several populist parties have raised concerns about vaccine safety (Guriev and Papaioannou, 2022, Kennedy, 2019). We also include municipality and year fixed effects (γ_m and θ_t, respectively). Finally, as public health measures and compliance with these measures might vary at the regional level, we include a set of region-specific time trends $\rho_r \times t$ (region×year).

This simple OLS fixed effects estimation is likely to produce biased estimates due to the crucial sources of endogeneity in our setting. In designing our IV approach, we take advantage of the granular level of detail offered by the Twitter data (i.e., the individual user level) to improve the efficiency of the first stage. However, since the outcome measures are available at the municipality level only, we employ the M2SLS approach, as proposed by Dhrymes and Lleras-Muney (2006). The first stage of the M2SLS, estimated using weighted least squares, is as follows:

First stage - (individual level)

$$s_{it} = \alpha + \beta ffs_{it} + \mathbf{T}'_{it}\zeta + \mathbf{C}'_{mt}\phi + \gamma_m + \rho_r \times t + \theta_t + \varepsilon_{it} \tag{4}$$

where s_{it} is the Twitter stance on vaccines of a user i in year t, and ffs_{it} denotes its indirect exposure to anti-vax content as described in subsection 5.1. Both variables range between 0 and 100, with 100 indicating the maximum level of vaccine skepticism. $\mathbf{T}_{it}$ are Twitter metrics, while $\mathbf{C}_{mt}$ are municipal characteristics. In our setting, there is a one-to-one mapping between the geotagged users and the municipality they reside in or tweet from. Equation (4) allows us to compute the instrumented $\widehat{s}_{it}$, which we then average out at the municipal level to obtain the main regressor for the second stage, which reads:

Second stage - (municipal level)

$$V_{mt} = \alpha + \lambda \overline{\widehat{s}}_{mt} + \overline{\mathbf{T}}'_{mt}\xi + \mathbf{C}'_{mt}\phi + \gamma_m + \rho_r \times t + \theta_t + \eta_{mt} \tag{5}$$

where the outcomes of interest (V_{mt}) are the vaccination rate, the number of vaccine-preventable hospitalizations in the targeted and non-targeted populations, or their total cost. $\overline{\widehat{s}}_{mt}$ is the averaged instrumented

[24] Birth rate, the percentage of people with at least lower secondary school, the mean age of females at first birth, and health costs per capita data come from the Italian National Institute of Statistics. Per-capita income data comes from the Ministry of Economy and Finance. Descriptive statistics are reported in Table A.1 in Appendix subsection A.1

[25] Following Albanese et al. (2022) methodology, parties coded as populist are the Movimento Cinque Stelle (Five Stars Movement) and Lega Nord (Northern League). The data comes from the Ministry of the Interior.

regressor computed in the first stage, weighted by the number of observations in the original cell (number of users at municipality/year level), $\overline{T}_{mt}$ is the average value of Twitter's control variables (T'_{it}), C'_{mt} is the vector of socioeconomic characteristics, γ_m, and θ_t are municipality and year fixed effects that account for time-invariant differences between municipalities and $\rho_r \times t$ (region×year) controls for region-specific trends. All estimates are weighted by municipality population size. We correct the variance-covariance matrix throughout the analysis by bootstrapping the standard errors. In the main specification, the parameter of interest λ captures the causal effect of anti-vax stances on vaccination rate at the municipality level.

6 Results

When presenting the results, we first review the baseline estimates for vaccination rates, distinguished by vaccination type. This allows us to analyze the differential impact of vaccine skepticism on mandatory or recommended vaccines as well as on the MMR, i.e., the vaccine specifically targeted by online disinformation. We then present the results on hospitalizations. We look at the number of hospitalizations for vaccine-preventable diseases and the related costs, rescaled per 100 thousand residents. We also distinguish between hospitalizations for the vaccine-targeted pediatric population versus those for non-target populations of vulnerable individuals (such as newborns, pregnant women, and immunocompromised patients).

To begin, we run a set of regression tests to assess the random assignment of the IV with respect to the contextual features of the user's geolocalized municipality. We do this by regressing the average Twitter stance on vaccines that user i in municipality m is indirectly exposed to through her friends-of-friends stances (ffs_{it}) on municipality characteristics such as income per capita, birth rates, public healthcare expenditure per capita, and education attainment. The identifying assumption requires that the variation in friends-of-friends stances is unrelated to the variation in these predetermined characteristics of municipalities - after controlling for municipality and year fixed effects. Table 5 provides these balance tests, showing that none of the estimated correlations are significantly different from zero.

Table 5: Instrument balance tests

	(1) Health public cost per capita (€)	(2) Income per capita (€)	(3) Lower secondary school att. (%)	(4) Avg. mother's age at birth	(5) Birth rate	(6) Populist party
Panel a: geolocated in the same user's municipality						
ffs_{it}	-0.0211	-0.403	0.0001	0.0001	-0.0002	0.0002
	[0.0246]	[0.442]	[0.0002]	[0.0001]	[0.0002]	[0.0002]
N	110,639	110,639	110,639	110,589	110,639	110,639
Panel b: geolocated in municipalities different from the user's municipality						
ffs_{it}	-0.0001	-0.447	-0.0001	-0.0001	-0.00002	0.0001
	[0.0126]	[0.337]	[0.0004]	[0.0001]	[0.0001]	[0.0001]
N	131,003	131,003	131,003	130,817	131,003	131,003
Panel c: not geolocated						
ffs_{it}	0.0037	1.001	-0.00004	-0.00001	0.0001	0.0002
	[0.0121]	[0.912]	[0.0002]	[0.00003]	[0.0001]	[0.0002]
N	130,977	130,977	130,977	130,791	130,977	130,977
CITY and YEAR FE	✓	✓	✓	✓	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
$Notes$: Regression of indirect exposure to anti-vax stance on municipal characteristics. Standard errors (in brackets) are clustered at the municipality level.

The M2SLS first stage results, shown in Table 6, suggest that the exposure to the vaccine-related stances of a user's friends-of-friends network is a strong predictor of users' own stance.[26] A one-unit increase in the anti-vaccination stance on the 0-100 scale leads to a 0.7-unit increase in the individual's vaccine-related stance, indicating that indirect exposure to anti-vaccination stances can lead users to engage in anti-vaccination activism.

Table 6: M2SLS Individual - First stage.

	(1) s_{it} (30.31)	(2) s_{it} (30.31)	(3) s_{it} (30.31)	(4) s_{it} (30.31)	(5) s_{it} (30.31)	(6) s_{it} (30.31)
ffs_{it} (28.77)	0.799***	0.751***	0.703***	0.703***	0.704***	0.704***
	[0.021]	[0.021]	[0.017]	[0.017]	[0.017]	[0.017]
N	127,754	127,754	127,754	127,754	127,754	127,754
CONTROL (Twitter)				✓		✓
CONTROL (socioeconomics)					✓	✓
YEAR FE	✓	✓	✓	✓	✓	✓
CITY FE		✓	✓	✓	✓	✓
Reg × Year			✓	✓	✓	✓
F-stat	1,501.16	1,288.96	1,765.22	1,763.52	1,755.84	1,757.86

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
$Notes$: The numbers refer to the sample of 830,253 tweets and to a population of 80,471 unique users across 4,220 municipalities. All estimates include municipal and year fixed effects as well as region specific time trends. Standard errors (in brackets) are clustered at the municipality level. The average values of s_{it} and ffs_{it} in parentheses are weighted by population size.

[26] Among the geolocalized tweets, 1% has an average of 1 user only tweeting about vaccines in a year. In the baseline analysis (Table 6), we drop the first percentile of municipalities. We test the results obtained on the full sample in Appendix A, Table A.9.

6.1 Vaccination rates

Table 7 reports the baseline IV results alongside those of the naive OLS model that does not account for endogeneity. Given that hexavalent and MMR vaccines are almost always administered through a single shot, the disease-specific vaccination rates are identical and we report the pooled figure for both. Their average coverage rates are given in parentheses, and the table reports the most demanding specifications, including all controls and fixed effects.

The coefficient estimated for mandatory vaccines (hexavalent) is not statistically distinguishable from zero, and there is no detectable difference between the OLS and the M2SLS approaches.Similarly, we estimate no effect of anti-vax stance on vaccination rate for the recommended vaccines against meningococcal and pneumococcal - in both specifications. On the other hand, when we look at the vaccine most targeted by the fake news, the MMR shot, we find i) a significant effect on coverage rates, and ii) a sizeable difference with respect to the OLS specification. We find that a 10 percentage point increase in the municipality-level anti-vaccination stance leads to a 0.43 percentage point decrease in the MMR coverage rate..[27]

Table 7: Results of the OLS and the Second stage of the M2SLS - Vaccination rates

	(1) OLS V_{mt}	(2) M2SLS V_{mt}
Panel a: Hexavalent (94.06)		
s_{mt}	-0.001	-0.023
	[0.002]	[0.015]
N	7,239	7,239
Panel b: MMR (89.53)		
s_{mt}	-0.005	-0.043**
	[0.003]	[0.021]
N	7,238	7,238
Panel c: Menigococcal (81.32)		
s_{mt}	-0.002	-0.040
	[0.008]	[0.054]
N	7,061	7,061
Panel d: Pneumococcal (82.64)		
s_{mt}	-0.0001	-0.029
	[0.008]	[0.052]
N	7,066	7,066
Controls (Twitter)	✓	✓
Controls (socioeconomics)	✓	✓
City and year FE	✓	✓
Reg × year	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
Notes: All estimates include city and year fixed effects as well as region specific time trends. Standard errors (in brackets) are clustered at the municipality level and have been corrected in the second stage. Estimates as well as averages of V_{mt} are weighted by the municipality population size.

[27] Table A.2 in Appendix A reports the reduced form estimates. Table A.4 in Appendix A reports the full set of estimates for the different models as specified in Table 6.

6.2 Hospitalizations

We also estimate the effect on hospitalizations due to vaccine-preventable conditions. We distinguish between two groups: the target pediatric population and non-target vulnerable individuals, as this distinction matters from a policy perspective. In fact, the number of hospitalizations for vaccine-preventable diseases among non-targeted patients measures the extent of low vaccination rates' negative spillovers on local communities. Quantifying the negative externalities of individuals opting out of immunization provides an objective argument in the policy debate on vaccine mandates that must be taken into consideration.

Table 8: Results of the OLS and the Second stage of the M2SLS - Hospitalizations .

	(1) OLS V_{mt} non-target pop.	(2) M2SLS V_{mt} non-target pop.	(3) OLS V_{mt} non-target pop.(MMR)	(4) M2SLS V_{mt} non-target pop.(MMR)	(5) OLS V_{mt} Children age 1-10 (MMR)	(6) M2SLS V_{mt} Children age 1-10 (MMR)
Panel a: Hopitalizations						
s_{mt}	0.0211	0.213*	0.018**	0.234***	0.007	0.145**
	[0.0159]	[0.113]	[0.00841]	[0.0601]	[0.008]	[0.065]
Panel b: Healthcare costs						
s_{mt}	129.8*	731.1**	71.96**	722.1***	47.13*	366.9**
	[66.39]	[353.8]	[30.92]	[243.1]	[25.95]	[161.1]
N	3,331	3,331	3,331	3,331	3,331	3,331
Controls (Twitter)	✓	✓	✓	✓	✓	✓
Controls (socioec.)	✓	✓	✓	✓	✓	✓
City and year FE	✓	✓	✓	✓	✓	✓
Reg $\times$ year	✓	✓	✓	✓	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
$Notes$: All estimates include city and year fixed effects as well as region specific time trends. Standard errors (in brackets) are clustered at the municipality level and have been corrected in the second stage. Estimates are weighted by the municipality population size.

In Table 8, we estimate the effect on the number of hospitalizations and the average annual cost for the two populations, expressed per 100 thousand residents. For vulnerable individuals (the non-target population), we find that a 1 percentage point increase in the municipality-level anti-vaccination stance leads to an additional 0.21 hospitalizations per 100 thousand residents (the baseline average being 22.21). This is also expressed in terms of excess healthcare expenditure of 731.1 euros, representing a 1.1% increase relative to the baseline. Specifically, in terms of hospitalizations due to MMR, the same increase in vaccine skepticism is associated with an additional 0.23 hospitalizations per 100 thousand residents (the baseline average being 4.99) and an additional expenditure of 722.1 euros, corresponding to a 4.6% increase. When looking at hospitalizations among the target pediatric population, our estimates (column 5) suggest that a 1 percentage point increase in the municipality-level anti-vaccination stance leads to an additional 0.145 hospitalizations per 100 thousand residents (the baseline average being 2.96) and an excess expenditure of 366.9 euros, corresponding to a 7.7%

increase.[28]

In line with the baseline results, Table A.5 in subsection A.1 shows no significant results for the non-target population and target pediatric population hospitalized for diseases preventable by hexavalent, meningococcus and pneumococcus vaccines, respectively.

To evaluate the efficacy of vaccinations in reducing the probability of mortality from vaccine-preventable diseases, Table A.7 in Appendix A shows the result of our estimates on the mortality cases of hospitalized patients with these diseases.

6.3 Robustness checks

We first check the robustness of our results to three potential confounders - either in the first stage (the introduction of a homophily-enhancing algorithm on Twitter) or in the second stage (pre-existing vaccine mandates or the influence of strong populist parties). Finally, we propose a reweighting of our estimates to account for the number of second-degree links between each user and her friends-of-friends network.

First, in 2016, Twitter introduced an algorithmic timeline that rearranges users' feed based on relevance rankings. This likely amplifies the impact of indirect exposure on user stance formation. To account for this change, we interact the instrumental variable (ffs_{it}) and a dummy variable ($TWalg$) that takes on a value of 1 from 2016 to account for the shift in the algorithm. Second, we use a similar approach to control for the existence of a vaccine mandate in a single region - Emilia-Romagna - since November 25th, 2016 (Regional Law n.19), which followed several outbreaks of infectious diseases affecting non-vaccinated individuals (Gori et al., 2020). The mandate required the vaccination to enroll in public school and kindergartens. We control for this through an interaction term between ffs_{it} and an indicator variable (ER) that takes on a value of 1 for individuals in Emilia-Romagna after the implementation of the regional law. Third, Italian populist parties have sometimes raised concerns about vaccine safety (Guriev and Papaioannou, 2022, Kennedy, 2019) - hence, our estimates could be capturing a differential effect of political stances rather than disinformation spread. We control for this potential confounder by interacting ffs_{it} with an indicator variable (PP) that takes on a value of 1 whenever there was a populist party ruling at the municipal level.

Table 9 reports the first stage results of the above exercises alongside the baseline model (column 1). While we find a significant impact of the introduction of the algorithm (column 2), neither the pre-existing mandate (column 3) nor the influence of populist parties (column 4) seem to play a significant role in affecting users' stance through the indirect exposure.

[28] Table A.3 in Appendix A reports the reduced form estimates.

Table 9: M2SLS Individual - First stage.

	(1) Main	(2) Twitter algorithm	(3) Emilia Romagna Law	(4) Populist party	(5) Network distance	(6) Excluding FoF geolocated in user's municipality
	s_{it}	s_{it}	s_{it}	s_{it}	s_{it}	s_{it}
	(30.33)	(30.33)	(30.33)	(30.33)	(30.33)	(30.33)
ffs_{it}	0.704***	0.528***	0.706***	0.691***	0.611***	0.731***
	[0.017]	[0.035]	[0.017]	[0.022]	[0.021]	[0.016]
$ffs_{it}\times$ TWalg		0.251***				
		[0.039]				
$ffs_{it}\times$ ER			0.005			
			[0.0742]			
$ffs_{it}\times$ PP				0.048		
				[0.043]		
N	127,754	127,754	127,754	127,754	127,754	127,746
Controls (Twitter)	✓	✓	✓	✓	✓	✓
Controls (socioec.)	✓	✓	✓	✓	✓	✓
City and year FE	✓	✓	✓	✓	✓	✓
Reg $\times$ year	✓	✓	✓	✓	✓	✓
F-stat	1,757.86	998.690	870.815	943.98	875.82	2102.95

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

$Notes$: The numbers refer to an initial sample of 830,253 tweets to a population of 80,471 unique users across 4220 municipalities. All estimates include city, region and year fixed effects and region-specific time trends fixed effects. Standard errors (in brackets) are clustered at the municipality level. Averages of s_{it} in parentheses are weighted by population size.

Additionally, in order to address any remaining concern about the exogeneity of the friends-of-friends network, we propose two alternative estimation strategies. Firstly, the network has a hierarchical structure with "ego" users, their passive friends, and the relative friends-of-friends connections. If a friend-of-friend is linked to the ego user through multiple passive lag-1 friends, this can weaken the intransitivity assumption underlying the validity of the IV. To account for this, we reweight the estimates with the inverse of the number of connections that a friend-of-friend shares with the ego user with the following equation:

$$w_i = \frac{1}{\sum_{j=1}^{n} f_{ij}} \tag{6}$$

where f_{ij} is the number of shared nodes between user i and each friend-of-friend j. The weight can be regarded as a measure of how long information will take to spread in the network. The first stage results (column 5) show a slightly decreased coefficient estimate, which however remains comparable to the original one in terms of both magnitude and statistical significance.

Secondly, we exclude all friends-of-friends geolocated in the user's municipality, in order to rule out the possibility that the network might be influenced by common local offline vaccine views. The respective first stage results (column 6) remain virtually unchanged.

Table 10: M2SLS Municipal - Second stage (Vaccination rate, hospitalizations and healthcare costs).

	(1) Main V_{mt}	(2) Twitter algorithm V_{mt}	(3) Emilia Romagna Law V_{mt}	(4) Populist Party Law V_{mt}	(5) Network distance V_{mt}	(6) Excluding FoF geolocated in the user's municipality V_{mt}
Panel a: MMR vaccination rate (89.53)						
s_{mt}	-0.043**	-0.047**	-0.048**	-0.055**	-0.050**	-0.042**
	[0.021]	[0.022]	[0.021]	[0.026]	[0.023]	[0.021]
N	7,238	7,238	7,238	7,238	7,238	7,238
Panel b: Non-target population						
Hopitalizations						
s_{mt}	0.213*	0.231*	0.204*	0.215*	0.220*	0.205*
	[0.113]	[0.121]	[0.112]	[0.112]	[0.115]	[0.108]
Healthcare costs						
s_{mt}	731.1**	821.3**	712.8**	746.5*	794.0**	909.9**
	[409.8]	[434.7]	[406.6]	[412.2]	[411.0]	[402.0]
N	3,331	3,331	3,331	3,331	3,331	3,331
Panel c: Non-target population (MMR)						
Hopitalizations						
s_{mt}	0.234***	0.256***	0.233***	0.231***	0.242***	0.211***
	[0.0601]	[0.0675]	[0.0596]	[0.0603]	[0.0621]	[0.0578]
Healthcare costs						
s_{mt}	722.1***	716.7***	725.1***	734.0***	743.7***	713.8***
	[243.1]	[250.6]	[242.8]	[247.7]	[247.1]	[235.6]
N	3,331	3,331	3,331	3,331	3,331	3,331
Panel d: Children age 1-10 (MMR)						
Hopitalizations						
s_{mt}	0.145**	0.150**	0.145**	0.146**	0.142**	0.115*
	[0.0650]	[0.0664]	[0.0651]	[0.0653]	[0.0659]	[0.0619]
Healthcare costs						
s_{mt}	366.9**	428.7**	366.5**	363.6**	390.2**	375.5**
	[161.1]	[171.8]	[160.9]	[163.9]	[163.7]	[162.3]
N	3,331	3,331	3,331	3,331	3,331	3,331
Controls (Twitter)	✓	✓	✓	✓	✓	✓
Controls (socioec.)	✓	✓	✓	✓	✓	✓
City and year FE	✓	✓	✓	✓	✓	✓
Reg × year	✓	✓	✓	✓	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Notes: All estimates include city and year fixed effects and region-specific time trends. Standard errors (in brackets) are clustered at the municipality level and have been corrected in the second stage. Estimates, as well as averages of V_{mt}, are weighted by the municipality population size.

Table 10 reports the second stage results for all the checks relative to the MMR vaccination rates (panel a), the hospitalization rates and costs for non-target population (panel b), the specific MMR non-target population (panel c) and for childen aged 1 to 10 (panel d).[29] All estimates are qualitatively and quantitatively in line with the baseline's.

[29] Table A.8 in subsection A.1 reports the (null) results on all vaccination types.

6.4 Non-linear effects and policy implications

To explore the potential policy implications of our results, we investigate whether there is any non-linearity in the effect of indirect exposure on user stances. Specifically, we look at whether the influence channeled through the friends-of-friends network varies depending on where a user falls in the stance distribution (i.e., whether they are vaccine supporters or skeptics).

Hence, we first re-run our main model specification while classifying user stances into two binary categories: pro-vax users (those with an average anti-vax stance of zero), and anti-vax users (those with an average anti-vax stance of 100). This allows us to better understand the factors that influence vaccine attitudes among these two sub-groups.

Table 11: M2SLS for pro-vax vs. anti-vax users - First stage.

	(1)	(2)
	Pro_{it}	$Anti_{it}$
	(0.495)	(0.204)
ffs_{it} (28.77)	-0.0076 ***	0 .0046***
	[0 .0003]	[0.0001]
N	127,754	127,754
Controls (Twitter)	✓	✓
Controls (Socioec.)	✓	✓
City and year FE	✓	✓
Reg $\times$ year	✓	✓
F-stat	1,765.22	1,763.52

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
$Notes$: The numbers refer to an initial sample of 830,253 tweets to a population of 80,471 unique users across 4220 municipalities. All estimates include city, region and year fixed effects and region specific time trends fixed effect. Standard errors (in brackets) are clustered on municipalities level. Mean values of Pro_{it}, $Anti_{it}$ and $\overline{ffs}_{it}$ in parentheses are weighted by population size.

According to the magnitude of the coefficient estimates presented in Table 11, the exposure to friends-of-friends stances has a stronger effect on pro-vax users compared to anti-vax users. Hence, each unit change in the exposure stance is more likely to increase hesitancy among pro-vax users rather than reduce it among anti-vax users.

Table 12: Results of the Second stage of the M2SLS for pro-vax vs. anti-vax users - Vaccination rates

	(1) M2SLS Pro_{mt} V_{mt}	(2) M2SLS $Anti_{mt}$ V_{mt}
Panel a: Hexavalent (94.06)		
	0.4567	0.0674
	[1.4333]	[2.1973]
N	7,239	7,239
Panel b: MMR (89.53)		
	3.9086*	-6.6162*
	[2.1978]	[3.5315]
N	7,238	7,238
Panel c: Menigococcal (81.32)		
	0.5034	-1.6496
	[4.8856]	[8.2071]
N	7,061	7,061
Panel d: Pneumococcal (82.64)		
	2.7584	-4.2443
	[5.3633]	[8.4350]
N	7,066	7,066
Controls (Twitter)	✓	✓
Controls (Socioec.)	✓	✓
City and year FE	✓	✓
Reg × year	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
Notes: All estimates include city and year fixed effects as well as region specific time trends. Standard errors (in brackets) are clustered at the municipality level. Estimates as well as averages of V_{mt} are weighted by the municipality population size.

In fact, the second stage results (Table 12) confirm that the effect on vaccine coverage is more strongly channeled through a shift of users towards anti-vax stances, rather than pro-vax ones. In turn, this suggests that policy interventions aimed at discouraging vaccine hesitancy should be targeted toward reducing the exposure to and the flow of anti-vax content, rather than increase pro-vax campaigns. However, social media censorship - although effective - has a number of political, social and ethical implications that go beyond the debate around vaccinations.[30] In addition, recent contributions have shown that these measures can backfire, leading to larger spread of censored information (Hobbs and Roberts, 2018).

We also consider the role of random events related to epidemics, scientific discoveries, court sentences, policies, and news in mitigating or reinforcing the influence of exposure on user stances. To do that, we hand-collect all of the significant events related to vaccines that were discussed in the media during the period of our analysis. These topics include issues such as deaths of children allegedly caused by vaccines or lack of vaccination, court rulings in favor of anti-vax or pro-vax views, the dissemination of scientific evidence for or against vaccines, and political debates about pro- and anti-vax stances. Following Athey et al. (2022), we

[30]Twitter acts on complaints by third parties - including governments - to remove illegal content from the platform. In addition, it runs its own content moderation policy, which includes actions like user suspension, content removal and permanent bans in response to violations of the terms of use (https://help.twitter.com/en/rules-and-policies/twitter-rules). Current allegations against Twitter policies include partisan implementation of moderation rules and arbitrary or politically biased use of bans.

manually classify these online debates into four broad domains: vaccine efficacy, statements from trustful sources, politics and mandates, and allegations that vaccines are unsafe.[31]

Table 13: User exposure to friends-of-friends stances and the role of online debates topics.

	(1) s_{it} (30.31)	(2) Pro_{it} (0.495)	(2) $Anti_{it}$ (0.204)
ffs_{it}	0.2884***	-0.3309***	0.2295***
	[0.0693]	[0.0757]	[0.0728]
$ffs_{it} \times Efficacy$	-0.3425	0.3765	-0.3548
	[0.2724]	[0.2754]	[0.2961]
$ffs_{it} \times Trustful\,Source$	-0.3136***	0.2656**	-0.3805***
	[0.0992]	[0.1127]	[0.1057]
$ffs_{it} \times Politics\,and\,Mandate$	-0.1749***	0.0660	-0.3899***
	[0.0530]	[0.0408]	[0.0589]
$ffs_{it} \times Vaccines\,Unsafe$	-0.0697	0.1369	-0.0387
	[0.2292]	[0.2442]	[0.2495]
N	531,352	531,352	531,352
User FE	✓	✓	✓
Date FE	✓	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

$Notes$: The numbers refer to an initial sample of 830,253 tweets to a population of 80,471 unique users across 4220 municipalities. All estimates include individual and daily date fixed effects. Standard errors (in brackets) are clustered at the individual. Mean values of s_{it}, Pro_{it}, and $Anti_{it}$ in parentheses are weighted by population size.

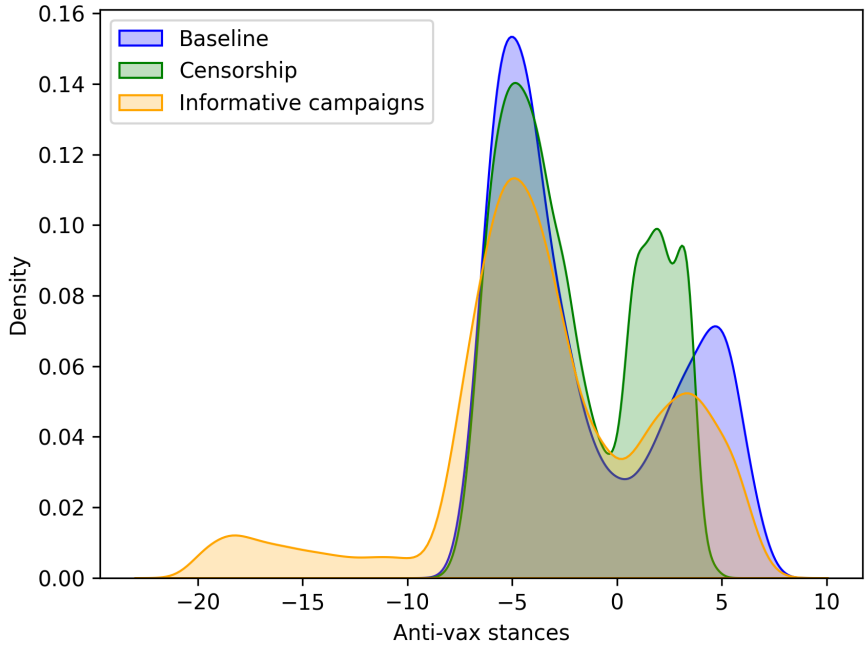
Table 13 shows estimates of daily-level user stances on vaccines, conditional on user and daily date fixed effects. These estimates show how individual stances fluctuate as a function of their friends-of-friends stances on regular days, and on days when specific events related to vaccines are debated on Twitter. Column 1 shows that, after controlling for individual fixed tendencies and day-specific features of Twitter activity, individual stances tend to evolve in response to the effect of their friends-of-friends stances - in line with our first stage baseline result. Exposure to anti-vax content tends to make individuals more lenient towards such stances. However, this relationship is mitigated (actually reversed) on days when a statement in favor of vaccines is issued by a trustworthy source, such as the World Health Organization, the academic or research community, the European Commission, or a court. A similar pattern is observed on days when political debates about the usefulness of vaccines are discussed on Twitter. When we classify user stances into two binary categories (pro-vax and anti-vax), we find that the effect of exposure to anti-vax content is mitigated to a greater extent in the anti-vax category (column 3). Events related to statements from trustworthy sources and political debates are generally able to offset the influence of exposure to anti-vax stances (or reinforce the influence of exposure to pro-vax content).

We make use of our sketched model to get a sense of the effects of the possible interventions on social media platforms. Based on the above results, we run two types of counterfactual exercises - symmetrical in

[31] The full list of events are reported in Table A.10 in Appendix A.

implementation, but rather different in interpretation. On the one hand, we look at the censorship effect on anti-vax stances by exogenously halving the activity rate of users whose stance exceeds the 90^{th} percentile (*Censorship*); on the other hand, we investigate the effects of broadening the reach of pro-vax activists by doubling the number of contacted users for pro-vax activists in the first decile of the distribution (*Infomative campaigns*).[32] The results of the exercises are reported in Figure 8, where we plot the converged density distribution of users' stances for the baseline exercise - panel (c) of Figure 5 - as well as for the Censorship and for the Informative campaign counterfactuals. The figure highlights two main facts - in line with our estimates: first, both interventions reduce the peak and the number of anti-vax users; second, the Informative campaigns, much more than the censorship intervention, leads to a decrease in the overall polarization. We stress that the different effects depend on the asymmetry that characterizes the starting distribution which, nonetheless, matches the one observed in the actual data.

Figure 8: Policy counterfactuals - Monte Carlo results



Notes: converged density distributions of users' stances (N=T=500) - average over 100 Monte Carlo runs. We report the baseline model (blue), the "Censorship" counterfactual exercise (green) in which we halve the activity rate of users in the upper decile of the stance distribution, and the "Informative campaigns" counterfactual exercise (orange) in which we double the activity rate for users in the first decile of the stance distribution.

In fact, informative campaigns about vaccines may be an effective and scalable intervention for shaping public health awareness, the more so when they are perceived as coming from trustful sources, and that they

[32] With these two exercises we implement two different policies: first, we make the platform flag and reduce the visibility of tweets based on their content; second, we simulate an informative pro-vax campaign - possibly sponsored by the government or other public entities - that by definition allows for increased reach.

are backed up by political interventions.

7 Conclusions

Between 2013 and 2018, pediatric vaccine coverage rates in Italy have undergone significant changes, partially due to the spread of misinformation regarding the safety of MMR vaccines. The vaccine hesitancy has contributed to outbreaks of several infectious diseases, leading to the expansion and legal enforcement of a mandate for a large number of pediatric vaccines in 2017. With the availability of a large amount of data on online interactions and a state-of-the-art NLP algorithm, we leverage this setting to investigate, in quantitative terms, the real costs that dis- and misinformation impose on society. The negative consequences of fake or unsubstantiated news spread has been largely discussed in academic fora, political circles and, during the COVID-19 crisis, in the mainstream media. Nonetheless, while it is known that the clusters of conspiracy are the breeding ground for fake news, and that online activities, especially on social media, can have nefarious effects such as hate crimes (Müller and Schwarz, 2021) or influence election outcomes (Fujiwara et al., 2021), this paper contributes to the debate by i) providing a method to estimate the causal effects of online interactions at the individual level on observable, aggregate outcomes; ii) estimating the actual costs imposed on healthcare systems by anti-vax online activity; and iii) proposing data- and simulation-driven insights on how to tame the spread of anti-scientific, or more generally, unsubstantiated content on social media. Regarding the last point, we show that pro-vaccine users tend to be more influenced by exposure to vaccine-related content than their anti-vax counterpart. Conversely, the latter seem to be effectively responsive to statements from trustworthy sources. Both elements suggest that informative campaigns about vaccines, online and offline, may be effective in contrasting vaccine hesitancy. Indeed, even if vaccine skeptics hardly change their views, such campaigns could counteract the persuasive effect of anti-vax content on pro-vaccine individuals.

In conclusion, we see our policy insights as a viable approach to contrast the decrease in vaccine coverage without resorting to coercitive measures such as vaccine mandates. Our findings suggest that while the legal enforcement may address the immediate effects of vaccine hesitancy on coverage rates and associated health costs, it also leads to polarization and radicalization of opinions, which are long-lasting and, when coupled with echo chambers, endogenously self-generating. Hence, policymakers should take these potential consequences into account, in order to avoid that vaccine-enhancing measures backfire once the legal enforcement is withdrawn. Baumann et al. (2021) suggest that when debated topics overlap thematically, increases in controversialness can lead to the emergence of ideological states where multiple stances align within a common, "political" stance. In their model, ideology emerges endogenously from uncorrelated polarization, achieved by

relaxing the unrealistic assumption of topic orthogonality. In this paper's analysis of pediatric vaccines from 2013 to 2018, fake news related to vaccinations was limited to the debate on the vaccine-autism causation. However, today the topic is no longer uncorrelated to other salient debates. The controversy surrounding the COVID-19 pandemic has created an ideological state that covers a wide range of topics including vaccines, face masks, mobility restrictions, and ultimately political opinions. Finding a way to deescalate the debates around scientifically grounded topics can prove to be a viable way to reduce the polarization and foster constructive discussions.

References

Abrevaya, J. and K. Mulligan (2011). Effectiveness of state-level vaccination mandates: evidence from the varicella vaccine. *Journal of health economics 30*(5), 966–976.

Acemoglu, D., A. Ozdaglar, and J. Siderius (2021). A model of online misinformation. Technical report, National Bureau of Economic Research.

Alatas, V., A. G. Chandrasekhar, M. Mobius, B. A. Olken, and C. Paladines (2019). When celebrities speak: A nationwide twitter experiment promoting vaccination in indonesia. Technical report, National Bureau of Economic Research.

Albanese, G., G. Barone, and G. de Blasio (2022). Populist voting and losers' discontent: Does redistribution matter? *European Economic Review 141*, 104000.

Allam, A., P. J. Schulz, and K. Nakamoto (2014). The impact of search engine selection and sorting criteria on vaccination beliefs and attitudes: two experiments manipulating google output. *Journal of medical internet research 16*(4), e100.

Allcott, H. and M. Gentzkow (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives 31*(2), 211–36.

Allcott, H., M. Gentzkow, and C. Yu (2019). Trends in the diffusion of misinformation on social media. *Research & Politics 6*(2), 2053168019848554.

Athey, S., K. Grabarz, M. Luca, and N. C. Wernerfelt (2022). The effectiveness of digital interventions on covid-19 attitudes and beliefs. Technical report, National Bureau of Economic Research.

Azzimonti, M. and M. Fernandes (2022). Social media networks, fake news, and polarization. *European Journal of Political Economy*, 102256.

Bailey, M., D. M. Johnston, M. Koenen, T. Kuchler, D. Russel, and J. Stroebel (2020). Social networks shape beliefs and behavior: Evidence from social distancing during the covid-19 pandemic. Technical report, National Bureau of Economic Research.

Baumann, F., P. Lorenz-Spreen, I. M. Sokolov, and M. Starnini (2020). Modeling echo chambers and polarization dynamics in social networks. *Physical Review Letters 124*(4), 048301.

Baumann, F., P. Lorenz-Spreen, I. M. Sokolov, and M. Starnini (2021). Emergence of polarized ideological opinions in multidimensional topic spaces. *Physical Review X 11*(1), 011012.

Berinsky, A. J. (2017). Rumors and health care reform: Experiments in political misinformation. *British Journal of Political Science 47*(2), 241–262.

Bessi, A., F. Petroni, M. D. Vicario, F. Zollo, A. Anagnostopoulos, A. Scala, G. Caldarelli, and W. Quattrociocchi (2016). Homophily and polarization in the age of misinformation. *The European Physical Journal Special Topics 225*(10), 2047–2059.

Bramoullé, Y., H. Djebbari, and B. Fortin (2009). Identification of peer effects through social networks. *Journal of econometrics 150*(1), 41–55.

Breza, E., F. C. Stanford, M. Alsan, B. Alsan, A. Banerjee, A. G. Chandrasekhar, S. Eichmeyer, T. Glushko, P. Goldsmith-Pinkham, K. Holland, et al. (2021). Doctors' and nurses' social media ads reduced holiday travel and covid-19 infections: A cluster randomized controlled trial. Technical report, National Bureau of Economic Research.

Burki, T. (2019). Vaccine misinformation and social media. *The Lancet Digital Health 1*(6), e258–e259.

Carpenter, C. S. and E. C. Lawler (2019). Direct and spillover effects of middle school vaccination requirements. *American Economic Journal: Economic Policy 11*(1), 95–125.

Carrieri, V., L. Madio, and F. Principe (2019). Vaccine hesitancy and (fake) news: Quasi-experimental evidence from italy. *Health economics*.

Chiou, L. and C. Tucker (2018). Fake news and advertising on social media: A study of the anti-vaccination movement. Technical report, National Bureau of Economic Research.

Cinelli, M., G. De Francisci Morales, A. Galeazzi, W. Quattrociocchi, and M. Starnini (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences 118*(9), e2023301118.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2018a). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2018b). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dhrymes, P. J. and A. Lleras-Muney (2006). Estimation of models with grouped and ungrouped data by means of "2sls". *Journal of econometrics 133*(1), 1–29.

Esposito, S., P. Durando, S. Bosis, F. Ansaldi, C. Tagliabue, G. Icardi, E. V. S. Group, et al. (2014). Vaccine-preventable diseases: from paediatric to adult targets. *European journal of internal medicine 25*(3), 203–212.

Flaxman, S., S. Goel, and J. M. Rao (2016). Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly 80*(S1), 298–320.

Fujiwara, T., K. Müller, and C. Schwarz (2021). The effect of social media on elections: Evidence from the united states. Technical report, National Bureau of Economic Research.

Gentzkow, M. and J. M. Shapiro (2011). Ideological segregation online and offline. *The Quarterly Journal of Economics 126*(4), 1799–1839.

Gori, D., C. Costantino, A. Odone, B. Ricci, M. Ialonardi, C. Signorelli, F. Vitale, and M. P. Fantini (2020). The impact of mandatory vaccination law in italy on mmr coverage rates in two of the largest italian regions (emilia-romagna and sicily): an effective strategy to contrast vaccine hesitancy. *Vaccines 8*(1), 57.

Grossman, G., S. Kim, J. M. Rexer, and H. Thirumurthy (2020). Political partisanship influences behavioral responses to governors' recommendations for covid-19 prevention in the united states. *Proceedings of the National Academy of Sciences 117*(39), 24144–24153.

Guriev, S. and E. Papaioannou (2022). The political economy of populism. *Journal of Economic Literature (forthcoming)*..

Hobbs, W. R. and M. E. Roberts (2018). How sudden censorship can increase access to information. *American Political Science Review 112*(3), 621–636.

Holtkamp, N. C. et al. (2021). *The Economic and Health Effects of the United States' Earliest School Vaccination Mandates*. Ph. D. thesis.

Huszár, F., S. I. Ktena, C. O'Brien, L. Belli, A. Schlaikjer, and M. Hardt (2022). Algorithmic amplification of politics on twitter. *Proceedings of the National Academy of Sciences 119*(1), e2025334119.

Jin, Z., Z. Peng, T. Vaidhya, B. Schoelkopf, and R. Mihalcea (2021). Mining the cause of political decision-making from social media: A case study of covid-19 policies across the us states. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 288–301.

Johnson, K. and D. Goldwasser (2016, November). Identifying stance by analyzing political discourse on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, Austin, Texas, pp. 66–75. Association for Computational Linguistics.

Jolley, D. and K. M. Douglas (2014). The effects of anti-vaccine conspiracy theories on vaccination intentions. *PloS one 9*(2), e89177.

Kennedy, J. (2019). Populist politics and vaccine hesitancy in western europe: an analysis of national-level data. *European journal of public health 29*(3), 512–516.

Kim, T. (2022). Measuring police performance: Public attitudes expressed in twitter. In *AEA Papers and Proceedings*, Volume 112, pp. 184–87.

Larsen, B., T. J. Ryan, S. Greene, M. J. Hetherington, R. Maxwell, and S. Tadelis (2022). Counter-stereotypical messaging and partisan cues: moving the needle on vaccines in a polarized us. Technical report, NBER Working Paper, Stanford University, Palo Alto, CA, 2022). https://www. nber ....

Lawler, E. C. (2017). Effectiveness of vaccination recommendations versus mandates: Evidence from the hepatitis a vaccine. *Journal of health economics 52*, 45–62.

Leask, J., S. Chapman, P. Hawe, and M. Burgess (2006). What maintains parental support for vaccination when challenged by anti-vaccination messages? a qualitative study. *Vaccine 24*(49-50), 7238–7245.

Lorenz-Spreen, P., B. M. Mønsted, P. Hövel, and S. Lehmann (2019). Accelerating dynamics of collective attention. *Nature communications 10*(1), 1–9.

Martinez, L. S., S. Hughes, E. R. Walsh-Buhi, and M.-H. Tsou (2018). "okay, we get it. you vape": an analysis of geocoded content, context, and sentiment regarding e-cigarettes on twitter. *Journal of health communication 23*(6), 550–562.

Michaels, D. (2008). *Doubt is their product: how industry's assault on science threatens your health*. Oxford University Press.

Mullainathan, S. and A. Shleifer (2005). The market for news. *American Economic Review 95*(4), 1031–1053.

Müller, K. and C. Schwarz (2021). Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association 19*(4), 2131–2167.

Opel, D. J., J. A. Taylor, R. Mangione-Smith, C. Solomon, C. Zhao, S. Catz, and D. Martin (2011). Validity and reliability of a survey to identify vaccine-hesitant parents. *Vaccine 29*(38), 6598–6605.

Perra, N., B. Gonçalves, R. Pastor-Satorras, and A. Vespignani (2012). Activity driven modeling of time varying networks. *Scientific reports 2*(1), 1–7.

Pierri, F., A. Artoni, and S. Ceri (2020). Investigating italian disinformation spreading on twitter in the context of 2019 european elections. *PloS one 15*(1), e0227821.

Polignano, M., P. Basile, M. De Gemmis, G. Semeraro, and V. Basile (2019). Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets. In *6th Italian Conference on Computational Linguistics, CLiC-it 2019*, Volume 2481, pp. 1–6. CEUR.

See, A., P. Liu, and C. Manning (2017). Get to the point: Summarization with pointer-generator networks. In *Association for Computational Linguistics*.

Shao, C., G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer (2018). The spread of low-credibility content by social bots. *Nature communications 9*(1), 1–9.

Siegal, G., N. Siegal, and R. J. Bonnie (2009). An account of collective actions in public health. *American Journal of Public Health 99*(9), 1583–1587.

Smith, L. E., R. Amlôt, J. Weinman, J. Yiend, and G. J. Rubin (2017). A systematic review of factors affecting vaccine uptake in young children. *Vaccine 35*(45), 6059–6069.

Sunstein, C. R. (2001). *Republic. com*. Princeton university press.

Sunstein, C. R. (2017). *# Republic: Divided democracy in the age of social media.* Princeton: Princeton University Press.

Sunstein, C. R. (2018). *The cost-benefit revolution*. MIT Press.

Vosoughi, S., D. Roy, and S. Aral (2018). The spread of true and false news online. *Science 359*(6380), 1146–1151.

Zhuravskaya, E., M. Petrova, and R. Enikolopov (2020). Political effects of the internet and social media. *Annual Review of Economics 12*(1), 415–438.

Appendix A

A.1 Additional Tables

Table A.1 provides an overview of the statistical data pertaining to the characteristics of the municipality.

Table A.1: Descriptive statistics of municipality's characteristics

	Median	Mean	sd	Min	Max
Avg. mother's age at birth	31.92	31.82	0.31	30.32	32.81
Health public cost pc (€)	1,911.00	1,903.89	56.37	1,662.00	2,515.00
Income pc (€)	9,183.32	10,854.95	3,786.64	1,986.88	84,253.34
Lower secondary school attainment (%)	86.41	85.30	2.22	74.36	87.73
Birth rate (%)	7.30	7.38	0.64	5.40	10.70
Populist party	1.00	0.58	0.49	0.00	1.00

Notes: The statistics are weighted by the municipality population size.

A.2 Reduced Form

Table A.2 and Table A.3 reports the reduced form results for the vaccination rates, hospitalizations and average annual costs for vaccine preventable diseases, respectively.

Table A.2: Reduced form - Vaccination rates

	(1) V_{mt} Hexavalent	(2) V_{mt} MMR	(3) V_{mt} Meningococcus	(4) V_{mt} Pneumococcus
ffs_{mt}	-0.007	-0.038**	-0.022	-0.071
	[0.012]	[0.019]	[0.049]	[0.055]
N	7,239	7,238	7,061	7,066
Controls (Twitter)	✓	✓	✓	✓
Controls (Socioec.)	✓	✓	✓	✓
City and year FE	✓	✓	✓	✓
Reg × year	✓	✓	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
$Notes$: All estimates include city, region and year fixed effects and region-specific time trends fixed effects. Standard errors (in brackets) are clustered on the municipality level. Estimates are weighted by municipality population size.

Table A.3: Reduced form - Hospitalizations.

	(1) V_{mt} non-target pop.	(2) V_{mt} non-target pop.(MMR)	(3) V_{mt} Children age 1-10 (MMR)
Panel a: Hopitalizations			
s_{mt}	0.123**	0.104***	0.0603*
	[0.0550]	[0.0309]	[0.0323]
Panel b: Healthcare costs			
s_{mt}	383.7*	326.7**	147.0*
	[203.9]	[146.2]	[78.60]
	(4) (Hexav.)	(5) (Meningo.)	(6) (Pneumo.)
	Non-target population		
Panel c: Hopitalizations			
s_{mt}	0.0266	-0.0002	-0.005
	[0.0432]	[0.0005]	[0.0079]
Panel d: Healthcare costs			
s_{mt}	-138.3	-5.515	-17.42
	[340.2]	[8.761]	[21.84]
	Children age 1-10		
Panel e: Hopitalizations			
s_{mt}	0.0005	0.00008	0.006
	[0.0097]	[0.0019]	[0.0074]
Panel f: Healthcare costs			
s_{mt}	-32.78	5.163	0.478
	[26.40]	[7.744]	[23.39]
N	5,136	5,136	5,136
CONTROL (Twitter)	✓	✓	✓
CONTROL (Socioec.)	✓	✓	✓
CITY and YEAR FE	✓	✓	✓
Reg × Year	✓	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. $Notes$: All estimates include city and year fixed effects and region-specific time trends. Standard errors (in brackets) are clustered at the municipality level. Estimates are weighted by the municipality population size.

A.3 Results - Vaccination Rates (full set of estimates)

Table A.4 shows second stage estimates related to vaccination rates under several specifications.

Table A.4: Results of the Second stage of the OLS and the M2SLS - Vaccination rates

	(1) OLS V_{mt}	(2) M2SLS V_{mt}	(3) M2SLS V_{mt}	(4) M2SLS V_{mt}	(5) M2SLS V_{mt}	(6) M2SLS V_{mt}	(7) M2SLS V_{mt}
Panel a: Hexavalent (94.06)							
s_{mt}	-0.001	-0.033	-0.005	-0.004	-0.003	-0.008	-0.023
	[0.002]	[0.026]	[0.015]	[0.015]	[0.015]	[0.016]	[0.015]
N	7,239	7,601	7,239	7,239	7,239	7,239	7,239
Panel b: MMR (89.53)							
s_{mt}	-0.005	-0.157***	-0.045*	-0.037	-0.041*	-0.048*	-0.043**
	[0.003]	[0.044]	[0.026]	[0.024]	[0.024]	[0.027]	[0.021]
N	7,238	7,600	7,238	7,238	7,238	7,238	7,238
Panel c: Menigococcal (81.32)							
s_{mt}	-0.002	-0.470***	-0.030	-0.001	-0.013	-0.009	-0.040
	[0.008]	[0.128]	[0.066]	[0.062]	[0.064]	[0.062]	[0.054]
N	7,061	7,438	7,061	7,061	7,061	7,061	7,061
Panel d: Pneumococcal (82.64)							
s_{mt}	-0.0001	-0.206**	-0.060	-0.032	-0.046	-0.071	-0.029
	[0.008]	[0.086]	[0.072]	[0.063]	[0.067]	[0.069]	[0.052]
N	7,066	7,429	7,066	7,066	7,066	7,066	7,066
CONTROL (Twitter)	✓				✓		✓
CONTROL (socioeconomics)	✓					✓	✓
YEAR FE	✓	✓	✓	✓	✓	✓	✓
CITY FE	✓		✓	✓	✓	✓	✓
Reg × Year	✓			✓	✓	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

$Notes$: All estimates include city and year fixed effects as well as region specific time trends. Standard errors (in brackets) are clustered at the municipality level and have been corrected in the second stage. Estimates as well as averages of V_{mt} are weighted by the municipality population size.

A.4 Results - Hexavalent, Meningococcus and Pneumococcus Hospitalizations

Table A.5 reports the estimates for the second stage, indicating the number of hospitalizations and the average annual costs associated with administering Hexavalent, Meningococcal, and Pneumococcal vaccinations.

Table A.5: Results of the OLS and the Second stage of the M2SLS - Hospitalizations.

	(1)	(2)	(3)	(4)	(5)	(6)
	OLS	M2SLS	OLS	M2SLS	OLS	M2SLS
	V_{mt}	V_{mt}	V_{mt}	V_{mt}	V_{mt}	V_{mt}
	(Hexav.)	(Hexav.)	(Meningo.)	(Meningo.)	(Pneumo.)	(Pneumo.)
				Non-target population		
Panel a: Hopitalizations						
s_{mt}	0.009	0.025	-0.0001	-0.0003	-0.0006	-0.021
	[0.012]	[0.092]	[0.0002]	[0.0009]	[0.002]	[0.015]
Panel b: Healthcare costs						
s_{mt}	102.0	-628.4	-4.756	-20.81	-10.53*	-46.519
	[100.6]	[700.3]	[3.976]	[16.46]	[6.103]	[37.26]
				Children age 1-10		
Panel a: Hopitalizations						
s_{mt}	-0.0001	0.002	0.0001	0.0003	-0.002	0.009
	[0.003]	[0.016]	[0.001]	[0.004]	[0.002]	[0.011]
Panel b: Healthcare costs						
s_{mt}	12.74	-66.18	-0.528	10.36	-3.788	-37.99
	[18.45]	[49.21]	[2.887]	[14.90]	[6.229]	[42.28]
N	3,331	3,331	3,331	3,331	3,331	3,331
Controls (Twitter)	✓	✓	✓	✓	✓	✓
Controls (Socioecon.)	✓	✓	✓	✓	✓	✓
City and year FE	✓	✓	✓	✓	✓	✓
Reg × year	✓	✓	✓	✓	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
Notes: All estimates include city and year fixed effects as well as region specific time trends. Standard errors (in brackets) are clustered at the municipality level and have been corrected in the second stage. Estimates are weighted by the municipality population size.

A.5 Results - Mortality among hospitalized

Descriptive statistics and second stage estimates for mortality cases of hospitalized patients with vaccine-preventable diseases are presented in Table A.6 and Table A.7, respectively.

Table A.6: Descriptive statistics of mortality among hospitalized due to vaccine-preventable diseases (2013-2016)

	Median	Mean	sd	Min	Max	N
Panel a: Hopitalizations						
Non-target population	0.00	1.00	2.08	0.00	97.56	31,760
Non-target population (Hexav.)	0.00	0.93	1.98	0.00	97.56	31,760
Non-target population (Meningo.)	0.00	0.00	0.15	0.00	29.02	31,760
Non-target population (MMR)	0.00	0.04	0.38	0.00	28.92	31,760
Non-target population (Pneumo.)	0.00	0.14	0.68	0.00	97.56	31,760
Children age 1-10	0.00	0.01	0.12	0.00	7.28	31,760
Children age 1-10 (Hexav.)	0.00	0.00	0.06	0.00	5.53	31,760
Children age 1-10 (Meningo.)	0.00	0.00	0.07	0.00	6.67	31,760
Children age 1-10 (MMR)	0.00	0.00	0.05	0.00	7.28	31,760
Children age 1-10 (Pneumo.)	0.00	0.00	0.04	0.00	2.81	31,760

Notes: The statistics refer to 7,940 municipalities for the time period between 2013-2016 and are weighted by the municipality population size.

Table A.7: Results of the Second stage of the M2SLS - Mortality among Hospitalized.

	(1) V_{mt} Main	(2) V_{mt} Hexavalent	(3) V_{mt} MMR	(4) V_{mt} Meningococcus	(5) V_{mt} Pneumococcus
Panel a: Non-target population					
Hospitalizations					
s_{mt}	-0.013	-0.019	0.005	-0.0006	-0.006
	[0.014]	[0.013]	[0.003]	[0.0004]	[0.005]
N	3,331	3,331	3,331	3,331	3,331
Panel a: Children age 1-10					
Hospitalizations					
s_{mt}	-0.000005	0.0002	0.00006	-0.0002	-0.0001
	[0.0005]	[0.0002]	[0.00005]	[0.0003]	[0.0002]
N	3,331	3,331	3,331	3,331	3,331
Controls (Twitter)	✓	✓	✓	✓	✓
Controls (socioec.)	✓	✓	✓	✓	✓
City and year FE	✓	✓	✓	✓	✓
Reg $\times$ year	✓	✓	✓	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

$Notes$: All estimates include city and year fixed effects and region-specific time trends. Standard errors (in brackets) are clustered at the municipality level and have been corrected in the second stage. Estimates are weighted by the municipality population size.

Results - robustness checks

For completeness, we report here the full tables of robustness checks, including the second-stage results on all vaccinations.

Table A.8: M2SLS Individual - Second stage (Vaccination rate)

	(1) Main	(2) Twitter algortithm	(3) Emilia Romagna Law	(4) Populist Party Law	(5) Network distance	(6) Excluding FoF geolocated in the user's municipality
	V_{mt}	V_{mt}	V_{mt}	V_{mt}	V_{mt}	V_{mt}
Panel a: Hexavalent (94.06)						
s_{mt}	-0.023	-0.021	-0.024	-0.022	-0.023	-0.018
	[0.015]	[0.016]	[0.015]	[0.019]	[0.015]	[0.016]
N	7,239	7,239	7,239	7,239	7,239	7,239
Panel b: MMR (89.53)						
s_{mt}	-0.043**	-0.047**	-0.048**	-0.055**	-0.050**	-0.042**
	[0.021]	[0.022]	[0.022]	[0.027]	[0.023]	[0.021]
N	7,238	7,238	7,238	7,238	7,238	7,238
Panel c: Menigococcus (81.32)						
s_{mt}	-0.040	-0.044	-0.041	-0.026	-0.038	-0.043
	[0.054]	[0.057]	[0.054]	[0.069]	[0.0559]	[0.056]
N	7,061	7,061	7,061	7,061	7,061	7,061
Panel d: Pneumococcus (82.64)						
s_{mt}	-0.029	-0.035	-0.031	-0.104	-0.027	
	[0.057]	[0.056]	[0.057]	[0.083]	[0.060]	[0.056]
N	7,066	7,066	7,066	7,066	7,066	7,066
Controls (Twitter)	✓	✓	✓	✓	✓	✓
Controls (Socioec.)	✓	✓	✓	✓	✓	✓
City and year FE	✓	✓	✓	✓	✓	✓
Reg × year	✓	✓	✓	✓	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

$Notes$: All estimates include city and year fixed effects and region-specific time trends. Standard errors (in brackets) are clustered at the municipality level and have been corrected in the second stage. Estimates, as well as averages of V_{mt}, are weighted by the municipality population size.

Results - full sample

For the sake of clarity, we present the findings for the full sample of geolocated tweets, including the first percentile of municipalities that were left out of our primary findings.

Table A.9: Results of the M2SLS - Vaccination rate, Hospitalizations and Costs (full sample)

Panel a: First stage				
	s_{it}	s_{it}	s_{it}	s_{it}
	(30.33)	(30.33)	(30.33)	(30.33)
ffs_{it} (28.77)	0.709***	0.709***	0.710***	0.710***
	[0.0164]	[0.0164]	[0.0164]	[0.0164]
N	130,896	130,896	130,896	130,896
Controls (Twitter)		✓		✓
Controls (socioec.)			✓	✓
City and year FE	✓	✓	✓	✓
Reg × year	✓	✓	✓	✓
F-stat	1,875.96	1,874.25	1,865.62	1,868.15

Panel b: Second stage - Vaccination Rate				
	(1)	(2)	(3)	(4)
	V_{mt}	V_{mt}	V_{mt}	V_{mt}
	Hexavalent	MMR	Meningococcus	Pneumococcus
OLS				
s_{mt}	-0.0003	-0.003	-0.005	-0.006
	[0.001]	[0.002]	[0.005]	[0.005]
M2SLS				
s_{mt}	-0.014	-0.030**	-0.047	-0.012
	[0.010]	[0.014]	[0.035]	[0.034]
N	10,281	10,275	9,978	9,994

Panel c: Second stage - Non-target Population					
	(1)	(2)	(3)	(4)	(5)
	V_{mt}	V_{mt}	V_{mt}	V_{mt}	V_{mt}
	Main	Hexavalent	MMR	Meningococcus	Pneumococcus
Hopitalizations					
s_{mt}	0.181**	0.0440	0.150***	-0.000370	-0.00788
	[0.0779]	[0.0610]	[0.0436]	[0.000677]	[0.0111]
Healthcare costs					
s_{mt}	585.2**	-191.6	464.3**	-8.797	-24.01
	[286.3]	[478.9]	[205.1]	[11.74]	[30.78]
N	5,136	5,136	5,136	5,136	5,136

Panel d: Second stage - Children age 1-10					
		(1)	(2)	(3)	(4)
		V_{mt}	V_{mt}	V_{mt}	V_{mt}
		Hexavalent	MMR	Meningococcus	Pneumococcus
Hopitalizations					
s_{mt}		-0.0003	0.0846*	-0.0001	0.008
		[0.0137]	[0.0452]	[0.00266]	[0.0104]
Healthcare costs					
s_{mt}		-48.82	206.8*	6.494	-2.958
		[37.29]	[110.0]	[10.97]	[33.27]
N		5,136	5,136	5,136	5,136
Controls (Twitter)	✓	✓	✓	✓	✓
Controls (socioec.)	✓	✓	✓	✓	✓
City and year FE	✓	✓	✓	✓	✓
Reg × year	✓	✓	✓	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

$Notes$: All estimates include city and year fixed effects as well as region specific time trends. Standard errors (in brackets) are clustered at the municipality level and have been corrected in the second stage. Estimates, as well as the s_{it} are weighted by the municipality population size.

List of events and classification

Table A.10: List of events

date	Classification	Description
14jan2013	Efficacy	Vaccino anti-meningite B,protezione in 95% vaccinati
28jan2013	Efficacy	Bimba di 3 anni muore di setticemia forse provocata da pneumococco
17apr2013	Efficacy	In Italia migliaia di casi di morbillo e rosolia evitabili
02may2013	Efficacy	Epidemia di morbillo nel Regno Unito: dobbiamo preoccuparci?
15may2013	Efficacy	Sanita': Napoli; vaccini gratis contro rosolia congenita
10jun2013	Efficacy	Veneto: casi di complicanze da morbillo e varicella in persone non vaccinate
13jun2013	Efficacy	Sanita': polmonite, in Lombardia costa 68 mln di euro l'anno
26jun2013	Efficacy	Arriva vaccino contro meningite ceppo B
27jun2013	Efficacy	Meningite: Veneto; e' 10% tutte malattie invasive da batteri
24sep2013	Efficacy	In Lombardia campagna vaccinazione Hpv anche per uomini
12may2016	Efficacy	Morbillo in Campania: allerta dei medici per le basse coperture vaccinali nella Regione
08jul2016	Efficacy	Documento sui vaccini della Fnomceo
23jun2017	Efficacy	Un bimbo malato di leucemia morto per il morbillo: "Contagiato dai fratelli non vaccinati"
21may2013	Politics and Mandate	Proposta di Legge d'iniziativa del deputato Burtone Istituzione della Giornata in ricordo delle persone decedute o rese disabili a causa di vaccinazioni
01jul2016	Politics and Mandate	introduzione obbligo vaccinale per gli asili nido in Emilia Romagna
23nov2016	Politics and Mandate	introduzione obbligo vaccinale per gli asili nido in Emilia Romagna
26jan2017	Politics and Mandate	accordo stato-regioni per una legge nazionale sui vaccini
03may2017	Politics and Mandate	Tutte le volte del Movimento 5 Stelle contro i vaccini, la Rete smentisce Grillo
20may2017	Politics and Mandate	Veneto Vaccini, torna l'obbligo. Zaia: «Misura inefficace». I virologi: «Salva la vita»
07jun2017	Politics and Mandate	decreto legge n 73 /2017 "legge lorenzin"
08jun2017	Politics and Mandate	Alto Adige, dove il consiglio provinciale ha approvato all'unanimità una mozione che chiede "lo stralcio delle misure coercitive previste dal decreto sui vaccini e una campagna di sensibilizzazione ampia ed equilibrata
28jul2017	Politics and Mandate	Il Decreto vaccini è legge, tutte le novità
10jan2018	Politics and Mandate	Vaccini, Salvini: "Con noi al governo via l'obbligo". Lorenzin: "Per qualche voto gioca con la salute dei bambini"
22jun2018	Politics and Mandate	Vaccini, Salvini: 'Inutili 10 vaccini obbligatori'. Burioni: 'Bugie pericolosissime'. Alt Di Maio e della Grillo
05aug2018	Politics and Mandate	Taverna, la sciamannata vicepresidente del Senato: i vaccini? Come i marchi alle bestie
12aug2018	Politics and Mandate	Salvini sa che i soldati devono vaccinarsi? Mettetevi d'accordo" Ironia social sulla # LevaObbligatoria
06sep2018	Politics and Mandate	Vaccini a scuola, colpo di scena: emendamento ripristina l'autocertificazione
08jan2013	Trustful Source	Pneumococco. L'EU estende l'uso di Prevenar 13 a bambini e adolescenti fino a 17 anni
17sep2013	Trustful Source	Oms, nessun legame tra vaccini e autismo
26nov2014	Trustful Source	Lorenzin condanna il giudice del tribunale del lavoro:"Quella sentenza sul vaccino è un attentato alla salute pubblica
17feb2015	Trustful Source	sentenza 1767/14 della corte d'Appello di Bologna nella causa d'appello alla sentenza 15.03.2012 Rimini
10jan2016	Trustful Source	La battaglia dei vaccini - presadiretta
27mar2016	Trustful Source	Robert De Niro ritira il film sul legame tra vaccini e autismo dal Tribeca Film Festival di New York
01jun2016	Trustful Source	Procura di Trani ha riconosciuto l'inconsistenza del presunto legame tra la vaccinazione trivalente MPR (contro morbillo, parotite e rosolia) e autismo
21apr2017	Trustful Source	L'Ordine dei medici di Treviso ha radiato Roberto Gava, considerato uno dei paladini dei no-vax in Italia
02may2017	Trustful Source	new york times pubblica "Populism, Politics and Measeles"
07sep2017	Trustful Source	TAR Lazio – decreto 7 settembre 2017: respinto il ricorso del Codacons riguardante le misure adottate per ottemperare agli obblighi di documentazione vaccinale
22nov2017	Trustful Source	la sentenza della Corte costituzionale considera legittimo l'obbligo dei vaccini nel contesto attuale definito dal Decreto 73/2017 e respinge i ricorsi presentati dalla Regione Veneto
05jul2018	Trustful Source	Giulia Grillo: «Vaccini, a scuola con autocertificazione. L'obbligo cambierà. Io incinta, vaccinerò mio figlio»

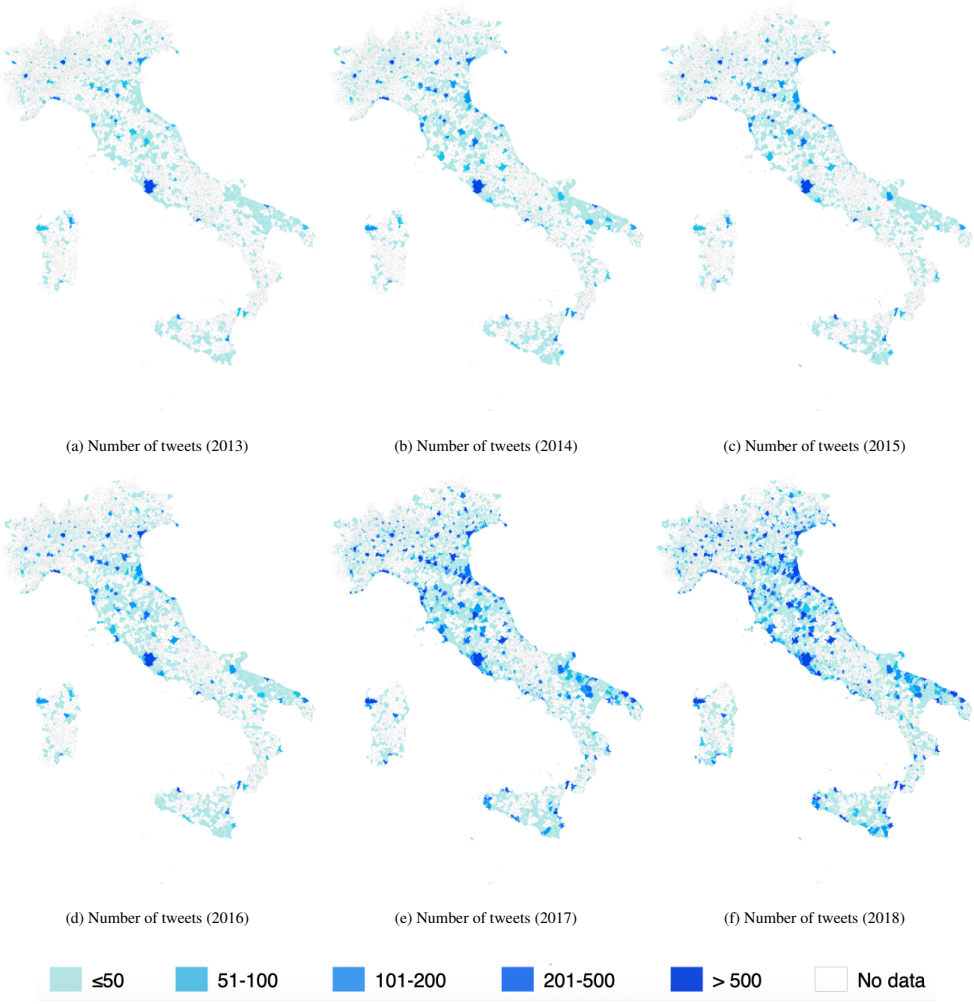
List of events - *continued from previous page*

date	Classification	Description
01jul2013	Vaccine Unsafe	Tribunale di Pesaro, 1 luglio 2013
07jul2013	Vaccine Unsafe	“Vaccine adverse events reporting system" pubblica uno studio dove Heidi Stevenson parla di migliaia di morti per colpa di vaccini, probabilità di morte aumenta del 50 per cento – e con ogni dose di vaccino supplementare"
10oct2013	Vaccine Unsafe	Vaccini:1 italiano su 2 è contrario, "inutili e poco sicuri"
11nov2013	Vaccine Unsafe	Tribunale di Pesaro, 11 novembre 2013
09jan2014	Vaccine Unsafe	Venti nuovi casi di danno da vaccini alla settimana per l’avvocato di Rimini
17mar2014	Vaccine Unsafe	Si vaccina poco? Big Pharma fa pressing sulle Asl e telefona alle famiglie
02jul2014	Vaccine Unsafe	Tribunale di Rimini, 2 luglio 2014, n. 217
23sep2014	Vaccine Unsafe	Tribunale del lavoro di milano: vaccino esavalente Infanrix Hexa Sk causa l’autismo
20oct2014	Vaccine Unsafe	La presenta di DNA fetale umano nei vaccini é una possibile causa di autismo
28nov2014	Vaccine Unsafe	Aifa: “Tredici casi di morte sospetta” vacccino antiinfluenzale
11mar2015	Vaccine Unsafe	ll ministero riconosce l’indennizzo per un bimba Catanzaro
01jul2015	Vaccine Unsafe	Jim Carrey tweet sul mercurio nei vaccini
03jul2015	Vaccine Unsafe	Jim Carrey causa l’autismo
25may2016	Vaccine Unsafe	Bimba muore a 2 mesi in culla dopo il vaccino: l’Asl sostituisce tutti i lotti
10nov2016	Vaccine Unsafe	Corte d’Appello di Milano, 10 novembre 2016, n.1255
01feb2017	Vaccine Unsafe	Corte di Cassazione, Sezione 6 civile, 1 febbraio 2017, n. 2684
17apr2017	Vaccine Unsafe	report vaccino HPV

Additional Figures

Figure A1 shows tweets’ distribution across municipalities over time.

Figure A1: Tweets over time (2013-2018)



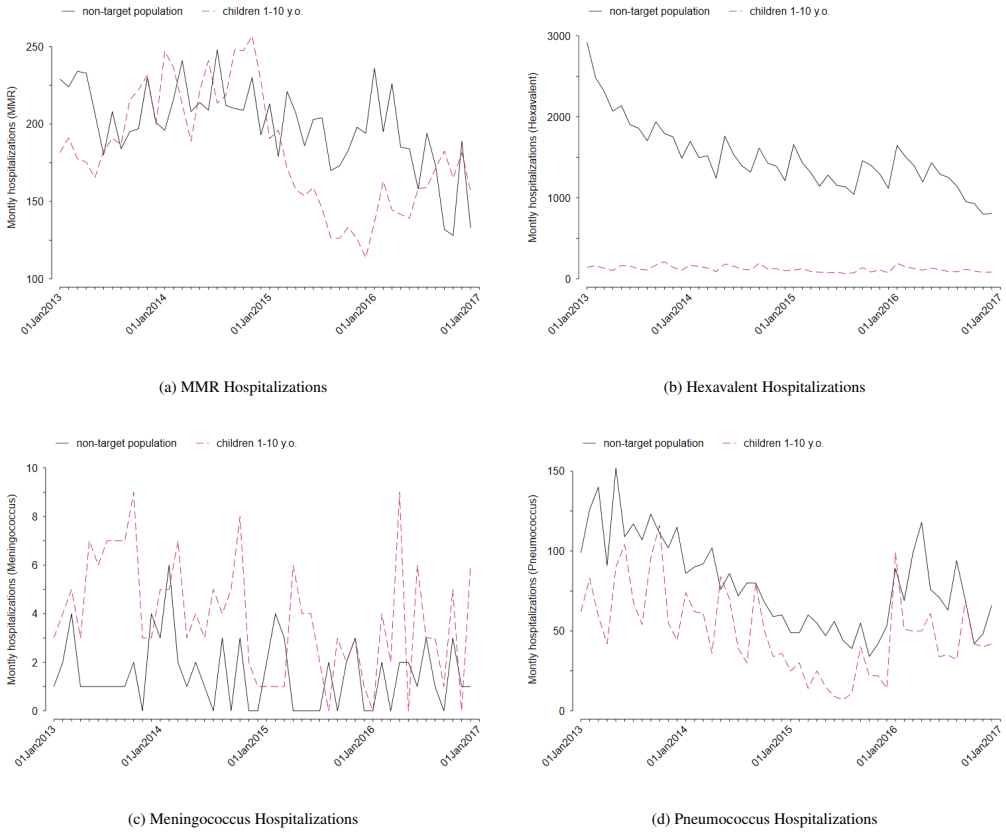
(a) Number of tweets (2013) (b) Number of tweets (2014) (c) Number of tweets (2015)

(d) Number of tweets (2016) (e) Number of tweets (2017) (f) Number of tweets (2018)

Notes: The sample consists of 830,253 tweets relative to a population of 80,471 unique users across 4,220 municipalities.

Figure A2 plots the monthly trends in hospitalizations among the vaccine-target population and in vulnerable populations that are not targeted by vaccines. The trends for hexavalent, pneumococcus, and meningococcus are generally comparable for the two groups. For diseases covered by the hexavalent vaccine, hospitalisation rates for the vaccine targeted population was stable for the entire study period, while hospitalization rates for the non-targeted population decreased from 3000 to 1000 during the study period. For Meningococcus, we see a higher overall number of hospitalizations for target population than for target population. Instead, the trend for pneumococcus is similar across groups. However, for the MMR vaccine, the hospitalization rate trends were opposite between January 2015 and January 2017, which was a period marked by several measles epidemic outbreakshe hospitalization rate for the group that received the vaccine decreased from 250 in 2015 to 170 in 2016, with the lowest point being in 2016. During the study period, the hospitalization rate for the vaccine non-targeted population remained steady at an average of 200 hospitalizations until july 2016, but then started to decrease after August 2016.

Figure A2: Hospitalization trends (2013-2016)



(a) MMR Hospitalizations

(b) Hexavalent Hospitalizations

(c) Meningococcus Hospitalizations

(d) Pneumococcus Hospitalizations

Notes: The hospitalization rate for the vaccine-targeted population is represented by the red dashed line, while the vaccine non-targeted vulnerable groups are represented by the black solid line.

Appendix B

The Model of Opinion Dynamics and Network Formation.

The model builds on Baumann et al. (2020)'s work on endogenous polarization dynamics in social networks. In the model we consider a continuum of individuals in a discrete, infinite time setting $[t = 0, 1, .., \infty]$. Each individual i has a stance on vaccinations $s_i^t = [\underline{s}, \bar{s}]$ which spans from unconditional support to hesitancy. We assume that the stance reflects individuals' opinions on the overall utility of vaccinations a one-to-one mapping between parents' and children's (perceived) utility.

Individual stances evolve over time from initial positions s_i^0, drawn from a distribution $S^0 \sim F_s(0)$, with finite first and second moments; in particular, $\mu^0 = \mathbb{E}(s_i^0)$, stands for the average initial stance in the society. To reflect the observed distribution of initial stances - on average pro-vaccines - in the baseline simulations $\mu^0 \leq 0$ and initial stances are drawn from a Gaussian distribution. We obtain qualitatively equivalent results when we move to a case where the initial distribution of opinions is centered around zero (i.e., $\mu^0 = 0$).

The opinion dynamics within the social network are entirely driven by the interactions among agents and are described by a system of N coupled differential equations:

$$\dot{s}_i = -s_i + \mathbb{I} \sum_{j=1}^{N} W_{ij}(t) tanh(\alpha_t s_j) \tag{B.1}$$

In Equation (B.1) $\mathbb{I}$ measures the strength of the interaction among users of the platform, $W(t)$ is a time-varying spatial contiguity matrix, whose i^{th}, j^{th} elements represent every link between individuals in the network - i.e., $w_{ij}(t) = 1$ if i interacts with j, $w_{ij}(t) = 0$ otherwise. The function $tanh(\cdot)$ is the hyperbolic tangent function, which provides a sigmoidal influence function of peers on individuals' stances, ensuring that i) an agent's i stance influences others monotonically and that ii) such influence "flattens" in the extremes. Finally, α_t is the degree of controversialness of the topic.

The contiguity matrix $W(t)$ evolves according to an activity-driven (AD) temporal network (Perra et al., 2012), where each agent is characterized by the propensity to interact with a share $\omega_i \in [\epsilon, 1]$ of other agents, and the probability of an interaction is driven by homophily (Bessi et al., 2016) - that is to say, individuals are more likely to interact with like-minded peers, and we model it as a decreasing function of the (absolute) distance between i and j's opinions, $p_{ij}(t) = \frac{|s_i(t) - s_j|^{-\beta}}{\sum_j |x_i - x_j|^{-\beta}}$. Note that the β parameter that informs the power law decay of interaction probability includes effects as diverse as the endogenous preferences for homophily (i.e., to what extent individuals dislike the interaction with people of different stances) or the exogenous settings embedded in the social networks' algorithms - e.g., how likely one's content is to appear in a like-minded peer's home newsfeed.

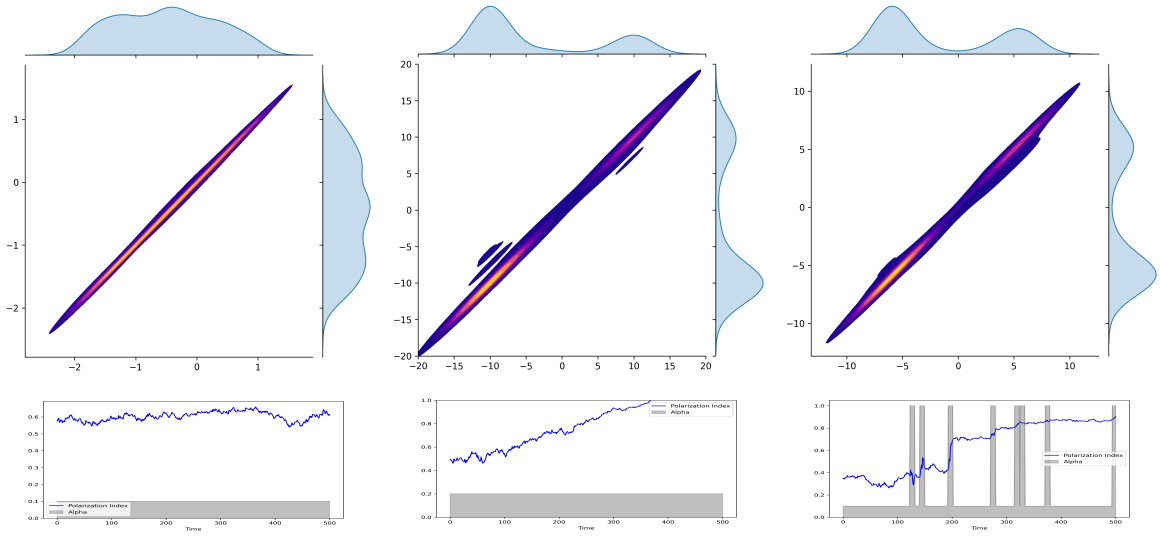
Figure 3 shows the predictions of the simulated models. The heatmaps show the distribution of stances for the users and their friends in a simulation for low controversialness ($\alpha = 0.1$ in panel a), relatively higher controversialness ($\alpha = 0.2$ in panel b), and time-varying controversialness (long periods of $\alpha = 0.1$ with short-lived outbursts of $\alpha = 1$ in panel c). The colors in the heatmaps represent the density of users, with lighter colors indicating a higher number of users. The marginal distribution of users' opinions and their friends' opinions are plotted on the x- and y-axis, respectively. The simulation shows that users are more likely to connect with peers who share similar opinions due to homophily.

In addition to homophily, higher controversialness strengthens the influence of peers' opinions on users who tend to form homogeneous groups. At the network level, this results in a correlation between users' and their friends' average opinions. When controversialness is low (panel a), the model converges to a bivariate Gaussian distribution centered at approximately (-.5,-.5); on the other hand, when the model is characterized by higher controversialness (panel b), it converges to a bivariate bimodal distribution with a high density of users

with like-minded friends, resulting in two echo chambers corresponding to opposite stances on vaccinations. In a more realistic simulation where long periods of low controversialness are interrupted by short-lived, high-controversialness outbursts (panel c), the model also generates echo chambers.

The figures below the heatmaps show the degree of polarization during the simulations. When controversialness is low, there is no trend in polarization within the population, but polarization increases with relatively high controversialness. Interestingly, with time-varying controversialness, polarization increases during the outbursts and remains stable at the new, higher level until the next shift.

Figure 3: Simulated distribution of stances



(a) Low controversialness (b) High controversialness (c) Time-varying controversialness

Notes: user (x-axis) and average friends' (y-axis) distribution of stances in a simulated model when controversialness is low ($\alpha = .1$ in panel a), high ($\alpha = .2$ in panel b), and low with short-lived outbursts ($\alpha = 0.1$ and $\alpha = 1$ in panel c). In all models, the number of individuals is $N = 500$ and the periods are $T = 5$ - divided in 100 subperiods. Initial values (s_0) are randomly drawn from a gaussian distribution with $\mu = -0.2$ and $\sigma = 0.5$ to match the asymmetry of the initial opinions in the data.