Identifying Vaccine-Hesitant Communities on Twitter and their Geolocations: A Network Approach

Jeanette B. Ruiz
University of California, Davis
jbruiz@ucdavis.edu

Jade D. Featherstone
University of California, Davis
jding@ucdavis.edu

George A. Barnett
University of California, Davis
gabarnett@ucdavis.edu

Abstract

Vaccine misinformation online may contribute to the increase of anti-vaccine sentiment and vaccine-hesitant behaviors. Social network data was used to identify Twitter vaccine influencers, their online twitter communities, and their geolocations to determine pro-vaccine and vaccine-hesitant online communities. We explored 139,433 tweets and identified 420 vaccine Twitter influencers—opinion leaders and assessed 13,487 of their tweets and 7,731 of their connections. Semantic network analysis was employed to determine twitter conversation themes. Results suggest that locating social media influencers is an efficient way to identify and target vaccine-hesitant communities online. We discuss the implications of using this process for public health education and disease management.

1. Introduction

Vaccine-hesitant parents have been shown to favor information from the internet rather than from health care providers or credible health organizations [1]. Similarly, vaccine-hesitant parents rely on information shared on social media platforms, specifically from family, friends, and social media influencers or opinion leaders, to inform vaccine decisions [2]. Unfortunately, online information and information coming from opinion leaders have been found to often provide inaccurate and misleading information [3]. Outside of the vaccine-hesitant community, information is driven by scientific evidence but this evidence is often misused in vaccine-hesitant communities [4]. Furthermore, research finds that if current trends continue, anti-vaccine views on social media will dominate the online vaccine discussion within 10-years [5]. This trend is based on data that shows that while online antivaccine groups have fewer followers than online pro-vaccine groups, the antivaccine groups are more numerous, more connected to undecided groups, and growing more connected at a faster pace.

The online antivaccine groups are more effectively connected at both a global and local level, unlike pro-vaccine groups who are less locally focused [5]. The antivaccine influencers then, are highly connected and occupy a central place in online forums. Generally, online opinion leaders have a lot of followers and tend to be central in their networks which results in their posts receiving a great number of responses in the form of likes/favorites, replies, and shares. Often social media influencers limit communication to specific topic areas and they dependably update "audiences" with consistent content but also are very responsive to their audience's diverse concerns all of which helps to increase their persuasive influence.

This persuasive influence can be seen on various social media platforms, including Twitter. Twitter, a popular microblogging site where users post short messages or "tweets" with a 280 character limit can also include multimedia content and hyperlinks to other content on various sites. Hashtags are also often included with tweets and these form hyperlinks that connect tweets and have the potential to reach a large audience.

In this study, Twitter was selected for assessment because of its popularity, wide-spread use, and the potential for users to follow anyone. We used information diffusion, a widely used method [6] where the reaction to influencer tweets is assessed to define social media influencers. Through social network analysis, we identified Twitter vaccine-information (pro- and antivaccine) communities and their geolocations. We focused on tweets for three childhood vaccines, measles, mumps, rubella (MMR); tetanus, diphtheria, pertussis (Tdap); and human papillomavirus (HPV). Our goal was to provide insights for public health researchers and health care professionals on efficient forms of locating vaccine-hesitant communities to better target vaccine communication for childhood vaccine promotion.

URI: https://hdl.handle.net/10125/71096
978-0-9981331-4-0

2. Methods

This study employed social network community detection and semantic network analysis (SNA) to identify pro- and antivaccine influencers, their communities, and geolocations on Twitter specific to the three childhood vaccinations previously mentioned. A sentiment analysis was performed to assess if Twitter conversations were negative, neutral, or positive in overall sentiment.

2.1. Data collection

Tweets about childhood vaccinations (MMR, Tdap, HPV) were collected from July 1, 2018, to October 15, 2018. These vaccines were selected to capture more widely discussed childhood vaccines (MMR and Tdap) as well as newer vaccine recommendations (HPV). This timeframe coincided with the peak period of a measles outbreak in Europe and the growing spread of measles in the U.S., as well as the start of the U.S. school-year, which requires parents to indicate child vaccination status for public school enrollment. This period included a more recent, at the time, a record-high measles outbreak in Europe which would provide information on how the growing concern was discussed on Twitter.

Data were collected from Twitter's Premium API using Boolean search methods with the keywords, "vaccine," "vaccination," "vax," "shot," "immunization," "immunisation," in combination with the three childhood vaccines selected for analysis (MMR, Tdap, HPV). The entire archive of English language tweets within the noted 15-week period was included along with tweet information (i.e. number of retweets, replies, and favorites), and sender information, such as geolocation and number of followers.

2.2. Identifying influencers

The tweet data was collected, organized, and cleaned using R (version 3.4.4). Social media influencers were identified by normalizing retweet counts, favorite counts, and reply counts of each tweet and multiplying the three values per the information diffusion method. Tweets with values greater than zero were selected as an influence measure. This measure followed a power-law distribution, from which we obtained the 420 senders with the greatest measured value and their lists of friends.

After collecting the sender's lists and extracting their social connections with followers an edge list of 7,731 connections was created. The edge list was imported into *Gephi* [7] for network detection. *Gephi* was also used to calculate and visualize the social networks of the 420 vaccine influencers.

2.3. Detecting communities and geolocations

Modularity, a community detection method that shows different clusters, or groups, by determining the fraction of links that fall within a given group, was employed to detect the communities among the 420 influencers. As a rule of thumb, modularity of .4 or greater indicates the presence of separate communities [8]. Based on their community, each sender's location information was extracted and summarized. The location summary included country, and state name, if the country was the U.S., for each community. Because our data was based on English language tweets, most locations identified were English-speaking countries.

2.4. Semantic network analysis (SNA)

After cleaning the tweet text data, it was separated into different files based on each sender's community. Text files were preprocessed using *ConText* [9] to remove syntactically functioning words and stem different forms of the same word. The remaining text was analyzed for word frequency. 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. The measurement of word co-occurrence is the standard for creating links between words in a semantic network. Words were considered linked if they co-occurred within three words of each other. The frequencies of word co-occurrence were then calculated and ranked. Word order, or direction, was not considered in the semantic network analysis. *Gephi* [7] was used to create semantic networks and their visualizations, as well as to assess network measures.

2.5. Sentiment analysis

Sentiment analysis from *IBM Watson Natural Language Understanding (NLU) [10]* was used to assess the percentages of positive, negative, and neutral tweets for each community. NLU uses deep learning to extract metadata from text and identifies the attitudes, opinions, or feelings in the text. This analysis considers both the polarity of individual words and the sequence of the text. Twitter data was used to train *NLU* making

it an especially appropriate sentiment analysis tool for this assessment [11].

3. Results

3.1. Community and geolocation detection

The community detection algorithm revealed three distinct communities among vaccine influencers (Figure 1). The modularity was .52, indicating meaningful community detection. While the global network density was 0.05, the within-community densities were 0.33 (the orange community), 0.16 (the green community), and 0.20 (the blue community), with an average of 0.23, 4.6 times greater than the overall density, a further indication of separate groups. Also, the pairwise density was 0.14 for orange and blue, 0.12 for orange and green, and 0.11 for blue and green. Lastly, the orange community (5243 tweets) consisted of 33.81% of the whole network, the green community (4263 tweets) was 38.57%, and the blue community (3981 tweets) took the rest at 27.62%.

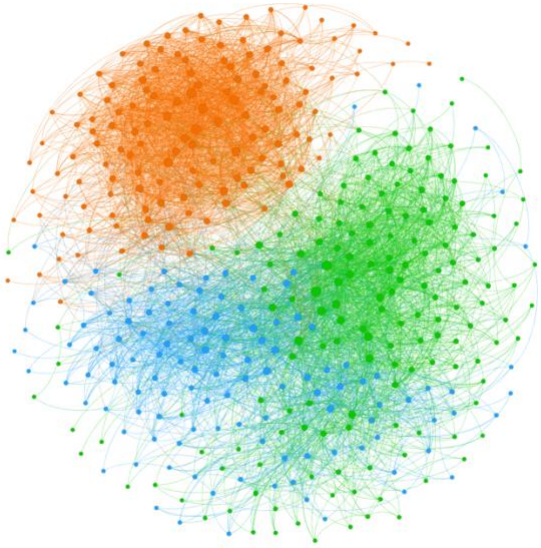

Figure 1. Influencer community detection results

Influencer geolocations for each community were extracted based on their Twitter personal information. The top three countries represented in these communities were the United States (USA), United Kingdom (UK), and Ireland (IE) (Figure 2). Both the orange and green communities were dominated by influencers from the U.S. The orange community was made up of influencers from California, New York, Texas, Georgia, and Florida. These states include some of the most populous states in the U.S. plus the home of the Centers for Disease Control and Prevention (CDC). The green community was dominated by influencers from California, New York, Texas, Washington D.C., and Maryland. Again, this community was comprised of the most populous states plus the home of both the National Institutes of Health (NIH) and the Federal government. The blue community represented influencers from IE and the UK.

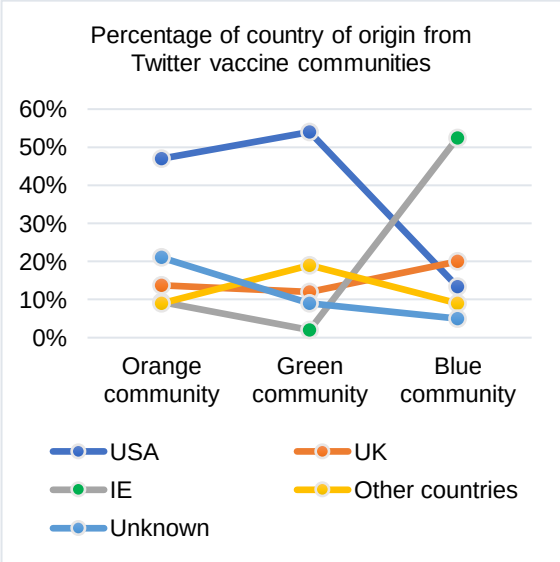

Figure 2. Country of origin by community

3.2. Pro- or antivaccine determination

Influencers were ranked based on their popularity score to locate the top 20 influencers from each community. The results of the influencer rankings were used to identify the community as pro- or antivaccine. The Children's Health Defense, an NGO focused on antivaccine activism and headed by Robert F. Kennedy, Jr. was the number one influencer in the orange community. This influencer was followed by three accounts held by individuals with obvious antivaccine stances. The green community revealed the World Health Organization as the top influencer followed by two media organizations and the National Cancer Institute. Finally, National Health Service, a UK health provider, was the blue community's top influencer followed by two more government accounts.

To further distinguish a community's stance on vaccination (i.e. pro- or anti-), we performed a descriptive analysis for all three identified communities based on the top 20 influencer accounts for each community. Results showed the orange community to be antivaccine comprised of more antivaccine advocates, organizations, and individuals.

The green and blue communities were largely pro-vaccine communities with the green community encompassing more diverse influencer types and the blue community focusing on vaccination promotion as a means for preventing cancer.

The theme(s) for each community were also identified based on the top associations in each community's semantic network. The orange community was mostly concerned with highlighting the dangers of childhood vaccines. The most central words were *vaccination*, *autism*, and *MMR* (Figure 3). The green community involved the promotion of the HPV and MMR vaccines specifically but vaccinations in general as a means to protect from and prevent disease. The most central words in the green community were *vaccination*, *get*, and *vaccineswork* (Figure 4). The blue community focused on extending and encouraging vaccination rates for boys in Ireland for the HPV vaccine. The most central words for this community were *HPV*, *vaccine*, and *boy* (Figure 5).

For Figures 3, 4, and 5, 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. The more closely related the words were, the shorter the link distance. The color of each semantic network (based on senders' network community) matches the color of their sender network community color.

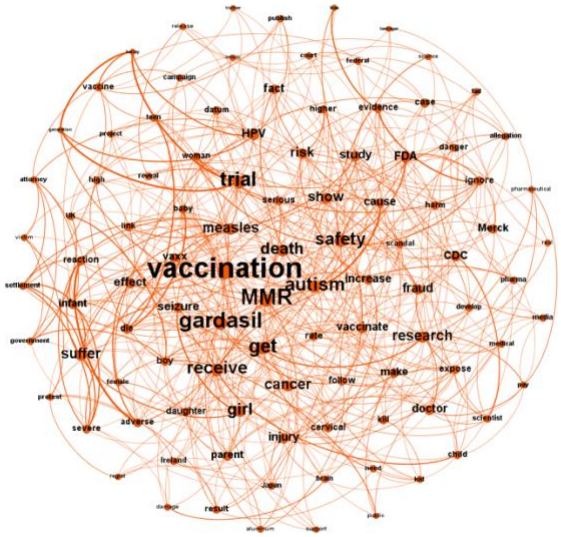

Figure 3. SNA results for the orange community

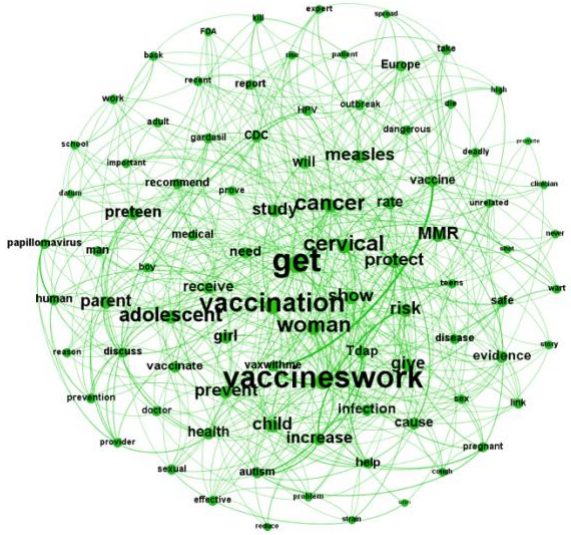

Figure 4. SNA results for the green community

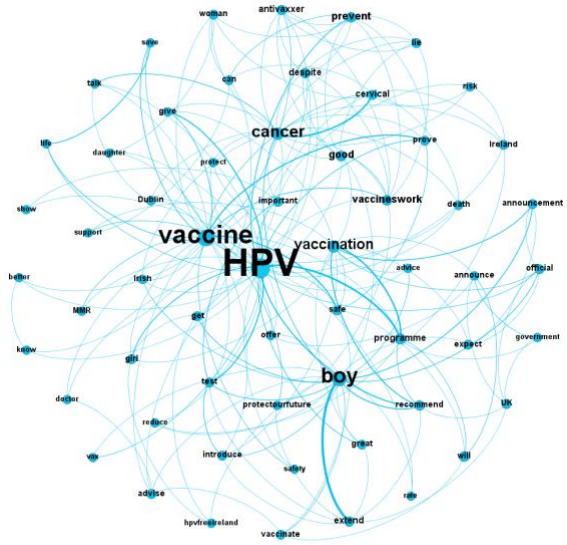

Figure 5. SNA results for the blue community

3.3. Sentiment analysis

Chi-square tests were conducted to evaluate the significance of proportional differences among the communities. Sentiment in the orange community was the most negative among the three communities ($p > .05$). The blue community displayed the most positive sentiment among the three communities ($p > .05$) and the green community presented the most neutral tweets among the three ($p > .05$) (Table 1.).

Table 1. Sentiment analysis results

	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%)

4. Discussion

This study used the identification of social media influencers to detect Twitter childhood vaccine communities and their geolocations. We confirmed that Twitter social media influencers formed independent communities online around the topic of childhood vaccination. The stated positions of each Twitter community were determined through semantic network analysis. The antivaccine community was more connected within when compared to the pro-vaccine communities. Pairwise density results did not show the antivaccine community as being independent of the pro-vaccine communities further highlighting the interconnectedness of the Twitter antivaccine community. This interconnectedness was lacking in the pro-vaccine Twitter communities. These results are in line with recent research that found pro-vaccine Facebook groups discuss vaccination issues mainly with each other rather than reaching out to vaccine-neutral groups or anti-vaccination groups [5]. This is unlike the antivaccine groups that do connect more widely to vaccine-neutral and pro-vaccination groups.

According to the semantic network analysis, the anti-vaccine community propagated misinformation about vaccines in addition to using anti-vaccine rhetoric. This conclusion supports previous research results that found online anti-vaccine information to include deceptive vaccine information [12, 13]. Sentiment analysis found the majority of tweets to be negative in sentiment. This can likely be attributed to the more rapid spread of negative emotions [14]. Considering that popular tweets tend to be more negative in sentiment, it can be expected to find more negative tweets across the communities assessed. Furthermore, vaccines treat diseases, a concept with a generally negative sentiment.

The proliferation of misinformation about vaccines in antivaccine communities can be largely attributed to the make-up of the different Twitter vaccine communities. For example, a descriptive analysis of the top 20 Twitter influencers on childhood vaccination found that the antivaccine community distributed information from emerging vaccine-information websites, personal blogs, and parent-organized groups. The pro-vaccine community, on the other hand, circulated news sourced from traditional mainstream media who obtain information from various reputable health organizations.

Our study shows that using social media influencers to identify antivaccine Twitter communities can be an effective strategy for targeting vaccine misinformation. Rather than monitoring large numbers of tweets, efforts can be focused on influencers and their communities. Moreover, promoting accurate vaccine information through social media influencer accounts will ensure a larger number of Twitter users receive the information. Their large and well-connected networks provide more efficient information coverage.

Like all studies, this research has its limitations. First, we concentrated our assessment on English-language tweets. Twitter conversations on the topic of childhood vaccinations may differ based on language. Considering tweets in additional languages would also likely change the geolocation of tweets. Similarly, this research focused on Twitter as opposed to information from other social media platforms such as Facebook, Instagram, or Weibo. The procedures described in this study may be generalized beyond public health to assess different topics within social media where the goal is social influence. One particular case would be to specify the “echo chambers” that exist in political-ideological communities.

Detecting online vaccine communities and their user geolocations through the identification of social media influencers provides an efficient means for public health officials to more accurately target antivaccine and vaccine-neutral groups to provide accurate vaccine information, answer vaccine safety concerns, counteract vaccine misinformation, and monitor vaccine misinformation spread more efficiently. Lastly, knowing the geolocation clusters for these communities provides valuable information for monitoring gaps in vaccination coverage and may assist in predicting disease outbreak.

10. References

[1] A. M. Jones, S. B. Omer, R. A. Bednarczyk, N. A. Halsey, L. H. Moulton, and D. A. Salmon, "Parents' source of vaccine information and impact on vaccine attitudes, beliefs, and nonmedical exemptions," *Adv Prev Med,* vol. 2012, pp. 1-8, 2012, doi: 10.1155/2012/932741.

[2] E. K. Brunson, "The impact of social networks on parents' vaccination decisions," *Pediatrics,* vol. 131, no. 5, pp. e1397-e1404, 2013, doi: 10.1542/peds.2012-2452.

[3] J. Leask, H. W. Willaby, and J. Kaufman, "The big picture in addressing vaccine hesitancy," (in English), *Hum Vacc Immunother,* vol. 10, no. 9, pp. 2600-2602, Sep 2014, doi: 10.4161/hv.29725.

[4] R. Getman, M. Helmi, H. Roberts, A. Yansane, D. Cutler, and B. Seymour, "Vaccine Hesitancy and Online Information: The Influence of Digital Networks," (in English), *Health Educ Behav,* vol. 45, no. 4, pp. 599-606, Aug 2018, doi: 10.1177/1090198117739673.
[5] N. F. Johnson *et al.*, "The online competition between pro-and anti-vaccination views," *Nature,* 2020, doi: 10.1038/s41586-020-2281-1.
[6] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi, "Measuring user influence in twitter: the million follower fallacy," in *Proceedings of the Fourth International Association for the Advancement of Artificial Intelligence Conference on Weblogs and Social Media*, *May 23-26, 2010, Washington, DC, 2010*. Menlo Park, CA: AAAI Press, 2010. Available: https://www.aaai.org/Library/ICWSM/icwsm10contents.php. [Accessed: September 9, 2018].
[7] M. Bastian, S. Heyman, and M. Jacomy, "Gephi: An open source software for exploring and manipulating networks.," in *Proceedings of the Third International Association for the Advancement of Artificial Intelligence Conference on Weblogs and Social Media*, *May 17 – 20, 2009, San Jose, CA, 2009.* Menlo Park, CA: AAAI Press, 2009. Available: https://gephi.org/users/publications/. [Accessed: November 11, 2009].
[8] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," (in English), *J Stat Mech-Theory E,* Oct 2008, doi: 10.1088/1742-5468/2008/10/P1008.
[9] J. Diesner, "ConText: Software for the Integrated Analysis of Text Data and Network Data," in *Proceedings of the Social and Semantic Networks in Communcation Research Preconference of the International Communication Association (ICA), May 22 – 26, 2014, Seattle, WA*. Available: http://context.ischool.illinois.edu/download.php#citing. [Accessed: November 11, 2009].
[10] IBM, "IBM Cloud API Docs: Natural Language Understanding." [Online]. Available: https://cloud.ibm.com/apidocs/natural-language-understanding/natural-language-understanding?code=python#sentiment
[11] S. Vergara, M. El-Khouly, M. El Tantawi, S. Marla, and L. Sri, *Building Cognitive Applications with IBM Watson Services: Volume 7 Natural Language Understanding*, ibm.com/redbooks: IBM Readbooks, 2017, p. 110.
[12] 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," (in English), *Vaccine,* vol. 30, no. 25, pp. 3778-3789, May 28 2012, doi: 10.1016/j.vaccine.2011.11.112.
[13] M. B. Moran, M. Lucas, K. Everhart, A. Morgan, and E. Prickett, "What makes anti-vaccine websites persuasive? A content analysis of techniques used by anti-vaccine websites to engender anti-vaccine sentiment," *Journal of Communication in Healthcare,* vol. 9, no. 3, pp. 151-163, 2016, doi: 10.1080/17538068.2016.1235531.
[14] S. Tsugawa and H. Osaki, "Negative Messages Spread Rapidly and Widely on Social Media," in *Proceedings of the Advancing Computing as a Science and Profession Conference on Online Social Networks*, *November 2 – 3, 2015*. Palo Alto, CA: Association for Computing Machinery, 2015. Available: https://dl.acm.org/doi/proceedings/10.1145/2817946. [Accessed: November 30, 2018].