Online Social Networks and Media 20 (2020) 100105

Contents lists available at ScienceDirect

Online Social Networks and Media

journal homepage: www.journals.elsevier.com/online-social-networks-and-media

Exploring childhood anti-vaccine and pro-vaccine communities on twitter – a perspective from influential users

Jieyu D. Featherstone *, George A. Barnett, Jeanette B. Ruiz, Yurong Zhuang, Benjamin J. Millam

University of California, Davis, One Shields Avenue, Kerr Hall 364, Davis, CA 95616, United States

ARTICLE INFO

Keywords:
Influencers
Twitter
Anti-vaccine
Pro-vaccine
Childhood vaccination

ABSTRACT

Anti-vaccine information online continues to deter optimum childhood vaccination coverage. Tweets from influential users about childhood vaccines are assessed to determine vaccine information on Twitter. Results indicate a well-connected anti-vaccine community where influential users widely share vaccine misinformation. Sentiment analysis finds negative tweets populate both pro- and anti-vaccine communities confirming the popularity of negative sentiment on social media. Geo-location clusters for influential users were identified. The identification of influential users and their geo-locations may provide useful information to assist with curving online vaccine misinformation and detecting areas of potential disease outbreak.

1. Introduction

The spread of anti-vaccine misinformation online threatens public health as it contributes to the increase of vaccine hesitant behaviors [29] such as vaccine refusal and scheduled vaccination delays [31]. These behaviors can lead to disease outbreaks, for example, the recent measles outbreaks in Europe and the United States [44] that claimed more than 140,000 lives in 2018 [48]. Public health officials indicate that most of the outbreaks started in communities with low vaccination rates and high numbers of vaccine-hesitant parents [5]. Unfortunately, vaccine-hesitant parents were influenced by anti-vaccine misinformation circulating on social media [11].

Vaccine-hesitant parents have shown distrust towards the medical community, public health officials, and the pharmaceutical industry. Survey results indicate that vaccine-hesitant parents tend to rely on vaccine information from the internet rather than information from health care providers or from credible health organizations [26]. Specifically, vaccine-hesitant parents are more likely to make vaccination decisions based on information shared through their social media networks, which include family members, friends, and opinion leaders (i.e. celebrities and influential online users) [7,17,50]. Vaccine information spread by online opinion leaders has been found to be mostly anti-vaccine, inaccurate, and misleading [30]. Comparatively, pro-vaccine parents tend to obtain their vaccine information from their pediatricians, health care professionals, or online and offline public health agencies [20].

Such source divergence has consequences. First, communication in anti-vaccination communities tend to occur within "echo chambers" where like-minded individuals share information consistent with their views and dismiss incongruent information [43]. This limits vaccine-hesitant communities from receiving accurate and reliable vaccine information [21] and gives rise to the use of strong and effective emotional appeals ([50]; Vosoughi et al., 2018). As such, these online communities cluster around online opinion leaders and help spread vaccine misinformation with strong anti-vaccine sentiments.

This study explores online vaccine content on Twitter by focusing on influencer tweets for three childhood vaccines (measles, mumps, rubella (MMR); tetanus, diphtheria, pertussis (Tdap); and human papillomavirus (HPV)). Twitter was selected for assessment because it is the most popular micro-blogging site where people exchange information and opinions about specific topics. The platform also allows people to follow anyone, thus, tweets have the potential to reach a wide audience base [2]. Knowing what information about vaccines influential users disseminate could help inform targeted public health communication campaigns about vaccines.

* Corresponding author.
E-mail addresses: jding@ucdavis.edu (J.D. Featherstone), gabarnett@ucdavis.edu (G.A. Barnett), jbruiz@ucdavis.edu (J.B. Ruiz), yrzhuang@ucdavis.edu (Y. Zhuang), bjmillam@ucdavis.edu (B.J. Millam).

https://doi.org/10.1016/j.osnem.2020.100105
Received 1 May 2020; Received in revised form 16 October 2020; Accepted 26 October 2020
Available online 5 November 2020

1.1. Influential users on social media

Opinion leaders on social media have been labeled "social media influencers" [8]. Social media influencers actively work to promote their posts, tend to have a lot of followers, and are relatively more central in their networks. This make-up results in greater public response to influencer posts through likes/favorites, replies, and shares. The reactions to their posts, especially shares, are indications of influence [10]. Studies have found that influencers are especially persuasive because their impact is gained by a concerted and consistent effort to promote a particular cause, position, and/or product. Their online use is unlike the average user who is much less direct in their approach to posting and sharing [1,12].

There are many methods to detect social media influencers, such as using social capital [45], information diffusion [13], and combining network typology with user behaviors [1]. The most efficient and widely used method is through information diffusion [12], where reaction to influencer posts, in this case, tweets, on specific topics are assessed. This includes the number of retweets, favorites, replies, and quotes an influencer receives in reaction to their tweets [2]. This research uses information diffusion, as previously described, to define social media influencers on Twitter for the topic of childhood vaccination.

1.2. Hypotheses and research question

Considering childhood vaccination communities on Twitter we hypothesize:

H1: Pro-vaccine and anti-vaccine Twitter communities are independent of each other in their influencers' social networks. There is no overlap between pro- and anti-vaccination media influencers on Twitter. Specifically, (H1a) members of anti-vaccine communities will be more connected with each other than with members of pro-vaccine communities; (H1b) members of pro-vaccine communities will be more connected with each other than with members of anti-vaccine communities.
H2: Twitter pro- and anti-vaccination communities discussing childhood vaccination are clustered around geo-location.
H3: Anti-vaccine Twitter influencers share more negative information about childhood vaccination than pro-vaccine Twitter influencers.

In addition to the above hypotheses, we are also interested in assessing childhood vaccine Twitter community characteristics, specifically:

RQ1: What/who are the top 20 pro- and anti-vaccine Twitter accounts about childhood vaccines?

2. Methods

This study employed three methods of data analysis: (1) social network community detection, (2) semantic network analysis (*SNA*), and (3) sentiment analysis of tweets about childhood vaccines. Tweets about childhood vaccinations were collected from July 1, 2018 to October 15, 2018. This timeframe included the peak period of a measles outbreak in Europe, the growing spread of measles in the United States, and the start of the school year when parents must indicate their children's vaccination status for school entry.

Social network community detection has been used to reveal highly interconnected groups and their functional characteristics [38]. Blondel, Guillaume, Lambiotte, and Lefebvre [6] have proposed the Louvain method to detect community based on modularity optimization. As such, this study applies the Louvain detection method via *Gephi* software [3].

SNA was used to analyze natural text [40]. Word frequency, word co-occurrence, and centrality measures reveal the position and importance of concepts within the text (Freeman, 1979; Wasserman and Faust, 1994). Potential meanings emerging in clusters were identified by analyzing relations among words based on word co-occurrence. The co-occurrence of words was set within a three-word window based on pre-established practice [15,42]. Sentiment analysis from *IBM Watson Natural Language Understanding (NLU)* was used to assess the percentages of positive, negative, and neutral tweets.

3. Data collection and analysis

Data were collected from Twitter's Premium API using Boolean search methods with the keyword's *vaccine, vaccination, vax, shot, immunization*, and *immunisation* in combination with childhood vaccine types *MMR, Tdap*, and *HPV*. Keywords were selected based on previous conclusions [49] that found vaccination information can be effectively obtained using a vaccine type (i.e. flu, influenza) with a synonym of the word *vaccine* (i.e. vaccination, immunization). Through this process, we obtained 18 possible search term combinations. The accuracy of these search methods was confirmed in a preliminary testing set. The entire archive of English language tweets within a 15-week period was collected. Data included tweets, tweets' information (i.e. number of retweets, favorites), and senders' information (i.e. geo-location, number of followers). In total, 139,433 tweets were collected and 14,735 tweets with influence were identified.

R (Version 3.4.4) was used to collect, organize, and clean the data. Using the *jsonlite* package [37], tweet data were converted and saved as text files in the Twitter API JSON format. Tweet and sender information were both extracted from the original file. Tweet information included the original tweet, retweet counts, favorite counts, reply counts. Sender information included the sender's name, location, and number of friends.

3.1. Methodology to identify influencers

To identify childhood vaccine social media influencers, we used the retweet counts, favorite counts, and reply counts of each tweet [2]. We added up the retweet, favorite, and reply counts for the same user, indicating the popularity of each user. In order to scale the data between 0 and 1, we multiplied the three counts (retweet, favorite, reply) and normalized the multiplied influence score by feature scaling: $X_{new} = \frac{X - X_{min}}{X_{max} - X_{min}}$. Zero indicates the lowest influence of tweets about vaccines by this user, whereas one indicates a user's greatest influence on vaccine tweets. The majority of users, 99.3% (N = 68,107), had a score of 0, indicating no influence. The remaining 0.7% of users (N = 482) had a score between 0 and 1. This confirms a power law distribution, with very few cases falling on the upper end of the distribution [36]. Given that Twitter influencers are defined as active users who are able to spread information and inspire others [4]and our data showed 99.3% of people had no such influence, we selected the top 0.7% as influencers targeted for further assessment. We traced these 482 influential users in order to collect their friend lists using Twitter's basic API. We were able to obtain friend lists for 420 senders[1]. 62 friend lists were unavailable due to account suspensions or private mode settings. We extracted social connections for the 420 influential senders and created an edge list of 7,731 connections where the followers are the source and the followed are the target. The edge list was imported into *Gephi* [3] for network detection. Then, the top 20 influencers for each detected community were ranked

[1] To check the robustness of feature scaling, we used Laplace-smoothing to generate the probabilities of each user being popular (a score of none zero) or not popular (zero). The Laplace-smoothing produced a list of 523 users in the 99th percentile, which included the 482 users detected using feature scaling. The 51 additional users had a low probability to be popular ($p < .01$). Therefore, we used the list of 482 users as our influencers.

based on their popularity score.

3.2. Methodology to detect community

The social networks of these 420 influencers were created and network statistics were calculated in *Gephi* [3]. Modularity is a community detection method that shows different clusters by determining the fraction of the links that fall within a given group. A threshold value of 0.4 (40% of the edges are within a given community) or above should be obtained for meaningful community detection [6]. Nodes from the same community are more densely connected with each other than with any other nodes in the network. Social networks were also visualized in *Gephi* [3]. Node size was ranked by in-degree centrality value and node color represents a detected community. An edge represents a connection between two nodes and the size of an edge represents the strength of their following (being followed and following).

Each sender's location information was extracted and summarized. User location information was extracted from their self-reported Twitter profile data. We manually checked each profile and entered the country or state names for U.S. locations after confirmation. Since our data was collected based on English language Tweets, most locations were English speaking countries. Bar charts were generated for each community to show the distribution of location based on country and state abbreviation for U.S. senders.

3.3. Methodology semantic network analysis (SNA)

Tweet text data were cleaned in R using both *tm* [19] and *qdap* packages [41], through which we removed URLs, converted to lowercase, expanded contractions, removed punctuation, and stripped whitespace. Tweet data was then saved into individual text files for analysis. Tweets were separated into different files based on each sender's community. For example, if sender A belongs to community 1, all of sender A's tweets were moved to file 1. Therefore, the numbers in the communities are equal to those of the text files.

Preprocessing procedures were conducted through *ConText* [16] for each text file individually, which provides a method for organizing large bodies of text into meaningful groupings of concepts. First, syntactically functional words (articles, conjunctions, prepositions) were removed and different forms of the same word (e.g. signify and signifies) were stemmed. The remaining text was analyzed for word frequency and word sentiment. Words that occurred with frequencies above the mean were included in the analysis in order to better represent the whole data set [25,27].

Next, semantic matrices were generated using the edited texts based on word co-occurrence. The basic network data set is an n x n matrix **S**, where n equals the number of nodes (words) in the analysis and s_{ij} is the measured relationship between nodes i and j with the node serving as the unit of analysis. Here, the nodes are identified based on the weighted frequencies of the words. Term frequency (TF) represents how much a word is mentioned in a corpus. Since our study aims to discover the most discussed themes in each community, TF is the most appropriate weighting method. The measurement of word co-occurrence is the standard for creating links between words in a semantic network. Miller [34] asserted that people can only process five to nine meaningful bits of information at a time; however, more recent studies suggest that this number may range from three to five words at a time [14]. Therefore, links were created for words that occurred either within three words of one another within each tweet or five words. A threshold of three words was picked as it generated a clearer theme and a cleaner output. The frequencies of word co-occurrence were then calculated and ranked.

The semantic networks were created using *Gephi* [3]. Words with frequencies above the mean were included in the network visualization. After importing the data, the network visualization was adjusted using the *ForceAtlas2* layout [24] to examine the spatialization between words. The size of the word label indicated how frequently the word occurred. The thickness of each link represented the weight or number of co-occurrences between two words. Closely related words were reflected in shorter distances between links. The color of each semantic network (based on senders' network community) matches the color of their sender network community color.

Network density was calculated to provide a depiction of how connected words are within the network. *Network density* is the number of connections divided by the total number of potential connections in the network (n*(n-1)/2) and can range from 0 to 1.0. Network density refers to how intertwined the word concepts are, indicating how complex discussions are surrounding an issue. *Degree* is the number of links connecting each word. *Eigenvector centrality* indicates a word's relative influence or how central it is in the network. All measures of centrality were normalized, such that degree was the values divided by the maximum possible values expressed as percentages.

3.4. Sentiment analysis

Lastly, sentiment analysis was conducted using IBM Watson *Natural Language Understanding (NLU)* [23]. *NLU* uses deep learning to extract metadata from text. The sentiment analysis feature identifies the attitudes, opinions, or feelings in the text. This analysis is not only based on the polarity of individual words but also aptly considers the sequence of the text. *NLU* is an effective sentiment analysis tool for this assessment as it was trained based on Twitter data and has reliably predicted social media posts on a variety of topics [47], with their results outperforming other competing models [9].

4. Results

4.1. Community detection

The influencers' network had a global network density of 0.05. The community detection algorithm revealed 3 distinct influencer communities (Fig. 1) with a modularity value of 0.52, indicating a meaningful community detection result. While the global network density was 0.05, the within-community densities were 0.33 (labeled orange), 0.16 (labeled green), and 0.20 (labeled blue), with an average of 0.23, 4.6 times greater than the overall density. In addition, the pairwise density was 0.14 for the orange and blue communities, 0.12 for the orange and green communities, and 0.11 for the blue and green communities. Lastly, the green community (4263 tweets) comprised the majority of the network (38.75%), followed by the orange community (5243 tweets) with 33.81% of the network, and the blue community (3981 tweets) took the rest with 27.62% of the overall network.

Influencer locations were extracted from their personal Twitter information. These were summarized to their country of origin for each community (Fig. 2). The top three countries represented in these communities were the United States (USA), United Kingdom (UK), and Ireland (IE). Both the orange and green communities were dominated by influencers from the United States. The top 5 U.S. states represented in the orange community were California, New York, Texas, Georgia, and Florida, whereas the top 5 states represented in the green community were California, New York, Texas, Washington D.C., and Maryland. The blue community was dominated by influencers from IE and the UK. The results confirm our second hypothesis that Twitter pro- and anti-vaccination communities discussing childhood vaccination are clustered around geo-location.

We ranked and identified influencers based on their popularity score and summarized their categories in each community (Table 1). Two general categories emerged: (1) organizations, including government, non-profit organizations (NGO), media, medical journals, professional organizations; and (2) individuals, including celebrities. The orange community is comprised of more individuals than the other two communities. The majority of these individuals were anti-vaccine advocates. The green community consisted of more diverse categories representing

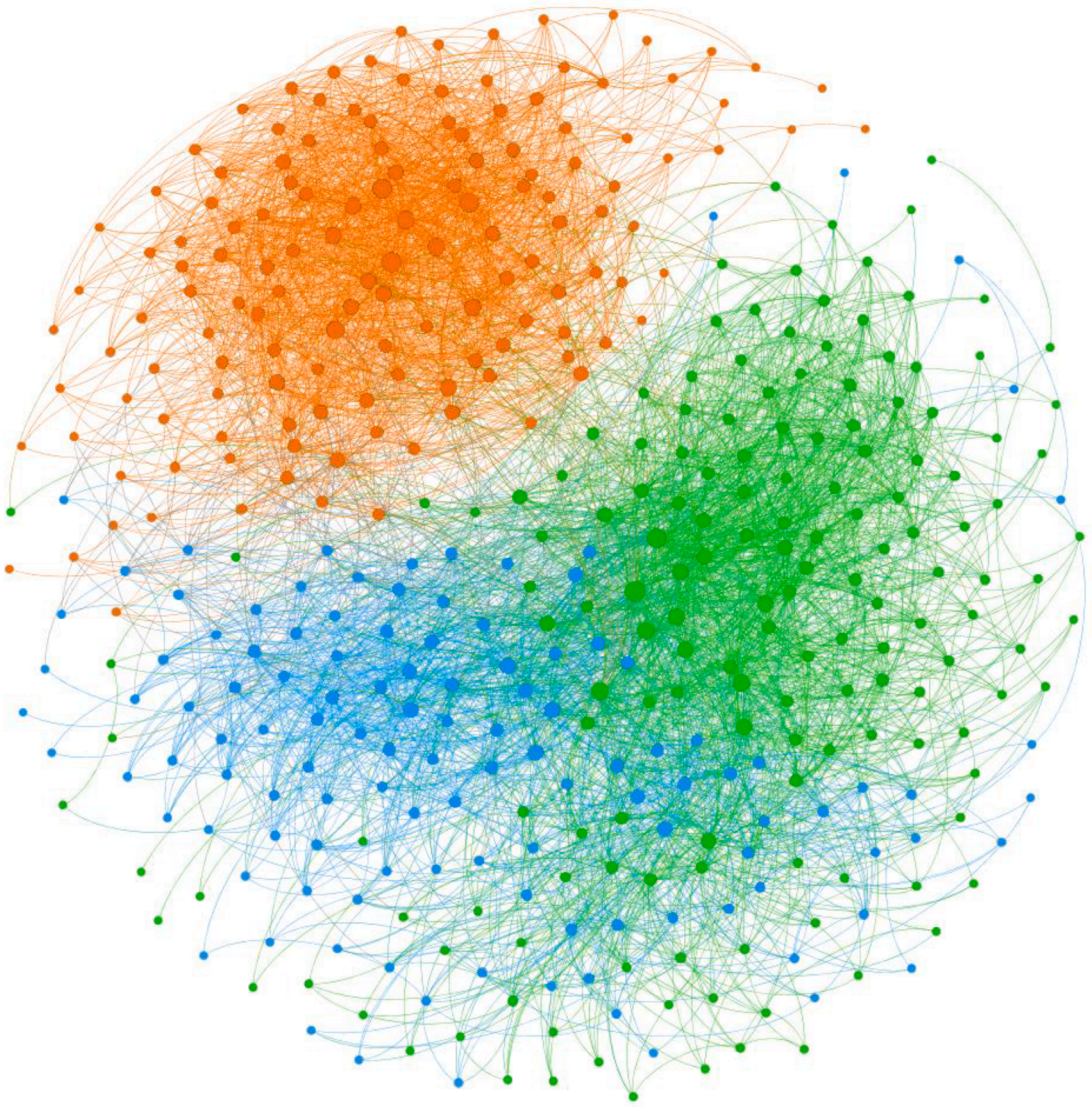

Fig. 1. Influencer social network community detection results.

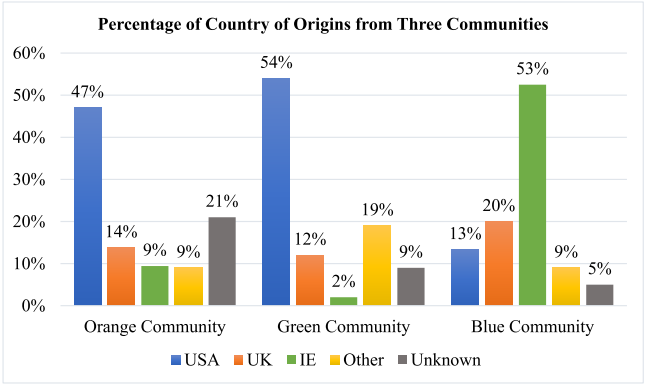

Fig. 2. Country of origin by community.

organizations and individuals that were largely promoting vaccination and disease prevention information. The blue community involved organizations run by government, professionals, and charities (NGOs) based out of Ireland who worked to promote vaccines and measures for preventing cancer.

4.2. Semantic networks

Three semantic networks were generated, one for each community. The orange community's semantic network was comprised of 103 nodes and 555 edges. The most central words in this community were *vaccination, Gardasil, MMR, get*, and *autism*. The most frequently occurring words were *vaccine, HPV, MMR, Gardasil*, and *cancer*.

The green community's semantic network consisted of 90 nodes and 563 edges. The most central words were *get, vaccination, vaccineswork, cervical, woman, adolescent*, and *protect*. The most frequently occurring words were *HPV, vaccine, get, cervical, cancer*, and *vaccineswork*.

The blue community's semantic network entailed 59 nodes and 185 edges. The most central words were *HPV, vaccine, boy, cancer, vaccination*, and *good*. The most frequently occurring words were *HPV, vaccine, boy, vaccination, cancer*, and *program*.

Table 2 indicates the top 20 central words based on both eigenvector centrality and degree for each community.

Semantic networks were constructed using the *Fruchterman-Reingold* algorithm in *Gephi* [24] with a minimum between-words tie strength of 3, which describes words that co-occurred at least 5 times within 3 words of each other. The semantic network for the orange community is presented in Fig. 3, the green community in Fig. 4, and the blue community in Fig. 5.

Table 1
Top 20 influencers and their category type by community.

	Orange Community		Green Community		Blue Community	
Rank	Name	Category Type	Name	Category Type	Name	Category Type
1	Children's Health Defense	NGO	WHO	Government	NHS	Government
2	Grace	Individual	Medscape	Media	Department of Health	Government
3	Kim Mack Rosenberg	Individual	WebMD	Media	Jen Keane	Individual
4	Jenna Jameson	Individual	National Cancer Institute	Government	HSE Ireland	Government
5	Learn the Risk	NGO	The BMJ	Medical Journal	Donal O'Keeffe	Individual
6	Erin-Health Nut News	Individual	CDC	Government	Peter Baker	Individual
7	Mairead Hilliard	Individual	Amer Acad Pediatrics	Professional Organization	PinkNews	Media
8	Chris Darnielle	Individual	The Lancet	Medical Journal	Cancer Research UK	Professional Organization
9	Marcus J	Individual	Gavi	Government	David Robert Grimes	Individual
10	Barb Loe, NVIC	NGO	Matthias Eberl	Individual	Mouth Cancer Action	NGO
11	Jonathan Irwin	Celebrity	STAT	Media	BDA	Professional Organization
12	Preventing Autism	NGO	HPV Roundtable	NGO	Newsworthyie	Media
13	Sharyl Attkisson	Celebrity	Terrence Higgins Trust	NGO	MD Anderson Cancer Center	Professional Organization
14	Dr. Sherri Tenpenny	Individual	Doc Bastard	Individual	Iver Hanrahan	Individual
15	VacTruth.com	NGO	Eric Cadesky, MD	Individual	Jabs for the Boys	NGO
16	Physicians for Info	NGO	Meenakshi Bewtra	Individual	Independent.ie	Media
17	IFICA	NGO	Jason Mendelsohn	Individual	Throat Cancer Foundation	NGO
18	No compulsory vaccines	NGO	CDC STD	Government	Jo's Trust	NGO
19	Health Impact News	Media	Debunking Denialism	Individual	Sakura No Seirei	Individual
20	Chris Collins	Individual	Slate	Media	Oral Health Foundation	NGO

Table 2
Summary output of the semantic network analysis (SNA).

Rank	Word	Degree	Eigencentrality	Rank	Word	Degree	Eigencentrality
Orange Community Semantics							
1	vaccination	30	1.00	11	risk	17	0.56
2	Gardasil	24	0.83	12	fraud	16	0.50
3	MMR	24	0.92	13	HPV	16	0.55
4	get	22	0.84	14	suffer	16	0.48
5	autism	21	0.77	15	seizure	15	0.48
6	safety	21	0.90	16	increase	14	0.53
7	death	19	0.75	17	injury	14	0.53
8	measles	18	0.47	18	Merck	13	0.53
9	cancer	17	0.72	19	Boy	12	0.49
10	girl	17	0.70	20	die	12	0.57
Green Community Semantics							
1	get	33	1	11	study	20	0.61
2	vaccineswork	29	0.99	12	child	19	0.58
3	vaccination	29	0.68	13	parent	19	0.58
4	cervical	23	0.64	14	MMR	19	0.54
5	woman	23	0.71	15	preteen	18	0.47
6	adolescent	22	0.73	16	girl	17	0.58
7	protect	22	0.67	17	increase	17	0.64
8	cancer	21	0.75	18	prevent	17	0.62
9	measles	20	0.72	19	cause	16	0.54
10	risk	20	0.64	20	rate	16	0.52
Blue Community Semantics							
1	HPV	38	1	11	safe	6	0.27
2	vaccine	27	0.73	12	recommend	6	0.26
3	boy	22	0.54	13	woman	6	0.24
4	cancer	19	0.60	14	prove	5	0.30
5	vaccination	15	0.36	15	advise	5	0.27
6	good	9	0.46	16	Ireland	5	0.27
7	vaccineswork	9	0.45	17	girl	5	0.24
8	prevent	9	0.38	18	important	5	0.22
9	cervical	8	0.35	19	introduce	5	0.22
10	official	6	0.24	20	get	5	0.17

The themes for each community were based on the top co-occurrences in their respective semantic networks (Table 3). The general danger of childhood vaccines was the dominant theme of the orange community. Vaccine promotion, specifically for HPV and MMR vaccines as preventatives, was the central theme of the green community. The importance of vaccinating boys against HPV was the principal theme of the blue community.

In order to identify pro-vaccine and anti-vaccine communities, we performed descriptive analysis on the top 20 accounts in each community and their semantic networks. Both results showed the orange community as more likely to be an anti-vaccine community as it was comprised of more anti-vaccine organizations, individuals, and tweets. Both the blue and green communities reflected more of a pro-vaccine stance.

4.3. Sentiment analysis

Chi-square tests were conducted to evaluate the significance of proportional difference among the sentiment for each community. The orange community was the most negative in sentiment ($p < .05$), the

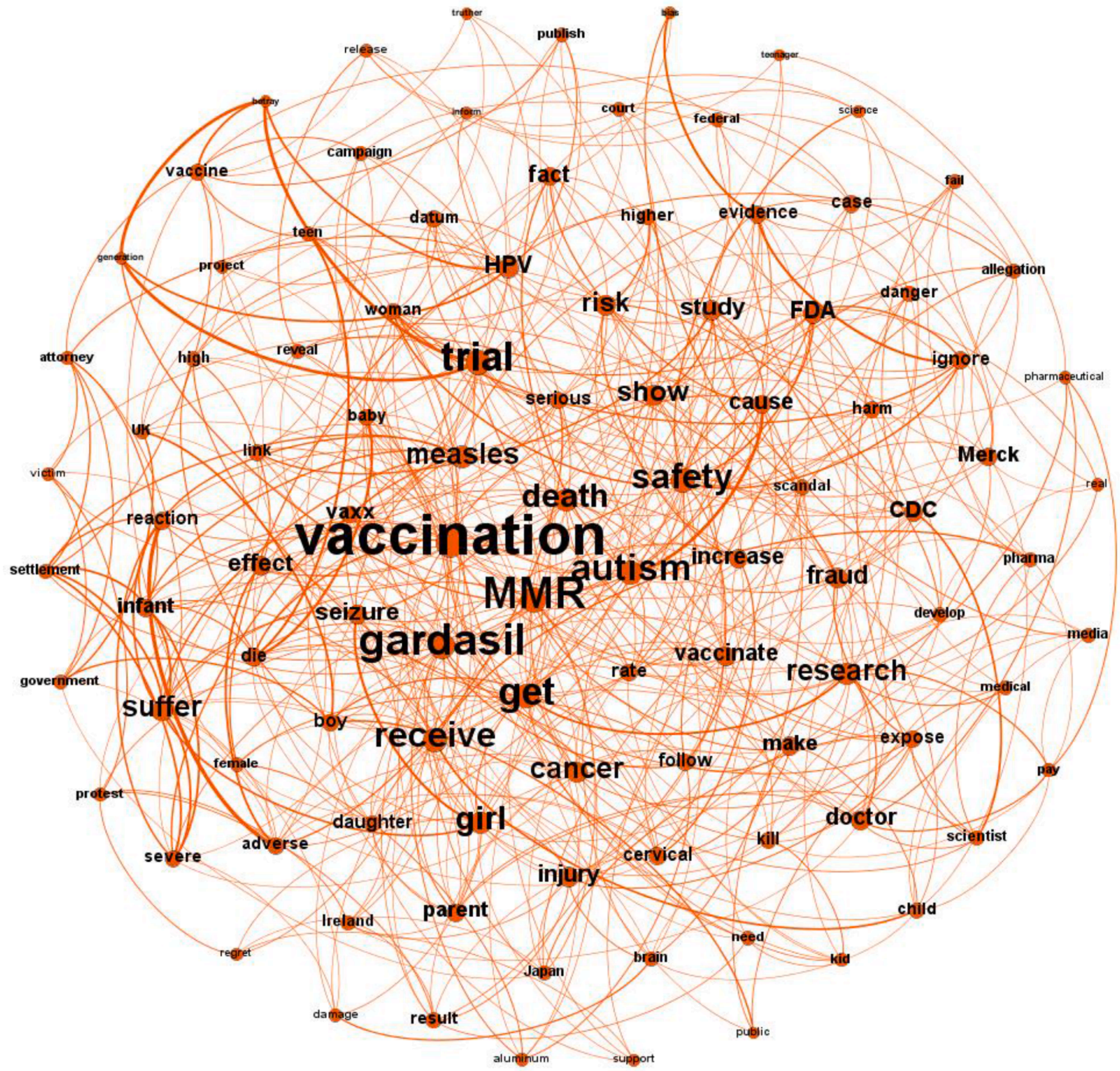

Fig. 3. Semantic network of the orange community. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

blue community revealed the most positive sentiment ($p < .05$), and the green community had the most neutral tweets among the three ($p < .05$) (Table 4). In order to check the robustness of IBM Watson's sentiment analysis, we conducted the same analysis using SentiStrength and found no significant differences between the two results ($p < .01$). Results support our third hypothesis and indicate that the anti-vaccine community spread more negative tweets than the pro-vaccine communities, while the pro-vaccine communities spread more positive tweets than the anti-vaccine community.

Within-community and pair-wise community densities as well as semantic network analysis describe the orange community as anti-vaccine and the blue and green communities as pro-vaccine, partially supporting our first hypothesis. The pro-vaccine and anti-vaccine communities are independent of each other, with the anti-vaccine community more densely connected with itself than the pro-vaccine communities. However, as indicated by the pair-wise densities, the pro-vaccine communities are less connected with each other than the anti-vaccine community is with the two pro-vaccine communities.

5. Discussion

5.1. Communities and discourses

This study utilized social media influencers to assess childhood vaccine information on Twitter. Social influencers were shown to form independent vaccine communities within the Twitter-sphere that were either pro- or anti-vaccine. Specifically, the anti-vaccine community we assessed was more connected within than either of the two pro-vaccine communities and was also more connected across all three communities. These results indicate that the anti-vaccine community was effectively connected within and received information from pro-vaccine communities. One possible explanation is that anti-vaccine influencers follow Twitter accounts related to vaccination in general. This tactic may provide information that can be used in order to attack claims made by pro-vaccine groups. These targeted attacks may help to form and strengthen anti-vaccine beliefs. Anti-vaccine groups have utilized vaccine promotion content from government sources (i.e. CDC) to counterargue these with popular anti-vaccine conspiracy theories [22,35]. On the other hand, the relatively less connections within the pro-vaccine communities reveals the limits of their reach and may be detrimental in promoting correct vaccine information to anti-vaccine communities.

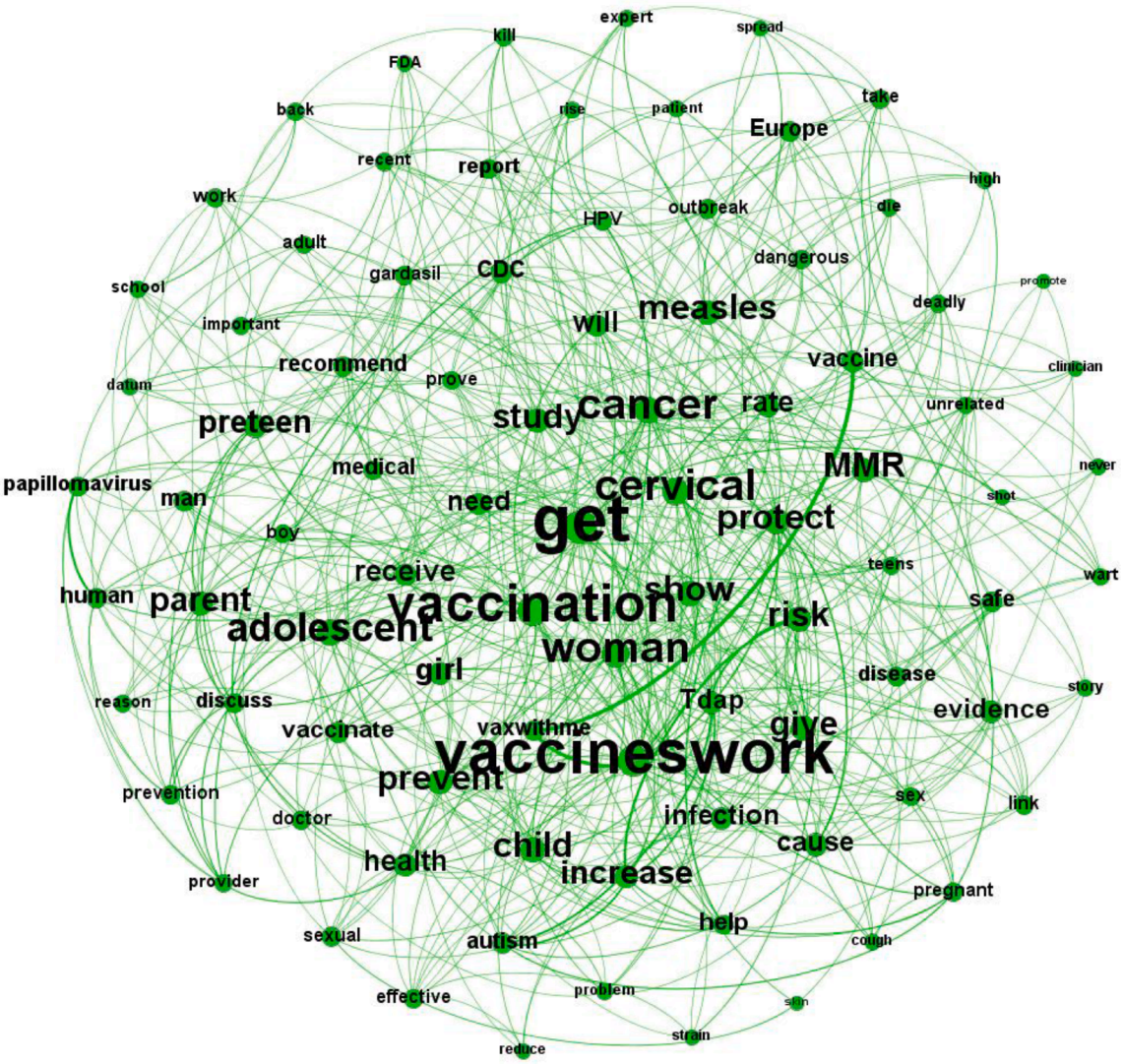

Fig. 4. Semantic network of the green community. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Descriptive analysis of the top 20 influencers showed community formation based on vaccine stances and geo-locations. Anti-vaccine tweets were more likely to be from new emerging news websites, such as personal websites and NGOs initiated by parents. This points to a growing use of personal narratives as effective persuasive tools. Whereas pro-vaccine tweets were sourced from traditional mainstream media, such as government, newspapers, and magazines [33]. These sources tend to focus on facts which may be less moving than the personal narrative. Unsurprisingly, geo-location information centered the communities across English-language speaking countries. However, the geo-location clusters in each community have the potential to provide valuable information for improved vaccination coverage monitoring and may possibly assist with disease outbreak predictions.

Influencers were defined by the popularity of their vaccine-related tweets and followed a power law distribution. Findings are similar to another study that showed a very small number of users had the most influence on MMR vaccine information on Twitter [39]. These results indicate that influential users dominate vaccine-related information on Twitter, be it anti-vaccine or pro-vaccine. This further illustrates the importance of studying influential users on the topic of vaccination.

Semantic network analysis results confirmed previous results that indicate anti-vaccine posts contain vaccine misinformation in addition to strong anti-vaccine sentiment [28,32,35]. The anti-vaccine community in our analysis (orange) revealed a circulation of false and inaccurate vaccine information. For example, “vaccine injury”, “vaccine autism”, and “vaccine fraud”. Identifying anti-vaccine communities through social media influencers can help target vaccine information efforts. Using influencers to identify these anti-vaccine communities may help to more efficiently detect vaccine misinformation online. Moreover, by locating anti-vaccine social media influencers, public health professionals can more accurately target anti-vaccine communities in order to educate them about vaccine safety, answer vaccine concerns, and monitor the spread of anti-vaccine misinformation.

Central to all three communities in our assessment was the mention of HPV vaccines. This is likely due to the recent and active promotion of HPV vaccination. All three communities mentioned the targeted population for HPV vaccination, such as *boy, teen, adolescent,* and *girl*. Unsurprisingly, the anti-vaccine community emphasized the harm, danger, and safety of the HPV vaccine, whereas the pro-vaccine communities stressed the benefits of preventing cervical cancer through vaccination. These results are similar to previous studies that found HPV vaccination at the center of vaccine discussions online, specifically on Twitter, with

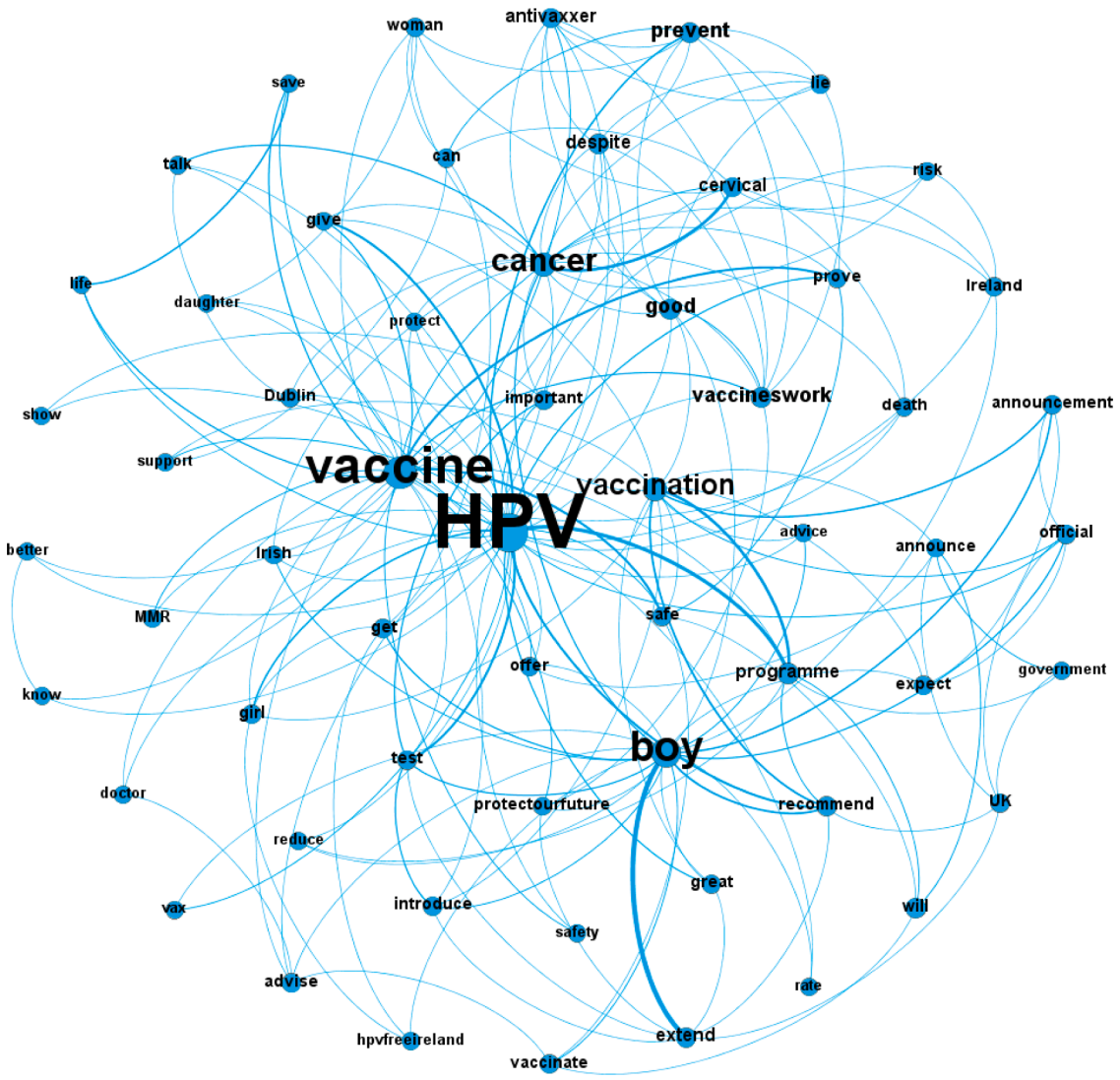

Fig. 5. Semantic network of the blue community.

tweets encouraging the vaccination of boys against HPV from pro-vaccines accounts and the spread of conspiracy rumors about vaccines from anti-vaccine Twitter users [18].

5.2. Sentiments

It is not surprising to see that negative tweets were overwhelmingly present in the anti-vaccine community. However, there are also more negative tweets than positive ones in the pro-vaccine communities. One explanation might be that negative emotions get spread more rapidly and widely than positive ones [46]. If popular tweets tend to be more negative in sentiment, it is not surprising to see more negative tweets in general across the various communities. Another explanation is that vaccinations treat diseases, a concept with negative association and general sentiment. Still, the percentages of positive-sentiment tweets were higher in the pro-vaccine communities than in the anti-vaccine community.

6. Limitations

This study is not without limitations. First, we only assessed tweets from a specific period of time. A longitudinal analysis can provide additional insights into changes that are likely to occur over time. While the time frame we assessed did include a period with increased measles spikes as well as the beginning of the U.S. school year, we're unable to compare network and community differences during different time-frames. Second, this study did not classify misinformation tweets from each community. Thus, we cannot conclude that the anti-vaccine community is *the* only community spreading misinformation. Although, knowing that the anti-vaccine community spread more misinformation than the other communities could prove a good source of investigation for future misinformation detection studies. Last, we only identified 20 influencers from each community, so we cannot generalize this finding to whole communities or to other communities. Other communities may have different influencer types.

7. Conclusion

Through the identification of social media influencers, we found an anti-vaccine community that was more connected than pro-vaccine communities. The anti-vaccine network circulated false and inaccurate information about vaccines through influential individuals such as celebrities. The MMR vaccine-autism link continues to be a topic of discussion among anti-vaccine networks. HPV vaccines also dominate the discussion among both pro- and anti-vaccine Twitter communities. In terms of sentiment, both pro- and anti-vaccine communities reflected

Table 3
Summary output of semantic top co-occurrences in each community.

Community Color	Themes	Top Co-Occurrences		Co-Occurrence Count
Orange	Danger of childhood vaccines	HPV	vaccine	2419
		MMR	vaccine	608
		cervical	cancer	409
		vaccine	safety	277
		vaccine	cause	274
		vaccine	injury	210
		vaccine	autism	200
		HPV	safety	196
Green	Get vaccines. Vaccines prevent and protect. HPV vaccines prevent cancer.	HPV	vaccine	1158
		cervical	cancer	259
		get	vaccine	250
		get	HPV	222
		MMR	vaccine	155
		prevent	cancer	146
		vaccine	protect	142
		vaccine	prevent	140
Blue	The program to extend HPV vaccine to boys. Prevent cervical cancer.	HPV	vaccine	83
		HPV	boy	37
		HPV	vaccination	31
		vaccination	boy	26
		extend	boy	19
		HPV	program	16
		cervical	cancer	15
		vaccine	prevent	13

Table 4
Sentiment analysis results by community.

	Sentiment Labels Frequency (%)		
Community (N)	Negative	Neutral	Positive
Orange (5243)	3108 (59.28%)	1677 (32%)	458 (8.7%)
Green (4263)	1925 (45.16%)	1880 (44.10%)	458 (10.74%)
Blue (3981)	2084 (52.42%)	1190 (29.9%)	704 (17.68%)

the popularity of negative-sentiment tweets. For future studies, using social media influencers to identify anti-vaccine communities may be an effective strategy for targeting anti-vaccine misinformation spread online.

CRediT authorship contribution statement

Jieyu D. Featherstone: Conceptualization, Methodology, Validation, Formal analysis, Data curation, Writing - original draft, Visualization. **George A. Barnett:** Conceptualization, Methodology, Validation, Resources, Writing - review & editing, Supervision, Project administration, Funding acquisition. **Jeanette B. Ruiz:** Conceptualization, Validation, Writing - review & editing. **Yurong Zhuang:** Software, Investigation, Resources, Data curation. **Benjamin J. Millam:** Software, Investigation, Resources, Data curation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Z.Z. Alp, S.G. Oguducu, Identifying topical influencers on twitter based on user behavior and network topology, Knowl.-Based Syst. 141 (2018) 211–221, https://doi.org/10.1016/j.knosys.2017.11.021.

[2] R. Bacha, T.T. Zin, A survey on influence and information diffusion in twitter using big data analytics, in: Paper presented at the International Conference on Big Data Analysis and Deep Learning Applications. ICBDL 2018, Singarpore, 2018.

[3] M. Bastian, S. Heyman, M. Jacomy, Gephi: an open source software for exploring and manipulating networks, in: Paper Presented at the International AAAI Conference on Weblogs and Social Media, San Jose, CA, 2009.

[4] L. Ben Jabeur, L. Tamine, M. Boughanem, Active microbloggers: identifying influencers, leaders and discussers in microblogging networks, String Process. Inf. Retrieval 7608 (2012) 111–117.

[5] O. Benecke, S.E. DeYoung, Anti-vaccine decision-making and measles resurgence in the United States, Global Pediatr. Health 6 (2019) 1–5, https://doi.org/10.1177/2333794X19862949.

[6] V.D. Blondel, J.L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, J. Statistic. Mech.-Theory Exper. 10 (2008) P10008.

[7] E.K. Brunson, The impact of social networks on parents' vaccination decisions, Pediatrics 131 (5) (2013) 1397–1404, https://doi.org/10.1542/peds.2012-2452.

[8] A. Burke-Garcia, Opinion Leaders for Health: Formative Research with Bloggers about Health Information Dissemination, George Mason University, Fairfax, VA, 2017.

[9] M. Canonico, L.D. Russis, A comparison and critique of natural language understanding tools, in: Paper Presented at the Cloud Computing 2018: The Ninth International Conference on Cloud Computing, GRIDs, and Virtualization, Barcelona, Spain, 2018.

[10] R. Cappelletti, N. Sastry, IARank: ranking users on twitter in near real-time, based on their information amplification potential, in: Proceedings of the 2012 Ase International Conference on Social Informatics (Socialinformatics 2012), 2012, pp. 70–77, https://doi.org/10.1109/SocialInformatics.2012.82.

[11] CDC, U.S. Public Health Response to the Measles Outbreak, Centers for Disease Control and Prevention Washington, Watshington, 2019. Retrieved from, https://www.cdc.gov/washington/testimony/2019/t20190227.htm.

[12] M. Cha, H. Haddadi, F. Benevenuto, K. Gummadi, Measuring user influence in twitter: the million follower fallacy, in: Paper presented at the the Fourth International AAAI Conference on Weblogs and Social Media, Washington, DC, 2010.

[13] W. Chen, C. Wang, Y. Wang, Scalable influence maximization for prevalent viral marketing in large-scale social networks, in: Paper presented at the the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2010.

[14] N. Cowan, Working Memory Capacity: Classic Edition, Routledge, 2016.

[15] J.A. Danowski, Network analysis of message content, in: W.D. Richard Jr., G. A. Barnett (Eds.), Progress in Communication Sciences: Advances in Communication Networks, 12, Norwood, NJ: Ablex, 1993, pp. 198–221.

[16] J. Diesner, ConText: software for the integrated analysis of text data and network, in: Paper Presented at the International Communication Association (ICA), Seattle, WA, USA, 2014.

[17] E. Dube, M. Vivion, N.E. MacDonald, Vaccine hesitancy, vaccine refusal and the anti-vaccine movement: influence, impact and implications, Expert Rev. Vaccines 14 (1) (2015) 99–117, https://doi.org/10.1586/14760584.2015.964212.

[18] J.D. Featherstone, J.B. Ruiz, G.A. Barnett, Exploring childhood vaccination themes on twitter: a semantic network analysis, in: Paper presented at the The 70th Annual Conference of the International Communication Association, Gold Coast, Australia, 2020.

[19] Feinerer, I., Hornik, K., and Meyer, D. (2008). Text mining infrastructure in R (Version 2008-03-18). Retrieved from https://www.jstatsoft.org/v025/i05.

[20] G.L. Freed, S.J. Clark, A.T. Butchart, D.C. Singer, M.M. Davis, Sources and perceived credibility of vaccine-safety information for parents, Pediatrics 127 (2011) 107–112, https://doi.org/10.1542/peds.2010-1722P.

[21] R. Getman, M. Helmi, H. Roberts, A. Yansane, D. Cutler, B. Seymour, Vaccine hesitancy and online information: the influence of digital networks, Health Educat. Behav. 45 (4) (2018) 599–606, https://doi.org/10.1177/1090198117739673.

[22] B.L. Hoffman, E.M. Felter, K.H. Chu, A. Shensa, C. Hermann, T. Wolynn, B. A. Primack, It's not all about autism: the emerging landscape of anti-vaccination sentiment on facebook, Vaccine 37 (16) (2019) 2216–2223.

[23] IBM. (2019). IBM cloud API docs: natural language understanding. Retrieved from https://cloud.ibm.com/apidocs/natural-language-understanding/natural-language-understanding?code=python#sentiment.

[24] M. Jacomy, T. Venturini, S. Heymann, M. Bastian, ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software, PLoS One 9 (6) (2014) e98679.

[25] K. Jiang, B.N. Anderson, P.C. Ronald, G.A. Barnett, Semantic network analysis reveals opposing online representations of the search term "GMO", Global Challenges 2 (1) (2018) 1–8.

[26] A.M. Jones, S.B. Omer, R.A. Bednarczyk, N.A. Halsey, L.H. Moulton, D.A. Salmon, Parents' source of vaccine information and impact on vaccine attitudes, beliefs, and nonmedical exemptions, Adv. Prev. Med. 2012 (2012) 1–8, https://doi.org/10.1155/2012/932741.

[27] G.J. Kang, S.R. Ewing-Nelson, L. Mackey, J.T. Schlitt, A. Marathe, K.M. Abbas, S. Swarup, Semantic network analysis of vaccine sentiment in online social media, Vaccine 35 (29) (2017) 3621–3638, https://doi.org/10.1016/j.vaccine.2017.05.052.

[28] A. Kata, Anti-vaccine activists, Web 2.0, and the postmodern paradigm - An overview of tactics and tropes used online by the anti-vaccination movement, Vaccine 30 (25) (2012) 3778–3789.

[29] H.J. Larson, The biggest pandemic risk? Viral misinformation, Nature 562 (2018) 309.

[30] J. Leask, H.W. Willaby, J. Kaufman, The big picture in addressing vaccine hesitancy, Human Vaccines Immunotherapeutics 10 (9) (2014) 2600–2602, https://doi.org/10.4161/hv.29725.

[31] N.E. MacDonald, S.W.G.V. Hesitancy, Vaccine hesitancy: definition, scope and determinants, Vaccine 33 (34) (2015) 4161–4164, https://doi.org/10.1016/j.vaccine.2015.04.036.
[32] K. Madden, X.L. Nan, R. Briones, L. Waks, Sorting through search results: a content analysis of HPV vaccine information online, Vaccine 30 (25) (2012) 3741–3746.
[33] C.Z. Meadows, L. Tang, W. Liu, Twitter message types, health beliefs, and vaccine attitudes during the 2015 measles outbreak in California, Am. J. Infect. Control 47 (11) (2019) 1314–1318, https://doi.org/10.1016/j.ajic.2019.05.007.
[34] G.A. Miller, The magical number seven plus or minus two: some limits on our capacity for processing information, Psychol. Rev. 63 (2) (1956) 81–97. https://doi.org/10.1037/h0043158.
[35] M.B. Moran, M. Lucas, K. Everhart, A. Morgan, E. Prickett, What makes anti-vaccine websites persuasive? A content analysis of techniques used by anti-vaccine websites to engender anti-vaccine sentiment, J. Commun. Healthcare 9 (3) (2016) 151–163.
[36] M.E.J. Newman, Power laws, pareto distributions and Zipf's law, Contemporary Phys. 46 (5) (2005) 323–351.
[37] Ooms, J. (2014). The jsonlite package: A practical and consistent mapping between json data and r objects. *arXiv preprint arXiv:1403.2805*.
[38] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, P. Spyridonos, Community detection in social media performance and application considerations, Data Mining Knowl. Discovery 24 (3) (2012) 515–554, https://doi.org/10.1007/s10618-011-0224-z.
[39] J. Radzikowski, A. Stefanidis, K. Jacobsen, A. Croitoru, A. Crooks, P. Delamater, The measles vaccination narrative in twitter: a quantitative analysis, JMIR Public Health Surveil 2 (1) (2016) e1.
[40] R.E. Rice, J.A. Danowski, Is it really just like a fancy answering machine? Comparing semantic networks of different types of voice mail users, Int. J. Bus. Commun. 30 (4) (1993) 369–397.
[41] Rinker, T. W. (2019). Quantitative discourse analysis package. Buffalo, New York. Retrieved from http://github.com/trinker/qdap.
[42] J.B. Ruiz, G.A. Barnett, Exploring the presentation of HPV information online: a semantic network analysis of websites, Vaccine 33 (29) (2015) 3354–3359.
[43] A.L. Schmidt, F. Zollo, A. Scala, C. Betsch, W. Quattrociocchi, Polarization of the vaccination debate on facebook, Vaccine 36 (25) (2018) 3606–3612, https://doi.org/10.1016/j.vaccine.2018.05.040.
[44] L. Selim, Measles explained: what's behind the recent outbreaks? UNICEF 10 (2019) P10008.
[45] K. Subbian, D. Sharma, Z. Wen, J. Srivastava, Finding influencers in networks using social capital, Soc. Network Anal. Mining 4 (1) (2014), https://doi.org/10.1007/s13278-014-0219-z. ARTN 219.
[46] S. Tsugawa, H. Osaki, Negative messages spread rapidly and widely on social media, in: Paper presented at the Conference on Online Social Networks, 2015.
[47] Vergara, S., El-Khouly, M., El Tantawi, M., Marla, S., and Sri, L. (2017). *Building cognitive applications with IBM watson services: volume 7 natural language understanding* (Vol. 7, pp. 110).
[48] WHO, More than 140,000 die from measles as cases surge worldwide, World Health Org. (2019).
[49] R.M. Wolfe, L.K. Sharp, Vaccination or immunization? The impact of search term on the Internet, J. Health Commun. 10 (2005) 537–551.
[50] X.Y. Yuan, A. Crooks, From cyber space opinion leaders and the diffusion of anti-vaccine extremism to physical space disease outbreaks, Social, Cult. Behav. Model. 10354 (2017) 114–119, https://doi.org/10.1007/978-3-319-60240-0_14.