Vax Populi: the Social Costs of Online Vaccine Skepticism*

Matilde Giaccherini[1,3] and Joanna Kopinska[2,1]

[1]*CEIS, Universitá degli studi di Roma "Tor Vergata"*

[2]*Universitá degli studi di Roma "La Sapienza"*

[3]*CESifo*

December 31, 2022

Abstract

We quantify the effects of online vaccine skepticism on vaccine uptake and health complications for individuals not targeted by immunization campaigns. We collect the universe of Italian vaccine-related tweets for 2013-2018, label anti-vax stances using NLP, and match them with vaccine coverage and vaccine-preventable hospitalizations at the most granular level (municipality and year). We propose a model of opinion dynamics on social networks that matches the observed data and shows that a vaccine mandate increases the average vaccination rate, but it also increases the controversialness around the topic, endogenously fueling polarization of opinions among users. We then leverage the intransitivity in network connections with "friends of friends" to isolate the exogenous source of variation for users' vaccine-related stances and implement an IV strategy. We find that a 10pp increase in the municipality anti-vax stance causes a 0.43pp decrease in coverage of the Measles-Mumps-Rubella vaccine, 2.1 additional hospitalizations every 100k residents among individuals untargeted by the immunization (newborns, the immunosuppressed, pregnant women) and an excess expenditure of 7,311 euro, representing an 11% increase in health expenses.

JEL Classification: I18, L82, Z18

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

*Corresponding author - Giaccherini: matilde.giaccherini@uniroma2.it. Kopinska: joanna.kopinska@uniroma1.it. We would like to thank Tiziano Arduini, Federico Belotti, Vincenzo Carrieri, Ludovica Gazze, Tommaso Orlando, Francesco Pierri, Francesco Sobbrio, Lucia Rizzica, Lorenzo Rovigatti, Gianluca Russo, for useful suggestions and the participants of the CESifo Area Conference in Economics of Digitization. Special thanks go to Fabian Baumann, Philipp Lorenz-Spreen, Igor M. Sokolov, and Michele Starnini for sharing their code before its publication. 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 wide diffusion of the internet and, more recently, social media has granted virtually unlimited access to information that is not subject to fact-checking or editorial judgment. As a result, the ability of consumers to discriminate between "good" and "fake" (or unsubstantiated) news has decreased substantially. Additionally, social networks are heavily influenced by ideological "echo chambers" (Cinelli et al., 2021), fueling 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 spread (Allcott and Gentzkow, 2017).

The phenomenon of misinformation is deeply ingrained in our society, impacting political, economic, and social well-being(Vosoughi et al., 2018). COVID-19 conspiracy aside, the link between vaccines and autism is one of the most propagated pieces of 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 misinformation continues to be perpetuated (see among others Allcott et al., 2019, Chiou and Tucker, 2018). The rise of social media has provided an unparalleled platform for the dissemination of misinformation about vaccines (Burki, 2019). 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). Due to safety concerns, an increasing number of parents are choosing not to vaccinate their children, consuming the benefits of the herd immunity granted by others (Esposito et al., 2014, Smith et al., 2017). In Italy, as in many other countries, this has led to decreasing vaccination rates and outbreaks of diseases such as measles. This has sparked a policy debate and led to the introduction of legal measures that impose costs on individuals who choose not to vaccinate. Although vaccine mandates curtailing individual freedom have always been controversial, opposed, and disputed, 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.

In the ongoing conflict between personal interest and public health endeavors, it is natural to ask what effects online skepticism about vaccines has 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, communicable diseases can have significant externalities, including higher hospitalization rates and costs for unvaccinated individuals. It is thus worth considering the extent to which these externalities impact individuals who are not targeted by vaccination campaigns.

In this study, we investigate the effects of online skepticism about vaccines on local public health outcomes such as vaccination rates, vaccine-preventable hospitalizations, and relative costs. We use Twitter's Academic Application Programming Interface (API) data to analyze the spread of vaccine skepticism and construct a measure of vaccine-related attitudes. This data includes all publicly available Italian tweets from 2013-2018. According to Kim (2022), Twitter data can be used to measure and track public attitudes towards policy-relevant topics over time and across different locations. This can provide valuable insight into how these attitudes evolve and where they are most prevalent. We rely on a Natural Language Processing (NLP) transfer learning model. In particular, we build our model as in AlBERTo, a pre-trained Bidirectional Encoder Representations from Transformers (BERT) to distinguish vaccine-skeptic tweets from other tweets (Polignano et al., 2019). We treat geolocated anti-vaccine Twitter messages as a proxy for the presence of the anti-vaccination movement in Italian municipalities.

Next, we analyze how individual social media attitudes towards vaccines evolved in Italy using a model of opinion dynamics in social networks. This model allows us to formalize the sources of endogeneity that affect the relationship between the spread of anti-vaccine opinions on social media and vaccine hesitancy. On the one hand, exposure to extreme opinions can influence an individual's stance (the "exposure effect"). On the other hand, people tend to form connections with like-minded peers, particularly when the topic is controversial (the "link formation effect"). The endogeneity in link formation among users poses a challenge for causal inference, and naive estimates of the impact of exposure to online vaccine skepticism on vaccine hesitancy are likely to be biased.

In this context, we provide an estimate of the monetary effects of online vaccine skepticism on society as a whole and on individuals who are not targeted by immunization campaigns. To measure the *exposure effect* net of the *link formation effect,* we use an Instrumental Variables approach. We exploit the complexity of the Twitter network structure and leverage the intransitivity of network

connections, as described in (Bramoullé et al., 2009), to build a valid instrument. Under certain assumptions, which we carefully examine, exposure to the vaccine-related stances of an individual's "friends of friends" can be considered an exogenous source of variation for an individual's stance. We assume that Twitter stances reflect individuals' opinions on the overall utility of pediatric vaccines.

Since data on individual vaccine hesitancy is typically unavailable and cannot be linked to social media accounts, we pair vaccine-related tweets with disease-specific vaccine coverage rates, vaccine-preventable hospitalizations, and relative costs at the municipality level for the years 2013-2018. Until 2017, Italy had four mandatory vaccines (for polio, diphtheria, tetanus, and hepatitis B, often combined with Haemophilus influenzae type b and whooping cough), although this mandate was not legally enforced. Vaccines for MMR, chickenpox, meningococcal, and pneumococcal diseases were only strongly recommended, so their use was at the discretion of parents. Only in late 2017 did the scope of mandatory pediatric vaccines in Italy expand and become legally enforceable upon school enrollment, with a transitional period allowing parents to comply with the new requirements over the following year. We focus on the period from 2013-2018, which allows for a significant amount of time for Twitter vaccine stances to potentially impact vaccination rates before the vaccine mandate was fully implemented. Importantly, within the administrative data on all Italian hospital discharges, we distinguish between the pediatric target population and the non-target, vulnerable population (infants aged 0-12 months, pregnant women, and immunosuppressed patients). This allows us to compute the prevalence and costs of vaccine-preventable hospitalizations for those affected by the vaccination campaign and its spillover effects.

We use a Mixed two-stage least squares approach to estimate the effects of online vaccine skepticism on vaccinations and hospitalizations. In the first stage, we use the user-specific "friends-of-friends" network, and in the second stage, we use municipality-level aggregated data on vaccinations and hospitalizations. This allows us to account for the complex relationships between social media activity and real-world health outcomes.

Our estimates consistently show that exposure to online vaccine skepticism reduced vaccination rates for MMR, the vaccine that received the most coverage from the anti-vax movement. We find that a 10 pp increase in average vaccine skepticism at the municipality level leads to a 0.43 pp decrease in vaccination coverage. Furthermore, this increased skepticism leads to higher rates of hospitalization

for vaccine-preventable diseases, as well as increased healthcare costs. Specifically, we find that a 10 pp increase in the average stance leads to 2.1 additional hospitalizations for every 100 thousand residents, as well as an excess expenditure of 7,311 euros, or an 11% increase in the respective healthcare costs. We perform several robustness checks to control for the impact of Twitter algorithm changes, local vaccine campaigns, and the impact of populist votes, and our results remain consistent. In addition to our baseline analysis, we present an alternative estimation strategy that addresses potential concerns about the exogeneity of our instrument, specifically the intransitivity of users' networks. Our results using this alternative strategy are comparable to our baseline results in terms of both magnitude and statistical significance. Finally, we examine the potential non-linear effects of lagged neighborhood stances on individual user stances in order to determine whether our results have implications for policymakers and public health agencies in terms of the measures that could be implemented on social networks to communicate with citizens.

We find that exposure to the stances of friends-of-friends has a stronger effect on pro-vaccine users compared to anti-vaccine users. This means that each unit change in the exposure stance is more effective at increasing vaccine hesitancy among pro-vaccine users rather than reducing it among vaccine skeptics. Our findings also indicate that political debates and statements from trustworthy sources can, on average, mitigate the negative effects of exposure to anti-vaccine viewpoints (or reinforce the effects of exposure to pro-vaccine content). These results suggest that informative campaigns about vaccines may be an effective and scalable intervention for shaping public health awareness.

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 fragile individuals not subject to the vaccination campaigns.

Additionally, we show that micro-interactions among Italian users in the period before the extension and legal enforcement of the vaccine mandate in 2017 were governed by a relative consensus. In

2017, the controversy around the mandate increased, leading to greater polarization and the formation of echo chambers. This network structure may have been reinforced by the Twitter amplification algorithm introduced in 2016, which we also found to have increased the effect of exposure to vaccine skepticism by friends-of-friends networks. These findings are consistent with the evidence presented by Acemoglu et al. (2021), who demonstrate that social media platforms interested in maximizing engagement tend to design their algorithms to create more homophilic communication patterns, or "filter bubbles."

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). Taken together, 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 has had a significant impact on self-reported beliefs and knowledge about COVID-19 vaccines through public health organization campaigns on Facebook and Instagram. 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 to COVID-19 were more likely to support social distancing and other public health behavior measures. Our findings on the effects of vaccine skepticism on public health provide direct evidence on the potential benefits of policies that raise awareness about the risks of infection with communicable diseases and promote preventive immunization.

2 Institutional background

Historically, the discovery of antibiotics and advances in vaccine technology have been major contributors to the improved life expectancy that we enjoy today. They have allowed us to protect ourselves from deadly infectious diseases, reducing their prevalence and saving countless lives. Paradoxically, due to the past success of collective vaccination efforts, individuals may underestimate the value of

immunization and be more willing to take the risk of being unprotected. The self-eroding nature of vaccination can lead to fluctuations in vaccine coverage, which can affect the level of protection for the entire community if herd immunity is not achieved (Siegal et al., 2009). As such, vaccine coverage is a classic example of a public good, where individuals may prioritize their own interests over those of the community when deciding whether or not to get vaccinated. This can lead to suboptimal participation in collective vaccination efforts and factors that may influence an individual's decision to get vaccinated include their perception of the risk of disease, cognitive biases, and local epidemiological conditions.

One of the turning points in the history of Italian vaccines 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 under the National Plan of Vaccine Prevention (PNPV). The PNPV establishes a vaccine calendar and offers eligible individuals free vaccines at Local Health Authorities (LHAs).[1]

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 as the so-called hexavalent or 6-in-1 vaccine. Vaccines for the trivalent MMR, chickenpox, meningo- and pneumo-coccal diseases were only strongly recommended, meaning that parents could choose whether or not to have their children vaccinated. Since 2012, when the Court of Rimini controversially confirmed a link between the MMR vaccine and autism, immunization rates have begun to decrease. The hesitancy has been fueled by the spread of non-scientific health information on the internet and the availability of low-quality news outlets promoting anti-scientific views. The lowest historical coverage rates were recorded in 2015, the year in which the Court of Bologna reversed the 2012 Rimini sentence (Carrieri et al., 2019). In response to 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,[2] which extended the scope of mandatory pediatric vaccines by enforcing them upon school enrollment and introduced sanctions against anti-vax doctors.

[1] The regions 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 (*Azienda Sanitaria Locale*). LHAs covering a population of 590,000 each 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.).

[2] The *Lorenzini*'s Decree.

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 controversial, opposed, and disputed, the PNPV proposers argued that falling uptake was driven by anit-vax sentiment and created sizable externalities, increasing the risk of infection not only for the eligible but also for fragile individuals not targeted by the vaccination.

3 Data

We collected the majority of 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 the users from 2013-2018.

We complement the Twitter data with 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 one contains yearly information on disease-specific vaccination rates provided by the *Local Health Authorities* (LHAs) and aggregated at the municipal level for the period 2013-2018.

The second one is an administrative dataset on the universe of Italian hospital admissions The second one is an administrative dataset on the universe of Italian hospital admissions, 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 users in 2018. Together with TikTok, Twitter has recently seen the fastest growth in its

user base. Twitter users tend to be older, with 39% of them being female and 61% being male.[3] Twitter is also the most followed by other 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 all Twitter users also using Facebook, 80% using YouTube, and 88% using Instagram [4].

Given the role of Twitter data for the information spread, we exploit the Twitter *Academic Research product track* to access the full archive of (as-yet-undeleted) tweets published on Twitter. In addition to the text of the tweets, the API provides additional 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.[5] This allows us to study the interplay between users' conversations on Twitter and their local environments.

In February 2016, Twitter introduced an "algorithmic timeline" which ordered tweets according to their relevance instead of appearing in the order they were posted. With this feature, Twitter intended to more effectively target tweets to individual users.

Download and filtering We collect all tweets containing the Italian correspondents of any of the following keywords: "vaccine(s)", "vaccination", "vaccinating", "anti-vax", "vax".[6] 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

[3] AgCom - *"Osservatorio sulle comunicazioni"*.

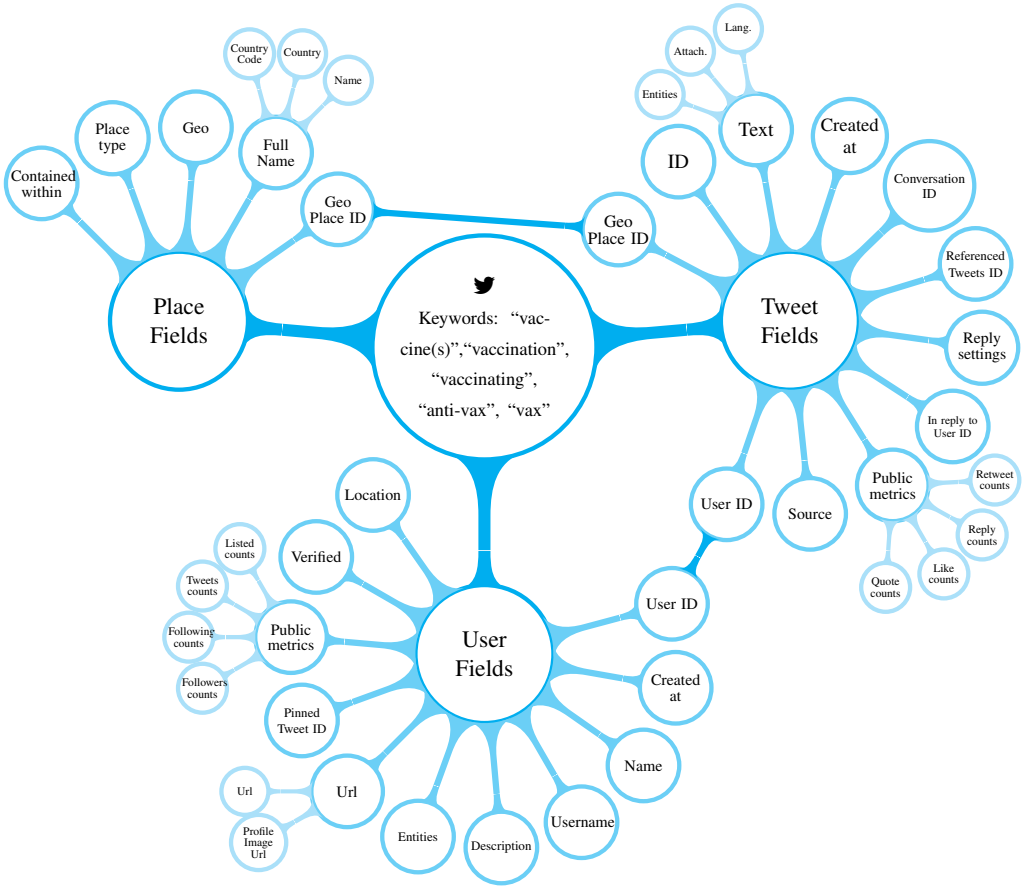
[4] Pew Research Center - *"Social Media Use in 2021"*

[5] 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.

[6] 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".

name, short bio and a few metrics - the number of friends, followers, and tweets posted - a verified status of the account, date when joined Twitter, and the location, when available (see Figure 1).[7]

Figure 1: Twitter objects



Notes: The figure shows: i) tweet, ii) user and iii) place fields for each retrieved tweet related to the keywords "vaccine(s)", "vaccination", "vaccinating", "anti-vax" and "vax".

We also collect information on when each account joined Twitter, as well as information on their *followers* and *following* (hereafter *friends*). *Followers* are Twitter users who follow a specific user, and *Friends* are the Twitter users that a specific user follows. We will discuss the specific aspects of the latter group in more detail in subsection 5.1.

Data cleaning We extract tweets' relevant content, excluding hashtags, special characters, emojis, and multimedia items. We omit *ex-post* all tweets containing only links or mentions[8] and those

[7] No personally identifiable information is included in this study.
[8] A tweet containing another user's username, preceded by "@".

produced by accounts that are temporarily unavailable due to violation of the Twitter media policy.[9] 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.

Within the Twitter sample, we geocode tweets in three consecutive steps: first, we use the tweet-specific geo-tag information ("Place fields" Figure 1); second, for the remaining tweets, we rely on the geo-tag information of the users ("location" in "User Fields" Figure 1); finally, we exploit Twitter 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.[10] Figure 9 in Table 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.[11].

Descriptive Statistics Out of 2.04 million tweets, the initial screening process leaves us with a sample of 2,017,539 tweets relative to 227,182 unique users, which through the process of geolocalization, is delimited to 830,253 tweets of 80,471 unique geotagged 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 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 ten times; only 0.7% of them have a verified account.[12]

[9] 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.

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

[11] 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

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

Table 1: Summary statistics: Twitter users.

	median	mean	sd	min	max
(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
(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
(c) Tweets' popularity					
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

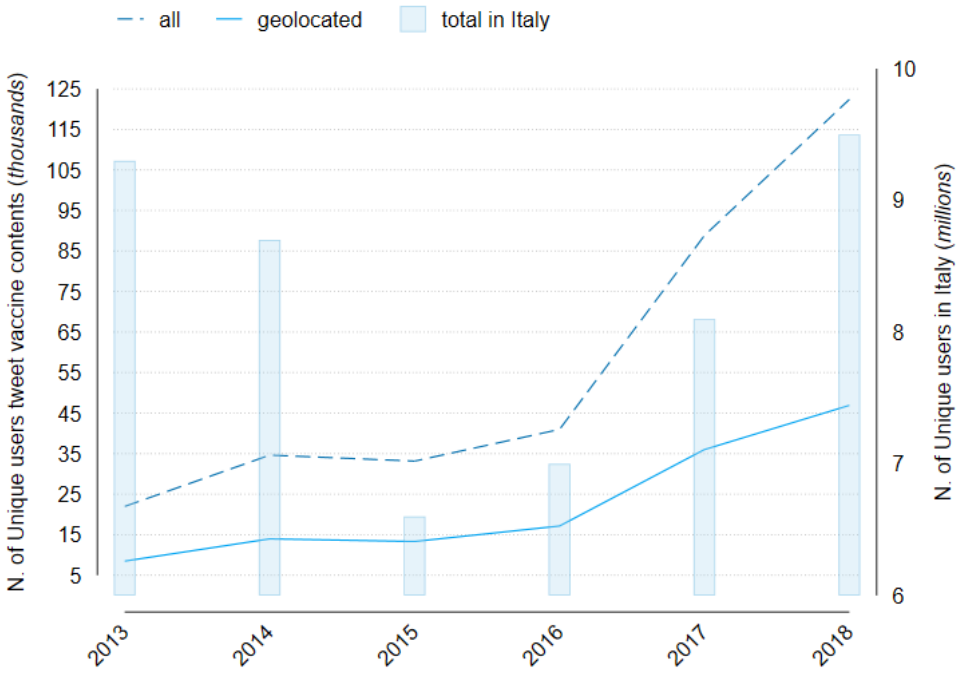
Notes: *(a)*: summary statistics of 80,471 geotagged unique users tweeting on vaccines (2013-2018); *(b)*: summary statistics of 830,253 geotagged tweets cleaned by hashtag, "RT @", "@", url and emoji; *(c)*: Tweet-related popularity metrics of 328,879 original tweets.

In the sample, 60% of the tweets are retweets or mentions, while 10% are replied to. 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 growing popularity of Twitter in the recent period.

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 Appendix A.

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 As the scope of the analysis is to study vaccine skepticism reflected in the corpus of tweets, we rely on a Natural Language Processing (NLP) transfer learning model. In particular, we build our model as in AlBERTo, a pre-trained Bidirectional Encoder Representations from Transformers (BERT) model in Italian proposed by Polignano et al. (2019), initially developed by Google.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" data needed to fine-tune the vaccine-specific task, saving computational time and data needs.

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 fake news users, pro-vaccine activists, and mainstream media outlets, and is labeled accordingly. 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%), created by 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"). [13] 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 the use of conspiracy theories or attacks on mainstream media outlets.

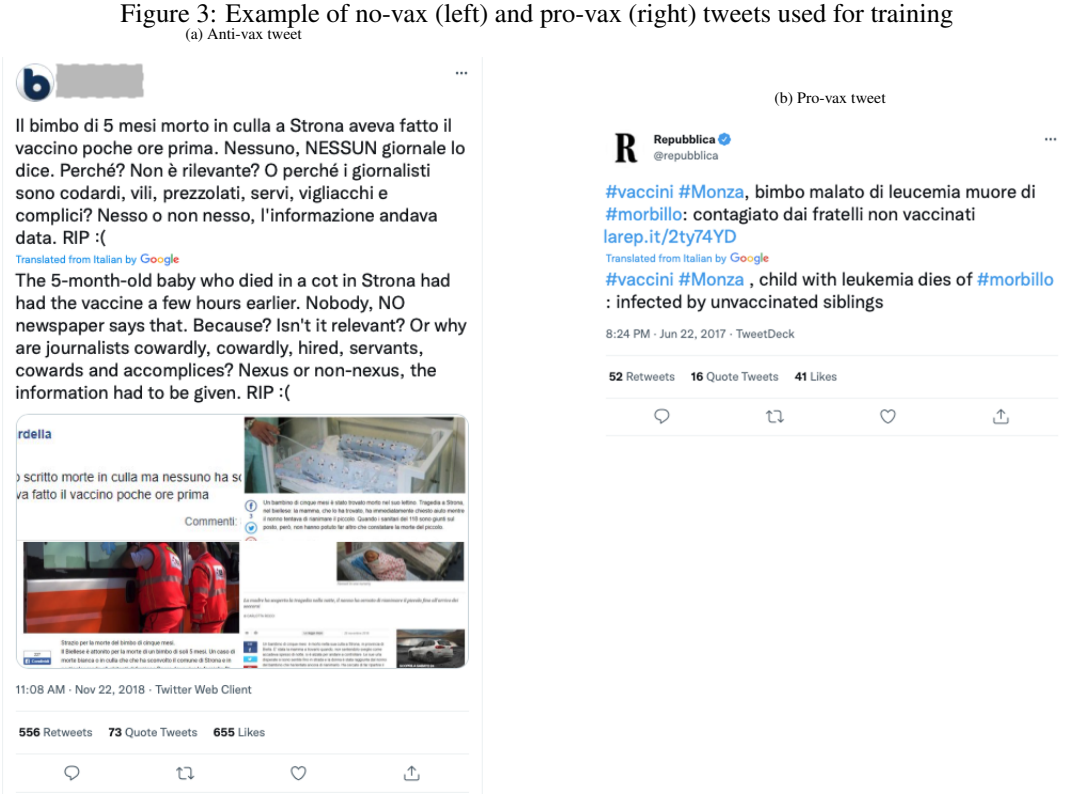
Once the model is trained, we classify all the tweets in our sample. Specifically, we generate a label $l_\tau \in \{0, 1\}$ for each tweet τ. Finally, we need to indentify the attitude of users. As in Cinelli et al. (2021), we define the leaning of a user as the average leaning of their tweets. Let 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 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).

[13] These linguistic con structs have been analyzed by Michaels (2008) in the tobacco industry.

Figure 3: Example of no-vax (left) and pro-vax (right) tweets used for training



Notes: Translation of Italian tweets is provided by Google.

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 the LHAs (*Azienda Sanitaria Locale*).[14] In Italy, public health policies are implemented by the regions through their health departments, while health protection and promotion are the responsibility of the Departments of Prevention within the 101 LHAs. LHAs are divided into 711 districts, each covering a population of around 84,000 on average. They are responsible for providing vaccinations free of charge to eligible individuals, such as children, the elderly, and other protected categories.

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** (hexavalent conjugate vaccine); Measles**, Mumps**, Rubella** (trivalent conjugate MMR vaccine), Meningococcal, and

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

Pneumococcal. One asterisk indicates vaccines which were compulsory in Italy between 2013 and 2017, and two asterisks indicate vaccines made mandatory by the "Lorenzin's Legge Vaccini" Law Decree 73, 2017. [15]

Table 2 shows the population-weighted average vaccination rates in the study period, along with their median, standard deviation, and minimum and maximum values (mostly 100%). 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 rates (around 94%), likely because it includes four mandatory shots, while the meningococcal vaccine has the lowest rate (81%).

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		94.88	94.24	3.14	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		91.02	89.50	5.97	10.72	100.00	44,752
Meningococcus		87.32	81.22	15.86	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 barriers. In addition, there are no differentials in the expected cost of treatment that could affect vaccine uptake. The records contain socio-demographic

[15] Note that we do not consider the vaccination for Chickenpox in our analysis, as a significant portion of the eligible population acquires immunity through natural infection, which also exempts them from the vaccine mandate for this disease.

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 include 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 fragile 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.[16] 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.

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

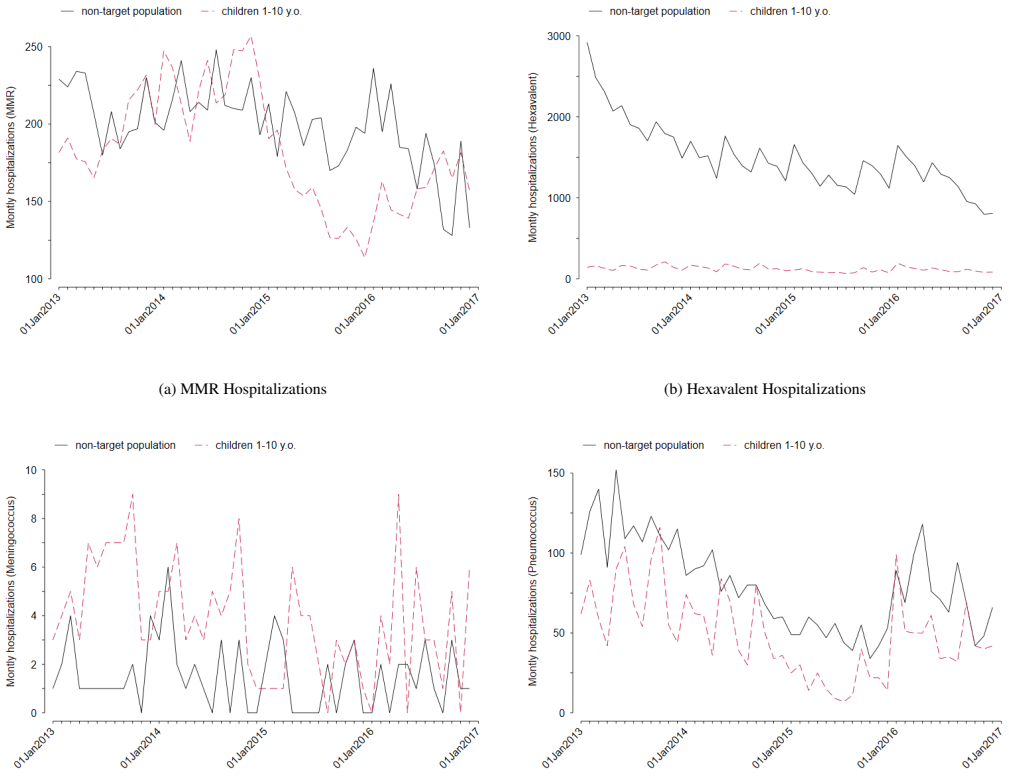
Notes: The statistics refer to 7,940 municipalities for the time period between 2013-2016 and are weighted by the municipality population size.

Figure 4 presents the monthly trends in hospitalizations among the vaccine-target population and in fragile populations that are not targeted by vaccines. In general, the two groups' trends for hexavalent, pneumococcus, and meningococcus are generally comparable. However, for the MMR vaccine,

[16] 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.

the hospitalization trends were opposite between January 2015 and January 2017, which was a period marked by several measles epidemic outbreaks.

Figure 4: Hospitalization trends (2013-2016)



(a) MMR Hospitalizations

(b) Hexavalent Hospitalizations

(c) Meningococcus Hospitalizations

(d) Pneumococcus Hospitalizations

Notes: The trend for the vaccine-target population is represented by the red dashed line, while the vaccine-untargeted fragile groups are represented by the black solid line.

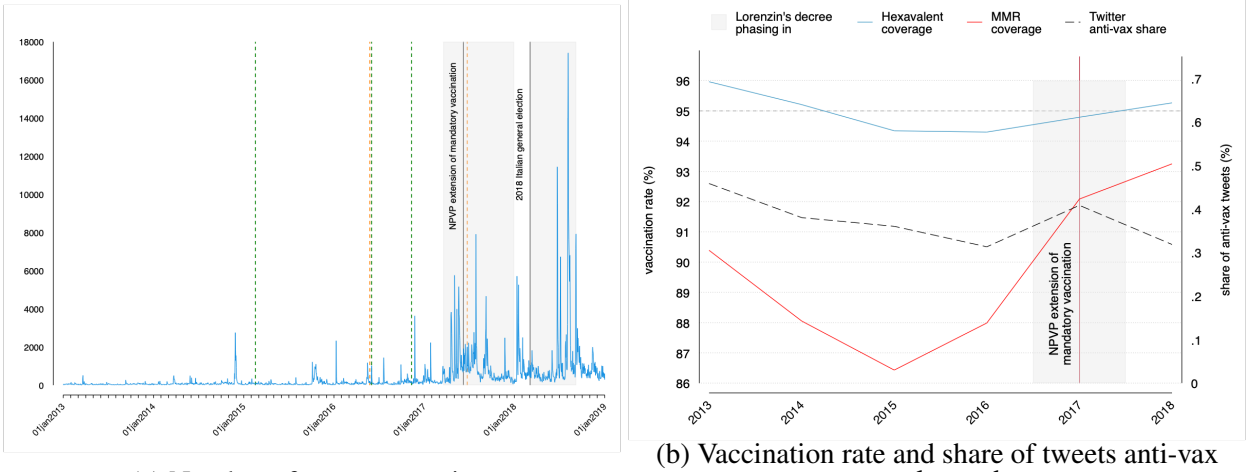
4 Twitter stances and user interactions

Many topics discussed on social media platforms 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 social 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 true for Twitter, where algorithmic amplification since 2016 in Italy has been designed to maximize users' exposure to ephemeral, captivating arguments (Huszár et al., 2022).

This 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. 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.

Twitter stances and vaccination rates. Figure 5, panel (a) shows 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 decree* on the expansion of the vaccine mandate 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 5: Number of tweets, vaccination rates and anti-vax sentiment in Italy



(a) Number of tweets over time

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

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.

This period also saw a sharp increase in measles cases in Italy. In 2017, the expansion and legal enforcement of mandatory vaccines under "Lorenzin's Law" led to a moderate increase in MMR coverage.

Figure 5, 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, as computed in Figure 3.1. Since 2012, there has been a progressive decline in the coverage of both vaccinations. On the one hand, the Rimini Court sentence in March 2012 formally confirmed the causal link between the MMR vaccine and autism, contributing to the growth of anti-vax sentiment. On the other hand, once collective vaccination efforts succeeded in eradicating certain infections, due to myopia and self-interest, individuals were more likely to skip vaccinations, leading to suboptimal uptake and fluctuations in immunization levels (Siegal et al., 2009). Coverage rates began to increase again in 2015 when the appeal court reversed the Rimini Court's sentence. This period also saw a sharp increase in measles cases in Italy. In 2017, the expansion and legal enforcement of mandatory vaccines under "Lorenzin's Law" led to a moderate increase in MMR coverage.

It is important to note that the dynamics of the average Twitter anti-vax sentiment do not necessarily correspond to 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.

The Model of Opinion Dynamics and Network Formation. We discuss and rationalize the evolution of anti-vax stances on social media in Italy based on a model of social networks opinion dynamics proposed by Baumann et al. (2020). The mechanics of the model allow us to formalize the sources of endogeneity that pervade the relationship between the spread of anti-vax opinions on social media and vaccine hesitancy, highlighting the role of the controversialness of vaccine-related topics. The following paragraphs provide a brief overview of the most relevant features of the model, while its complete representation is described in Appendix B.

In the model, there are two channels through which the spread of anti-vax content can affect vaccine hesitancy. On the one hand, exposure to extreme views can influence users' stances (*exposure effect*). On the other hand, the controversialness of a vaccine-related topic can endogenously exacer-

bate 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, our 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 assumption of randomness.

In the model, we consider a continuum of individuals i, each with their own stance on vaccinations $s_i^t = [\underline{s}, \bar{s}]$ that can range from unconditional support to hesitancy.[17] 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.

The opinion dynamics within the social network are entirely driven by time-varying interactions among agents, where each agent's i stance influences others in a monotonic manner, and this influence "flattens" at the extremes. Importantly, the influence of individual stances on others is tuned by controversialness, so that for controversial topics, even moderate opinions can capture the beliefs of their peers.

Each agent is characterized by their propensity to interact with a certain number of other agents, and the probability of interaction is driven by the concept of homophily (Bessi et al., 2016) - the tendency for individuals to associate and bond with others who have similar beliefs and characteristics. This is modeled as a decreasing function of the distance between opinions. Since we are interested in capturing the possible exchange of opinions between users, we assume that links are the medium through which information can flow. For example, if user i follows user j on Twitter, user i can see tweets 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 can reveal the presence of echo chambers, where users are surrounded by 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.

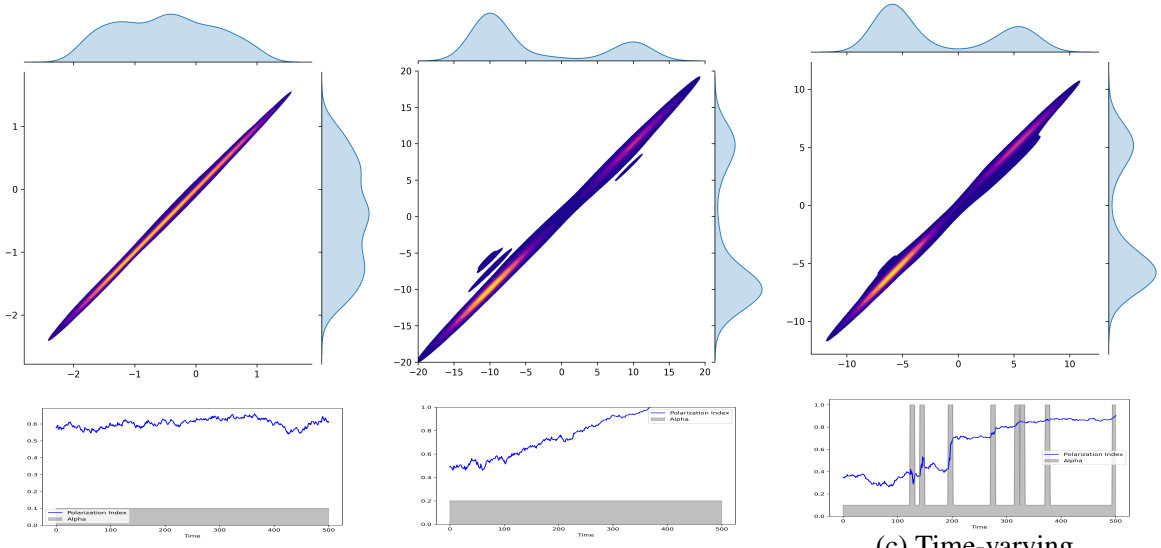
[17] Throughout the paper and the model, we assume that the stance reflects individuals' opinions on the overall utility of vaccinations. Before COVID-19, the debate on vaccines in Italy focused almost exclusively on pediatric shots, where parents make the vaccination decision on behalf of their offspring. We thus assume a one-to-one mapping between parents' and children's (perceived) utility.

Model Simulations and Data. Figure 6 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 6: Simulated distribution of stances

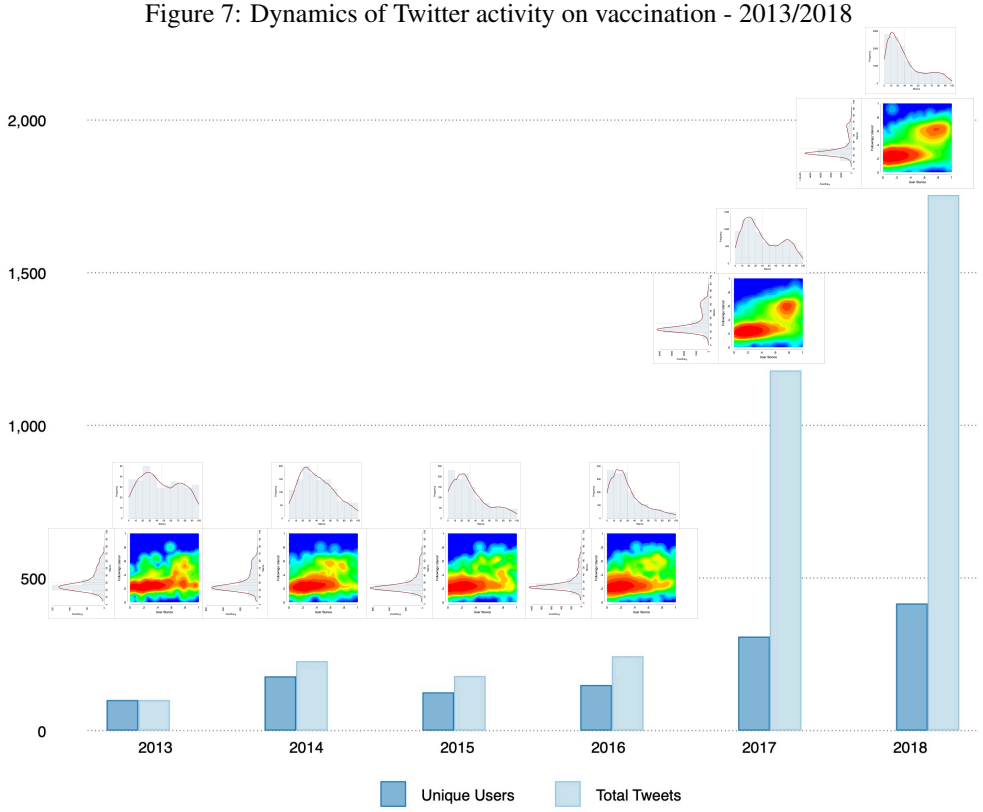


(a) Low controversialness (b) High controversialness (c) Time-varying controversialness

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.

Given the model simulations, in the figures below, we examine the evolution of the Italian vaccine debate on Twitter. In Figure 7, we show the number of unique users (blue bars, 2013=100), unique tweets (light blue bars, 2013=100), and the joint distribution of users' and their friends' observed stances. Between 2013 and 2016, the activity was relatively stable. The number of tweets per user, which can serve as a proxy for the degree of controversialness of the topic, was roughly constant, and the heatmaps do not show any evidence of polarization. From 2017, when the vaccination mandate was extended, the number of users and tweets significantly increased, and the heatmaps show the formation of echo chambers. The two opposite clusters suggest an endogenous rise in the radicalization of opinions among users. 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. In fact, as argued by Acemoglu et al. (2021), social media platforms typically interested in maximizing engagement tend to design their algorithms to create more homophilic communication patterns ("filter bubbles").

Figure 7: Dynamics of Twitter activity on vaccination - 2013/2018



Notes: The blue bars correspond to unique users, and the orange bars to unique tweets on vaccination-related topics between 2013 and 2018. 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 endogenous link formation leading to echo chambers among users and their friends 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. This set of model predictions thus motivates an identification strategy capable of estimating the empirical counterpart of the "$exposure\ effect$" of the spread of vaccine-skeptic content on vaccine hesitancy, which we will address using an Instrumental Variables approach in section 5.

5 Empirical strategy

The aim of this paper is to investigate the impact that anti-vaccination attitudes on social media can have on public health efforts to promote collective immunity. In the late-2000s, platforms like Twitter became widely used, influencing the work of journalists and the distribution of various types of content. These platforms put audiences at the center of content discovery and distribution and make them active participants in the production of news. The user-to-user sharing that is central to the social media news distribution system can lead to the offline spread of viral content, potentially affecting the real-life beliefs and behaviors of users.

One of the key assumptions underlying this paper is that Twitter anti-vaccination activity can be used as a proxy for online anti-vaccination activism. Furthermore, we assume that the extent of anti-vaccination persuasion among Twitter users living in Italian municipalities is representative of the pressure exerted by vaccine skeptic activists on parents who use other media outlets, both online and offline.

The goal of this paper is to quantify the relationship between exposure to the anti-vaccination movement - specifically, the production and dissemination of vaccine skepticism online - and vaccination rates among children. We ideally 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. Without loss of generality, we assume that there exists 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, *ceteris paribus*, 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.

We face two challenges in estimating the relationship in (2). First, as discussed in section 4, the

presence of homophily and the controversial nature of the vaccine topic can lead to the creation of echo chambers on Twitter, where users with similar vaccine stances are more likely to interact with each other. This endogeneity in the formation of social connections poses a challenge for our identification strategy. To address this problem, we use an instrumental variables (IV) approach. We leverage the local-average model of Bramoullé et al. (2009), which assumes the presence of intransitivity in the friendship network. Our theoretical framework suggests that the observed correlation in vaccine stances across users may be due to unobserved characteristics that are correlated with the endogenous choice of friends - for example, user i may form a link with user j *because* they both hold similar views on vaccines. We use the Twitter network structure and the intransitivity of network connections to construct a valid instrument. User i's "friends of friends" are not her direct connections (i.e., they are not chosen endogenously to be part of her network), but they can still have an impact on her exposure to vaccine skeptic content through their online 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 degree of indirect 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 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 was proposed by Dhrymes and Lleras-Muney (2006) for grouped data. The following paragraphs explain our approach in more detail.

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 a directed graph-

based 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 the "ego" user before engaging 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. Among this restricted group of friends, we focus only on passive users: those who do not tweet about vaccines, but only react to others' tweets (by liking, retweeting, or replying). We exclude active users: those who create original tweets about vaccines.

Within each user's ego network, for each of her passive friends who were already part of her network before she became active on Twitter in the vaccine debate, we define the *friends of friends* group. We also ensure that the link between the user and the *friends of friends* was established before their vaccine-related activity, by excluding those who created their account after the "ego" user or whose first tweet about vaccines was published after the user's first vaccine-related tweet.

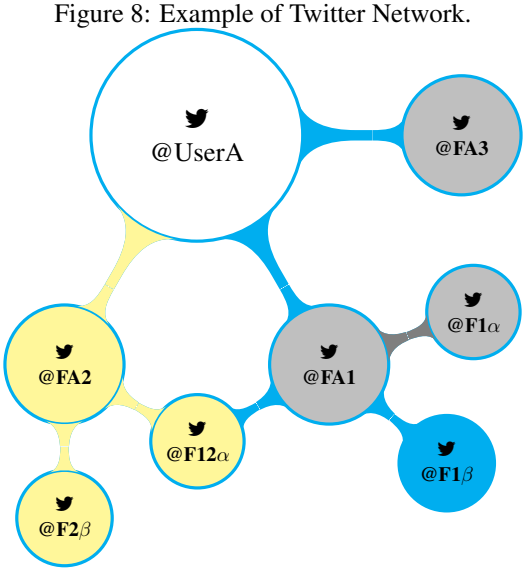
Figure 8 illustrates the selection of friends and friends of friends for a "ego" user (@UserA). An edge connects each pair of users. The gray nodes represent the links we use in our analysis, which are the first-lag neighborhood and form a 1-step ego-centered network made up of *passive friends* (e.g., @FA1, @FA3). We then define the indirect exposure of @UserA as described by the blue nodes, representing the second-lag neighborhood or *friends of friends* (e.g., @F12α and @F1β). The yellow nodes represent the links we exclude due to their potential endogeneity. They are part of the first-lag neighborhood and consist of *active friends* and their respective *friends of friends* (i.e., @FA2 and @F2β).

.

Using our sample of vaccine-related tweets, we filter for the first layer of friends and identify 65,673,913 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 a few users are central nodes in the network.

The final sample of the second layer of friends consists of approximately 2 billion nodes, corresponding to an average of 12,556 *friends of friends* per user. In our sample of users engaged in vaccine-related tweets, the median user has 469 passive friends and a median of 7,687 active *friends of friends*. The user's second layer of connections produces, on average, 142,261 tweets about vac-

Figure 8: Example of Twitter Network.



Notes: The figure plots the architecture of our network: in white *"Twitter unique user"* (@UserA), in gray *"passive friend"* (@FA1, @FA3) and *"passive friends lag 2"* (@F1α), in yellow *"active friends"* (@FA2), in blue light *"active friends lag 2"* (@F1β, @F12α).

cines (see Table 4).

Table 4: Descriptive statistics of ego network

	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 statistics refer to 80,471 geotagged unique users who tweeted on vaccines (2013-2018) for 132,190 observations.

Finally, for each set of *friends of friends*, we compute the average anti-vaccination stance. This allows us to define each user i's indirect exposure to the anti-vaccination stances of her N *friends of friends* in year t as $ffs_{it} = \frac{\sum_{\tau=1}^{N} s_\tau}{N_{it}}$, where the value ranges from 0 to 100. This measure serves as our instrumental variable in the estimation of the effect of the online anti-vaccination movement on vaccine hesitancy.

5.2 The Mixed two-stage least squares estimation

In a naive OLS version of our estimates, without taking into account endogeneity, we would measure the impact of anti-vaccination skepticism on Twitter 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: 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).[18] 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.[19] Several populist parties have raised concerns about vaccine safety (Guriev and Papaioannou, 2022, Kennedy, 2019). We also include city 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).

As previously mentioned, this simple OLS fixed effects estimation is likely to produce biased estimates due to important sources of endogeneity in our setting. Therefore, we use an Instrumental Variables approach to identify an exogenous source of variation in exposure to anti-vaccination content on Twitter. In designing our approach, we want to take advantage of the granular level of detail in our Twitter data and improve the efficiency of the first stage. However, our outcome measures are available at the municipality level. To address this, we use an 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)

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

where s_{it} is the Twitter stance on vaccines of a unique user i in year t. $\bar{fs}_{it}^{ind}$ denotes the aver-

[18] 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.2 in appendix A

[19] 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.

age Twitter stance on vaccines that the unique user i is indirectly exposed to through her *friends of friends* stances as proposed in Figure 8. Both variables range between 0 and 100, with 100 indicating the maximum level of vaccine skepticism. $\mathbf{T}_{mt}$ are m's 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 $\overline{\widehat{s}}_{it}$, which we can 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 non-targeted population, or their total cost. $\overline{\widehat{s}}_{mt}$ is the averaged instrumented 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$\times$year) controls for spatially-varying effects. All estimates are weighted by municipality population size. We correct the variance-covariance matrix throughout the analysis by bootstrapping the standard errors. In our main specification, the parameter of interest λ captures the effect of anti-vax stances on the vaccination rate.

6 Results

When presenting our results, we first review the baseline estimates of our IV strategy for vaccination rates, which we distinguish by disease type. This allows us to analyze the differential impact of vaccine skepticism on mandatory and recommended vaccines. We then present the results on hospitalizations. We look at the number of hospitalizations for vaccine-preventable diseases and the related costs, all rescaled for each 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 ($\bar{ffs}_{it}^{ind}$) on municipality characteristics such as income per capita, birth rates, public healthcare expenditure per capita, and education attainment. Our 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 almost none of the estimated correlations are significantly different from zero, supporting the assumption that our model specification identifies a source of variation unrelated to municipality characteristics.

Table 5: Balance test

	(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						
$\bar{ffs}_{it}^{ind}$	-0.0211	-0.403	0.0001	0.0001	-0.0002	0.0002
	[0.0246]	[0.442]	[0.0002]	[0.0001]	[0.0002]	[0.0002]
	110639	110639	110639	110589	110639	110639
Panel b: geolocated in municipalities different from the user's municipality						
$\bar{ffs}_{it}^{ind}$	-0.0001	-0.447	-0.0001	-0.0001	-0.00002	0.0001
	[0.0126]	[0.337]	[0.0004]	[0.0001]	[0.0001]	[0.0001]
	131003	131003	131003	130817	131003	131003
Panel c: not geolocated						
$\bar{ffs}_{it}^{ind}$	0.0037	1.001	-0.00004	-0.00001	0.0001	0.0002
	[0.0121]	[0.912]	[0.0002]	[0.00003]	[0.0001]	[0.0002]
	130977	130977	130977	130791	130977	130977
CITY and YEAR FE	✓	✓	✓	✓	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
Notes: Standard errors (in brackets) are clustered at the municipality level.

The first stage results, shown in Table 6, suggest that exposure to the vaccine-related stances of a user's *friends of friends* network is a strong predictor of the user's own anti-vaccination activity. 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 leads users to engage in anti-vaccination activism.

The results are robust under a number of model specifications which we present in subsection 6.3.

Table 6: Mixed 2SLS Individual - First stage.

	(1) s_{it} (30.31)	(2) s_{it} (30.31)	(3) s_{it} (30.31)	(4) s_{it} (30.31)
$\bar{ffs}_{it}^{ind}$ (28.77)	0.703***	0.703***	0.704***	0.704***
	[0.017]	[0.017]	[0.017]	[0.017]
N	127754	127754	127754	127754
CONTROL (Twitter)		✓		✓
CONTROL (socioeconomics)			✓	✓
CITY and YEAR FE	✓	✓	✓	✓
Reg × Year	✓	✓	✓	✓
F-stat	1765.22	1763.52	1755.84	1757.86

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

$Notes$: 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 s_{it} and $\bar{ffs}_{it}^{ind}$ in parentheses is weighted by population size.

6.1 Vaccination rates

When examining the impact of Twitter anti-vaccination activism on vaccination rates, we find no effect on the uptake of mandatory vaccines (see Table 7). The estimates for the various disease-specific compulsory vaccination rates are identical, as all of the shots are delivered in a single hexavalent vaccine. For this reason, we report the estimates for the pool of vaccines. The average coverage rates are given in brackets. The models reported in the table are the most demanding specifications, including all controls and fixed effects. The coefficient estimates are not statistically distinguishable from zero for both the OLS and M2SLS approaches.

On the other hand, when examining the effect of anti-vaccination social media activism on recommended vaccines, particularly the MMR shot, which has been most frequently linked to autism, we find a statistically significant effect on coverage rates. 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. This effect is only statistically significant in the M2SLS framework due to the measurement error and endogeneity present in the OLS estimates.

Table 7: Results of the OLS and the Second stage of the Mixed 2SLS - Vaccination rates

	(1) OLS V_{mt}	(2) Mixed 2SLS V_{mt}
Panel a: Hexavalent (94.06)		
s_{mt}	-0.0005	-0.002
	[0.002]	[0.015]
	7239	7239
Panel b: MMR (89.53)		
s_{mt}	-0.005	-0.043**
	[0.003]	[0.022]
	7238	7238
Panel c: Menigococcal (81.32)		
s_{mt}	-0.006	-0.008
	[0.007]	[0.055]
	7074	7074
Panel d: Pneumococcal (82.64)		
s_{mt}	-0.0001	-0.029
	[0.007]	[0.054]
	7066	7066
CONTROL (Twitter)	✓	✓
CONTROL (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. Estimates as well as averages of V_{mt} is weighted by the municipality population size.

6.2 Hospitalizations

We then estimate the impact of anti-vaccination social media activism on hospitalizations due to vaccine-preventable conditions. We distinguish between two groups: the target pediatric population and non-target vulnerable individuals. This distinction is important from a policy perspective. Hospitalizations for vaccine-preventable diseases among non-targeted patients measure the extent of negative spillovers or by-products of not reaching herd immunity thresholds in 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.

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 is 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 measles, mumps, or rubella (MMR), the same increase in vaccine skepticism is associated with 0.23 hospitalizations per 100 thousand residents (the baseline average is 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 is 2.96) and an excess expenditure of 366.9 euros, corresponding to a 7.7% increase.

Table A.3 in appendix shows no significant results for non-target population and target pediatric population hospitalized for diseases preventable by hexavalent, meningococcus and pneumococcus vaccines, respectively.

Table 8: Results of the OLS and the Second stage of the Mixed 2SLS - Hospitalizations .

	(1) OLS V_{mt} non-target pop.	(2) Mixed 2SLS V_{mt} non-target pop.	(3) OLS V_{mt} non-target pop.(MMR)	(4) Mixed 2SLS V_{mt} non-target pop.(MMR)	(5) OLS V_{mt} Children age 1-10 (MMR)	(6) Mixed 2SLS V_{mt} Children age 1-10 (MMR)
Panel a: Hopitalizations						
s_{mt}	0.0211	0.213*	0.0182**	0.234***	0.00712	0.145**
	[0.0159]	[0.113]	[0.00841]	[0.0601]	[0.00780]	[0.0650]
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	3331	3331	3331	3331	3331	3331
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 as well as region specific time trends. Standard errors (in brackets) are clustered at the municipality level. Estimates are weighted by the municipality population size.

6.3 Robustness checks

In addition to our main set of estimates, we conduct a series of sensitivity checks to confirm the robustness of our findings. The additional estimates are reported in Table 9, Table 10, and Table 11. The first columns report the baseline estimates, and the others report estimates that are robust to various potential threats, as outlined below.

It is important to consider the effect of Twitter's algorithmic amplification on the impact of vaccine

stances on users. In 2016, Twitter introduced an algorithmic timeline that rearranges tweets based on their relevance to the user. This feature likely amplifies the impact of indirect exposure on user stance formation. To account for this change, in our IV strategy, we introduce an interaction term between our instrumental variable, $\bar{ffs}_{it}^{ind}$, and a dummy variable ($TWalg$) that takes on a value of 1 from 2016 to account for the shift in the algorithm. The results in the second column of Table 9 show that the role of the indirect impact of friends of friends' stances increased after the algorithm change.

Additionally, Emilia-Romagna is a region that has experienced vaccine hesitancy and several outbreaks of infectious diseases affecting non-vaccinated individuals (Gori et al., 2020). In response to this, the authorities enacted the Regional Law n.19 on November 25th, 2016, which imposed a vaccine mandate before the national implementation of the "Lorenzin Law" in 2017. This mandate applied to public school and kindergarten enrollment. To analyze the interference of this policy shift, in our IV strategy, we introduce an interaction term between our instrumental variable, $\bar{ffs}_{it}^{ind}$, and a dummy variable (ER) that takes on a value of 1 from 2016 in Emilia-Romagna to account for the implementation of the regional law. The results in column 3 of Table 1 show that the Emilia-Romagna mandate did not affect how users in the region were influenced by their friends of friends' vaccine stances.

Finally, Italian populist parties have raised concerns about vaccine safety (Guriev and Papaioannou, 2022, Kennedy, 2019). In our IV strategy, we introduce an interaction term between our instrumental variable, $\bar{ffs}_{it}^{ind}$, and a dummy variable (PP) that takes on a value of 1 for the rule of populist parties at the local level (column 3). As explained in section 5, we identify municipalities where mayors are affiliated with a populist party in order to test whether the impact of online activism is likely to be enforced at the local level. If this is the case, the first stage estimate of the interaction term should be positive and statistically significant, but it is not. This suggests that our identification strategy strengthens the interpretation of our baseline results as being related to an online impact rather than a parallel effect of offline movements.

Table 9: Mixed 2SLS Individual - First stage.

	(1) Main	(2) Twitter algortithm	(3) Emilia Romagna Law	(4) Populist party	(5) Network distance
	s_{it}	s_{it}	s_{it}	s_{it}	s_{it}
	(30.33)	(30.33)	(30.33)	(30.33)	(30.33)
$\bar{ffs}_{it}^{ind}$	0.704***	0.528***	0.706***	0.691***	0.611***
	[0.017]	[0.035]	[0.017]	[0.022]	[0.021]
$\bar{ffs}_{it}^{ind}$ × TWalg		0.251***			
		[0.039]			
$\bar{ffs}_{it}^{ind}$ × ER			0.005		
			[0.0742]		
$\bar{ffs}_{it}^{ind}$ × PP				0.048	
				[0.043]	
N	127,754	127,754	127,754	127,754	127,754
CONTROLs	✓	✓	✓	✓	✓
CITY and YEAR FE	✓	✓	✓	✓	✓
REG × year	✓	✓	✓	✓	✓
F-stat	1757.86	998.690	870.815	943.98	875.82

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

$Notes$: 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. Mean values of s_{it} in parentheses are weighted by population size.

Finally, in order to address potential concerns about the exogeneity of the *friends of friends* network, we offer an alternative estimation strategy. Our network has a hierarchical structure with *unique users*, passive *friends*, and active *friends of friends*. If a user in a unique user's *friends of friends* network is linked to them through several direct friends, this can weaken the intransitivity assumption behind our IV definition. To account for this, we propose an estimation strategy that penalizes our estimates using weights defined as the inverse of the number of nodes a friend of a friend is distant from the unique user. This is given by the following equation:

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

where f_{ij} is the number of nodes between the unique user i and each friend of a friend j. As a result, this weight can be regarded as a measure of how long information will take to spread in the network. This alternative set of results (column 4) in Table 9 shows that the reweighing slightly decreases the first stage coefficient estimate, which remains comparable to the original estimate in terms of both magnitude and statistical significance.

The second stage results relative to the for alternative estimation sets for vaccination rates are reported in Table 10, and for hospitalization rates in Table 11. Columns 2-5 in both tables show that

the main results (column 1) are robust to our additional results.

Table 10: Mixed 2SLS Individual - Second stage (Vaccination rate.

	(1) Main V_{mt}	(2) Twitter algortithm V_{mt}	(3) Emilia Romagna Law V_{mt}	(4) Populist Party Law V_{mt}	(5) Network distance V_{mt}
Panel a: Hexavalent (94.06)					
s_{mt}	-0.001	-0.001	-0.003	-0.00393	-0.001
	[0.014]	[0.017]	[0.014]	[0.0157]	[0.014]
	7239	7239	7239	7239	7239
Panel b: MMR (89.53)					
s_{mt}	-0.041**	-0.039*	-0.043*	-0.0440*	-0.028*
	[0.019]	[0.024]	[0.025]	[0.0236]	[0.013]
	7238	7238	7238	7238	7238
Panel c: Menigococcus (81.32)					
s_{mt}	-0.040	-0.0113	-0.0109	-0.0127	-0.035
	[0.043]	[0.057]	[0.058]	[0.0552]	[0.039]
	7074	7074	7074	7074	7074
Panel d: Pneumococcus (82.64)					
s_{mt}	-0.010	-0.010	-0.018	-0.0386	-0.008
	[0.018]	[0.019]	[0.021]	[0.0594]	[0.010]
	7079	7079	7079	7079	7079
CONTROL (Twitter)	✓	✓	✓	✓	✓
CONTROL (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 and 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.

Table 11: Mixed 2SLS Individual - Second stage (Hospitalizations).

	(1) Main V_{mt}	(2) Twitter algortithm V_{mt}	(3) Emilia Romagna Law V_{mt}	(4) Populist party V_{mt}	(5) Network distance V_{mt}
			Non-target population		
Panel a: Hopitalizations					
	0.213*	0.231*	0.204*	0.215*	0.220*
	[0.113]	[0.121]	[0.112]	[0.112]	[0.115]
Panel b: Healthcare costs					
	731.1**	821.3**	712.8**	746.5*	794.0**
	[409.8]	[434.7]	[406.6]	[412.2]	[411.0]
			Non-target population (MMR)		
Panel c: Hopitalizations					
	0.234***	0.256***	0.233***	0.231***	0.242***
	[0.0601]	[0.0675]	[0.0596]	[0.0603]	[0.0621]
Panel d: Healthcare costs					
	722.1***	716.7***	725.1***	734.0***	743.7***
	[243.1]	[250.6]	[242.8]	[247.7]	[247.1]
			Children age 1-10 (MMR)		
Panel e: Hopitalizations					
	0.145**	0.150**	0.145**	0.146**	0.142**
	[0.0650]	[0.0664]	[0.0651]	[0.0653]	[0.0659]
Panel f: Healthcare costs					
	366.9**	428.7**	366.5**	363.6**	390.2**
	[161.1]	[171.8]	[160.9]	[163.9]	[163.7]
	3331	3331	3331	3331	3331
CONTROL (Twitter)	✓	✓	✓	✓	✓
CONTROL (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 and region-specific time trends. Standard errors (in brackets) are clustered at the municipality level. Estimates are weighted by the municipality population size.

6.4 Non-linear effects and policy implications

Based on the results discussed in the previous section, we want to understand if our findings can provide insights for policymakers and public health agencies regarding potential mitigation measures that could be implemented on social networks to effectively communicate with citizens. As noted in (Athey et al., 2022), advertising on social media became a widely used method to rapidly reach large audiences during the COVID pandemic, and has been utilized by public health organizations to convey important information and influence behavior.

To explore the potential policy implications of the vaccine-related Twitter interactions we observe in Italy, we investigate whether there are any non-linearities in the effect of lagged neighborhood

stances on user stances. Specifically, we look at whether the persuasiveness of friends-of-friends' stances varies depending on where a user falls in the stance distribution (i.e., whether they are vaccine supporters or skeptics). Additionally, on top of the network structure, we consider the role of random events related to epidemics, scientific discoveries, court sentences, policies, and news in mitigating or reinforcing the influence of vaccine stance exposure on user stances.

To speak to these issues, in the final part of our paper, 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 one). This will allow us to better understand the factors that influence vaccine attitudes among these two groups.

Table 12: Mixed 2SLS for pro-vax vs. anti-vax users - First stage.

	(1) Pro_{it} (0.495)	(2) $Anti_{it}$ (0.204)
$\bar{ffs}_{it}^{ind}$ (28.77)	-0.0076 ***	0 .0046***
	[0 .0003]	[0.0001]
N	127754	127754
CONTROL (Twitter)	✓	✓
CONTROL (socioeconomics)	✓	✓
CITY and YEAR FE	✓	✓
Reg × Year	✓	✓
F-stat	1765.22	1763.52

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

$Notes$: 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 $\bar{ffs}_{it}^{ind}$ in parentheses are weighted by population size.

Table 13: Results of the OLS and the Second stage of the Mixed 2SLS for pro-vax vs. anti-vax users - Vaccination rates

	(1) Mixed 2SLS Pro_{mt} V_{mt}	(2) Mixed 2SLS $Anti_{mt}$ V_{mt}
Panel a: Hexavalent (94.06)		
	0.4567	0.0674
	[1.4333]	[2.1973]
	7239	7239
Panel b: MMR (89.53)		
	3.9086*	-6.6162*
	[2.1978]	[3.5315]
	7238	7238
Panel c: Menigococcal (81.32)		
	0.5034	-1.6496
	[4.8856]	[8.2071]
	7074	7074
Panel d: Pneumococcal (82.64)		
	2.7584	-4.2443
	[5.3633]	[8.4350]
	7066	7066
CONTROL (Twitter)	✓	✓
CONTROL (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. Estimates as well as averages of V_{mt} is weighted by the municipality population size.

According to the magnitude of the coefficient estimates presented in Table 12, the exposure to friends-of-friends' stances has a stronger effect on pro-vax users compared to anti-vax users. This means that 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. However, in the second stage (Table 13), the actual effect on vaccine coverage is more strongly channeled through a shift of users towards anti-vax stances, rather than pro-vax ones. This suggests that the relationship between online stances and actual vaccine uptake is stronger for those at the top of the vaccine-related stance distribution. Furthermore, our results show that the only vaccination that may have been impacted by the online vaccine debate is the MMR vaccine, with no effect on the other vaccinations.

If anti-vax positions on social media are more likely to translate into actual changes in vaccine coverage, a further question is how the network can be influenced to reduce users' anti-vaccine beliefs. To explore this, 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 a 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 manually classify these online debates into four broad domains: vaccine efficacy, statements from trustful sources, politics and mandates, and allegations that vaccines are unsafe.

Table Table 14 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' neighborhood stances on regular days, and on days when specific events related to vaccines are debated on Twitter. The first column of the table 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' neighborhood stances. Exposure to anti-vax content tends to make individuals more lenient towards such stances. However, this relationship is mitigated (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). These estimates suggest that informative campaigns about vaccines may be an effective and scalable intervention for shaping public health awareness.

Table 14: User xposure 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)
$\bar{ffs}_{it}^{ind}$	0.2884***	-0.3309***	0.2295***
	[0.0693]	[0.0757]	[0.0728]
$\bar{ffs}_{it}^{ind} \times Efficacy$	-0.3425	0.3765	-0.3548
	[0.2724]	[0.2754]	[0.2961]
$\bar{ffs}_{it}^{ind} \times TrustfulSource$	-0.3136***	0.2656**	-0.3805***
	[0.0992]	[0.1127]	[0.1057]
$\bar{ffs}_{it}^{ind} \times PoliticsandMandate$	-0.1749***	0.0660	-0.3899***
	[0.0530]	[0.0408]	[0.0589]
$\bar{ffs}_{it}^{ind} \times VaccinesUnsafe$	-0.0697	0.1369	-0.0387
	[0.2292]	[0.2442]	[0.2495]
N	531352	531352	531352
User FE	✓	✓	✓
Daily 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}, $ffsPro_{it}$, and $ffsAnti_{it}$ in parentheses are weighted by population size.

7 Conclusions

Italy's pediatric vaccine coverage rates have seen significant changes between 2013 and 2018, in part due to the spread of misinformation about the safety of vaccines. This vaccine hesitancy has contributed to outbreaks of several infectious diseases, leading to the extension and legal enforcement of a mandate for a large number of pediatric vaccines in 2017. Despite the decrease in immunization, the low cost of engaging in debates has led to an increase in interaction rates among individuals. Heterogeneous and bimodal distributions of opinions are common features of controversial issues like pediatric vaccines. It is likely that the online activity of vaccine-skeptic users can affect the stances of other users, but understanding the extent to which users tend to form endogenous links with like-minded peers and the extent to which they are truly exogenously exposed to anti-vax activism is more complex.

We use Twitter data to analyze the spread of novax propaganda across time and space (municipality and year), distinguishing novax tweets using a Natural Language Processing pre-training for Italian. Exposure to like-minded peers is likely to reinforce opinions and lead to the radicalization of stances, particularly in the case of controversial issues. If segregation in the opinion space is reflected

in interactions among users, echo chambers can emerge and individuals' opinions can resonate with those of their social network contacts. To formalize these ideas, we develop a model that combines opinion dynamics, topic-related controversialness, and network formation, showing that controversial topics can foster the creation of echo chambers, leading to opinion polarization and radicalization. As predicted by the model and confirmed by our empirical data, micro-interactions among users around controversial topics can lead to transitions from a relative consensus to opinion polarization and the formation of endogenous links.

Due to the homophily and controversial nature of the vaccine topic, Twitter users tend to create echo chambers and interact with peers who have similar vaccine stances. To address the endogeneity in link formation, we employ an instrumental variables (IV) strategy. Following Bramoulle et al. (2009), we exploit exogenous variation in novax stances due to a user's exposure to the stances of her friends' friends. We treat novax stances of local Twitter users as a proxy for the penetration of the anti-vaccination movement in Italian municipalities, which is likely to affect other online social networks and the parallel offline novax propaganda. By pairing Twitter data with vaccine coverage rates, as well as rates and costs of hospitalization due to vaccine-preventable diseases, we find that exposure to novax propaganda caused a reduction in MMR immunization coverage in the period before the vaccine was mandatory. A 10 pp increase in the anti-vax stance caused a 0.43 pp drop in MMR coverage, 2.1 additional hospitalizations for every 100 thousand residents due to health complications among fragile individuals who were not targeted by the immunization, such as newborns, immunosuppressed patients, and pregnant women, as well as an excess expenditure of 7,311 euros, representing a 11% increase in relevant healthcare costs. We did not find any effect of novax propaganda on mandatory vaccines.

In addition to our main findings, we show that the exposure stance is more effective at increasing vaccine hesitancy among pro-vaccine users rather than reducing it among vaccine skeptics. However, we also show that political debates and statements from trustworthy sources can, on average, mitigate the negative effects of exposure to anti-vaccine viewpoints. This suggests that informative campaigns about vaccines may be an effective and scalable intervention for shaping public health awareness. If vaccine skeptics are resistant to changing their views, pro-vaccine individuals may be influenced by exposure to anti-vaccine viewpoints, hence providing them with accurate and reliable information can

help to counteract the skepticism effect and reinforce the support for vaccination.

In conclusion, our findings suggest that while legally enforced preventive vaccination may address the direct effects of novax propaganda on coverage rates and associated health costs, it may also lead to potential negative effects typical of controversial topics, such as the creation of echo chambers and the polarization and radicalization of opinions. Policymakers should take these potential consequences into account when implementing vaccination mandates to prevent them from backfiring once the legal enforcement of the policy is withdrawn. In fact, Baumann et al. (2021) suggest that for certain topics that overlap thematically, further evolution of controversialness can lead to the emergence of ideological states with issue alignment. In their model, ideology emerges from uncorrelated polarization simply by relaxing the assumption of topic orthogonality. While in our analysis of pediatric vaccines from 2013 to 2018, the fake news related to vaccination was limited to the anti-scientific views on the vaccine-autism causation, today the topic is no longer uncorrelated to other salient debates. The controversy surrounding the COVID-19 pandemic has given rise to an ideological state that covers a wide range of topics, including vaccines, face masks, mobility restrictions, and political views, which have the potential to hinder a wide range of deliberative processes.

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.

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.

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.

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

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.

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.

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.

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.

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). Hashtag republic.

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

Additional Tables

Table A.1: Ego-Network

User	Friend	Friend of friends	Friend of friends included
@A	if @FA1 no vaccine's tweets	and @F1α vaccine's tweets	✓
@A	if @FA1 retweets	and @F1α vaccine' tweets	✓
@A	if @FA1 reply on vaccine	and @F1α tweets on vaccine	✓
@A	if @FA1 tweets on vaccine	and @F1α tweets on vaccine	✗
@A	if @FA1 tweets on vaccine	and @F1α no tweets on vaccine	✗
@A	if @FA1 reply on vaccine	and @F1α no tweets on vaccine	✗

Table A.2: 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.

Results - Hexavalent, Meningococcus and Pneumococcus Hospitalizations

Table A.3: Results of the OLS and the Second stage of the Mixed 2SLS - Hospitalizations.

	(1)	(2)	(3)	(4)	(5)	(6)
	OLS	Mixed 2SLS	OLS	Mixed 2SLS	OLS	Mixed 2SLS
	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}	-0.00007	0.002	0.00005	0.0003	-0.002	0.009
	[0.003]	[0.016]	[0.0006]	[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	3331	3331	3331	3331	3331	3331
CONTROL (Twitter)	✓	✓	✓	✓	✓	✓
CONTROL (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 and region-specific time trends. Standard errors (in brackets) are clustered at the municipality level. Estimates are weighted by the municipality population size.

Reduced Form

Table A.4: Reduced form - Vaccination rates.

	(1)	(2)	(3)	(4)
	V_{mt}	V_{mt}	V_{mt}	V_{mt}
	Hexavalent	MMR	Meningococcus	Pneumococcus
$\bar{iffs}_{mt}^{ind}$	0.00239	-0.0297*	-0.00732	-0.0238
	[0.01]	[0.016]	[0.038]	[0.038]
N	7239	7238	7074	7066
CONTROL (Twitter)	✓	✓	✓	✓
CONTROL (socioeconomics)	✓	✓	✓	✓
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.

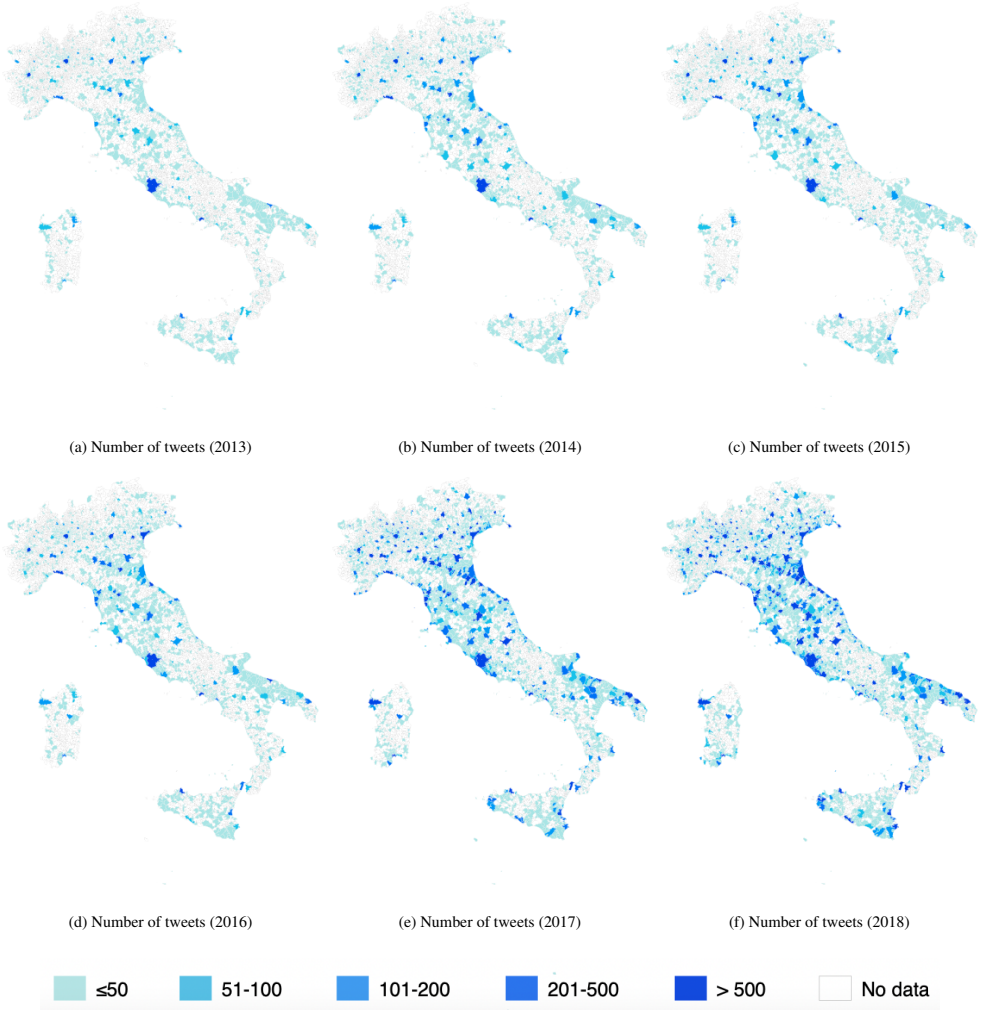
Table A.5: Reduced form - Hospitalizations.

	(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	-0.000222	-0.00549
	[0.0432]	[0.000484]	[0.00793]
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}	0.000486	0.0000780	0.00629
	[0.00970]	[0.00189]	[0.00739]
Panel f: Healthcare costs			
s_{mt}	-32.78	5.163	0.478
	[26.40]	[7.744]	[23.39]
N	5136	5136	5136
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.

Additional Figures

Figure 9: 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)

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.