

Semantic network analysis of vaccine sentiment in online social media



Gloria J. Kang ^{a,b}, Sinclair R. Ewing-Nelson ^b, Lauren Mackey ^b, James T. Schlitt ^{a,b}, Achla Marathe ^b, Kaja M. Abbas ^a, Samarth Swarup ^{b,*}

^a Department of Population Health Sciences, Virginia Tech, USA

^b Biocomplexity Institute, Virginia Tech, USA

ARTICLE INFO

Article history:

Received 22 December 2016

Received in revised form 29 April 2017

Accepted 17 May 2017

Available online 27 May 2017

Keywords:

Vaccine hesitancy

Vaccine sentiment

Semantic network analysis

Online social media

Twitter

ABSTRACT

Objective: To examine current vaccine sentiment on social media by constructing and analyzing semantic networks of vaccine information from highly shared websites of Twitter users in the United States; and to assist public health communication of vaccines.

Background: Vaccine hesitancy continues to contribute to suboptimal vaccination coverage in the United States, posing significant risk of disease outbreaks, yet remains poorly understood.

Methods: We constructed semantic networks of vaccine information from internet articles shared by Twitter users in the United States. We analyzed resulting network topology, compared semantic differences, and identified the most salient concepts within networks expressing positive, negative, and neutral vaccine sentiment.

Results: The semantic network of positive vaccine sentiment demonstrated greater cohesiveness in discourse compared to the larger, less-connected network of negative vaccine sentiment. The positive sentiment network centered around *parents* and focused on communicating health risks and benefits, highlighting medical concepts such as *measles*, *autism*, *HPV vaccine*, *vaccine-autism link*, *meningococcal disease*, and *MMR vaccine*. In contrast, the negative network centered around *children* and focused on organizational bodies such as *CDC*, *vaccine industry*, *doctors*, *mainstream media*, *pharmaceutical companies*, and *United States*. The prevalence of negative vaccine sentiment was demonstrated through diverse messaging, framed around skepticism and distrust of government organizations that communicate scientific evidence supporting positive vaccine benefits.

Conclusion: Semantic network analysis of vaccine sentiment in online social media can enhance understanding of the scope and variability of current attitudes and beliefs toward vaccines. Our study synthesizes quantitative and qualitative evidence from an interdisciplinary approach to better understand complex drivers of vaccine hesitancy for public health communication, to improve vaccine confidence and vaccination coverage in the United States.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

1.1. Vaccine hesitancy

Suboptimal vaccination coverage in the United States continues to pose significant risk of disease outbreaks, in part, due to vaccine hesitancy [1]. Vaccine hesitancy refers to a combination of beliefs, attitudes, and behaviors that influence an individual's decision to vaccinate despite vaccine availability; these behaviors include refusal, delay, or reluctant acceptance despite having active concerns [2,3]. Strategies to address vaccine refusal have focused on

individual reasons for not vaccinating, however, evidence of successful interventions remains limited. A review of vaccine hesitancy interventions expressed weak support for current strategies in mitigating vaccine resistance [4]; interventions targeted toward anti-vaccination groups are likely to be ineffective, unsustainable, and potentially more detrimental compared to no intervention at all [4–6].

Vaccine hesitancy stems from socio-cultural, political, and otherwise non-medical factors that are poorly understood [7]. The underlying causes of vaccine hesitancy should not be attributed to scientific illiteracy alone [8], but rather viewed as a deliberative and structured process that requires contextualized examination at local levels [9,10]. In the case of our study, we focus on semantic and rhetorical qualities of vaccine communication

* Corresponding author at: Biocomplexity Institute, Virginia Tech, 1015 Life Science Circle, Blacksburg, VA 24061, USA.

E-mail address: swarup@vt.edu (S. Swarup).

amongst the general public within contexts of differing vaccine sentiment.

1.2. Social network analysis and digital epidemiology

The advent of the Internet and social media has provided new platforms for persuasion and rapid spread of (mis)information, bringing forth new challenges and opportunities to an age-old public health problem. Social Network Analysis (SNA) broadly studies social interactions of contact networks with significant implications for public health [11], such as contributing evidence that belief systems are a primary barrier to vaccination [12]. Novel public health tools such as SNA employ computational frameworks in the context of digital epidemiology [13]. Online social media such as Twitter are novel avenues to acquire real-time data of attitudes, beliefs, and behaviors, particularly for underrepresented demographic groups who disproportionately comprise Twitter users [14]. By leveraging online data, studies can examine the dynamics of massively interacting populations, such as online health sentiment and its potential impact on infectious disease outbreaks [15,16].

1.3. Semantic networks

Semantic networks are graphical representations of knowledge based on meaningful relationships of written text, structured as a network of words cognitively related to one another [17,18], in this study, vaccine information. Within the semantic network, nodes are words that represent concepts found in text. The connections between nodes are referred to as edges which represent relationships between connected concepts. Semantic networks allow extraction of meaningful ideas by identifying emergent clusters of concepts rather than analyzing frequencies of isolated words [19]; in this way, analyzing online social media can enhance understanding of complex health behavior, particularly for vaccine hesitancy.

Similar studies have analyzed websites using search engine results and natural language processing (NLP) [20,21]. Text network analysis traditionally employs semi-automated techniques in which information is extracted and analyzed using both human and computerized methods, dealing with challenges such as coreference resolution, synonym resolution, and ambiguity [22]. To

Table 2

Summary of measures for article text networks and sentiment group networks. The table describes network characteristics of extracted web documents; joint semantic networks of positive, negative, and neutral vaccine sentiment; and the corresponding greatest connected component. Measures describe network size, density, and average centrality.

Vaccine sentiment	Positive	Negative	Neutral
<i>Document text networks</i>			
Number of documents (total = 50)	23 documents	21 documents	6 documents
Average number of nodes (per document)	53.1 nodes	90.9 nodes	43.8 nodes
Average number of edges (per document)	49 edges	90.7 edges	39.7 edges
Average degree (per document)	1.9	1.98	1.8
<i>Vaccine sentiment networks</i>			
Average degree	3.356	2.95	2.348
Number of connected components	21	49	12
<i>Greatest component subgraph</i>			
Nodes/total network nodes	585/652 nodes	1140/1257 nodes	171/201 nodes
Edges/total network edges	1042/1094 edges	1783/1854 edges	216/236 edges
Average degree	3.562	3.128	2.526
Diameter	12	13	17
Density	0.0061	0.0027	0.0149
Number of communities	21	31	10
Average path length	4.492	4.77	6.78
Average degree centrality	0.0061	0.0027	0.0149
Average betweenness centrality	0.006	0.0033	0.0342
Average closeness centrality	0.2292	0.2161	0.1533
Average node connectivity	1.3117	1.1835	1.035
Average clustering coefficient	0.196	0.14	0.131

limit these issues, we constructed semantic networks manually and then performed network analysis within our study.

Both proximate and non-proximate determinants of vaccine hesitancy necessitate an interdisciplinary approach [23,24]. Our study presents a novel framework that applies methods of network analysis to semantic networks [25] within the context of vaccine sentiment.

Table 1

Summary of sampled documents. The table summarizes article characteristics by vaccine sentiment group and describes document type, article source, target vaccine population, vaccine type focus, and specific vaccine topics.

Vaccine sentiment articles (total n = 50)	Positive (n = 23)	Negative (n = 21)	Neutral (n = 6)
Document type			
Blog =	8 (34.8%)	Blog =	15 (71.4%)
News =	7 (30.4%)	Alternative News =	2 (9.5%)
Magazine =	5 (21.7%)	Magazine =	2 (9.5%)
Informational =	3	Commercial =	1 (4.8%)
	(13.0%)	News =	1 (4.8%)
Article source type			
Media =	9 (39.1%)	Media =	15 (71.4%)
Government =	8 (34.8%)	Industry =	3 (14.3%)
News =	4 (17.4%)	Personal =	2 (9.5%)
Industry =	1 (4.4%)	Forum =	1 (4.8%)
Resource =	1 (4.4%)		
Target vaccine population			
Childhood =	16 (69.6%)	Childhood =	15 (71.4%)
Adolescent =	3 (13.0%)	Adolescent =	0
Adult =	1 (4.4%)	Adult =	0
Multiple =	3 (13.0%)	Multiple =	6 (28.6%)
Vaccine type focus			
General =	8 (34.8%)	General =	14 (66.7%)
Specific =	15 (65.2%)	Specific =	7 (33.3%)
Specific vaccines			
Measles/MMR =	9	Shingles =	1
HPV =	3	Polio =	1
Influenza =	1	Gardasil =	1
Meningococcal =	1	Measles =	1
Rubella =	1	Swine flu =	1
		Tdap =	1
		Hepatitis B =	1

1.4. Study objective

Our objective was to examine current vaccine sentiment on social media by constructing and analyzing semantic networks of vaccine information from highly shared websites of Twitter users in the United States.

1.5. Public health significance

The Strategic Advisory Group of Experts on Immunization (SAGE) Working Group on Vaccine Hesitancy (WG) reported specific research needs to better understand context-specific causes underlying vaccine hesitancy [26]. To help address this gap, we utilized quantitative network methods in analyzing qualitative aspects of vaccine information—an efficient approach to investigating the scope and variability of current attitudes and beliefs toward vaccines. Such findings are pivotal in informing and improving public health communication of vaccine confidence.

2. Methods

2.1. Data retrieval and document selection

We used ChatterGrabber [27], a web-scraping tool that randomly samples public tweets of Twitter users in the United States. (Details on ChatterGrabber including search term conditions, qualifiers, and exclusions are in [Appendix A](#)). Webpage links from col-

lected tweets identified current sources of vaccine information based on the frequency of link shares during the time of data collection. Our analysis focuses on the textual content of relevant webpage articles (also referred to as documents) and not the tweeted text per se. Document types selected for analysis included blog posts, media stories, informational articles, and news reports. We excluded academic publications, court documents, and media formats such as images, PDF files, and videos.

A total of 26,389 tweets were collected between April 16, 2015 and May 29, 2015 from which we obtained 8416 unique web links. To generalize findings from a representative pool of popular vaccine articles, we screened the top 100 most shared links for relevance from which we randomly sampled 50 for analysis; we excluded articles concerning non-human vaccines.

2.2. Vaccine sentiment coding

Articles were read for content and manually coded as having either positive, negative, or neutral sentiment toward vaccines. Coding was determined by whole-text assessment which included examining the title/headline and the source/domain of articles. In general, differences between sentiment were determined based on consistency of statements that clearly identified group affiliation, such as encouraging vaccination and highlighting benefits (positive sentiment) or discouraging vaccination and highlighting risks (negative sentiment). Articles that were ambiguous or mixed in sentiment were coded as neutral. Three researchers (GJK, SRE,

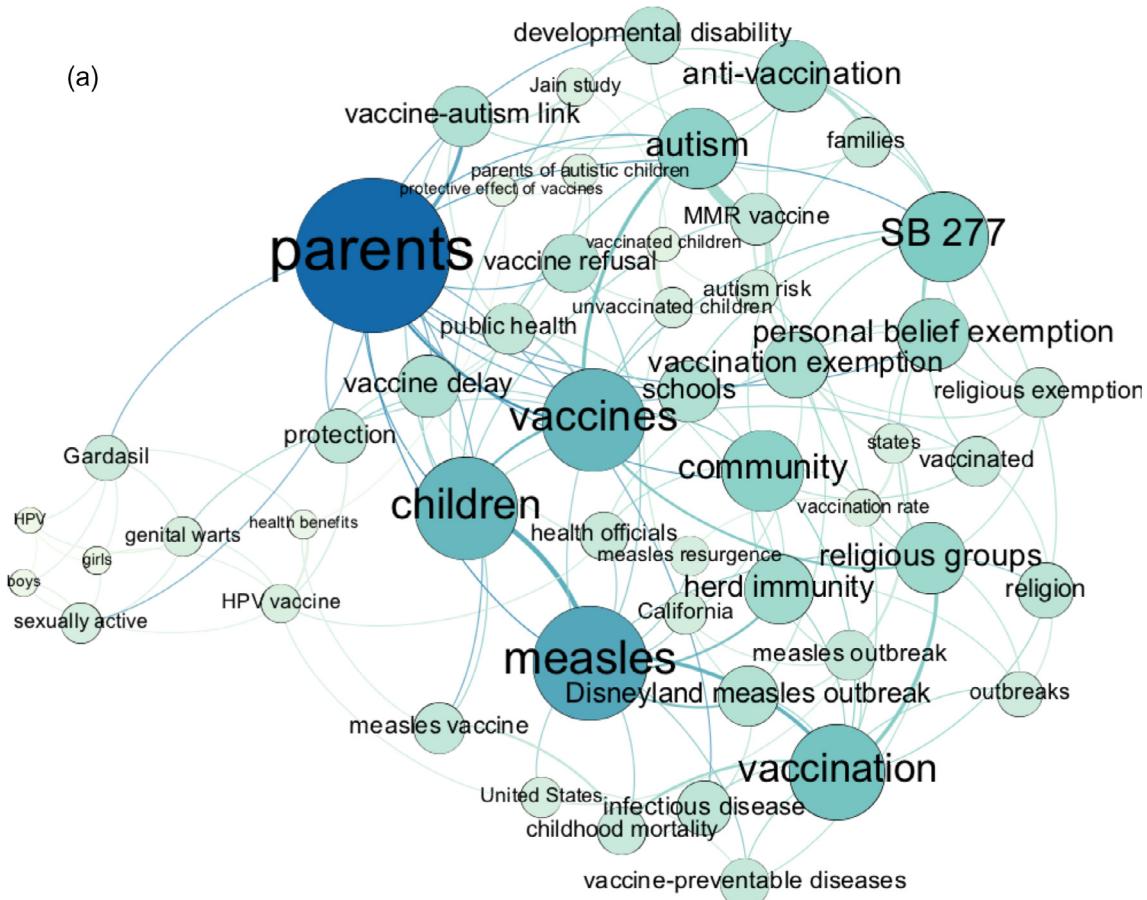


Fig. 1. Maximum k-core subgraphs show clusters of significant vaccine concepts within the semantic networks. Visualizations of maximum k-cores (i.e., the maximal connected subgraph in which all nodes have degree of at least k) for networks of [a] positive vaccine sentiment ($k = 4$), [b] negative vaccine sentiment ($k = 4$), and [c] neutral vaccine sentiment ($k = 2$) where increasing node and text size represents increasing betweenness centrality. [a] Maximum k-core ($k = 4$) subgraph show clusters of significant network concepts within the positive sentiment network. [b] Maximum k-core ($k = 4$) subgraph show clusters of significant network concepts within the negative sentiment network. [c] Maximum k-core ($k = 2$) subgraph show clusters of significant network concepts within the neutral sentiment network.

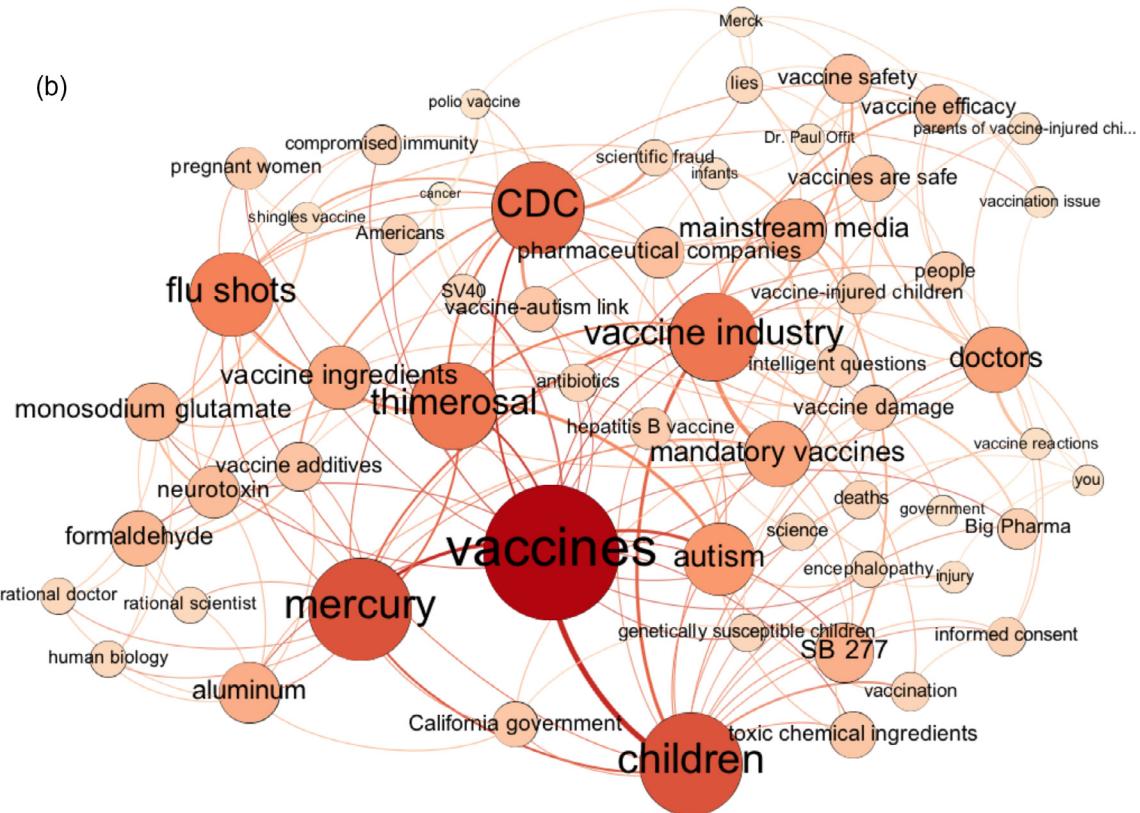


Fig. 1 (continued)

LM) independently coded a subset of 10 articles for sentiment; there was no inter-annotator variability and resulted in consistent sentiment coding.

2.3. Construction of vaccine sentiment networks

Document text networks were merged by sentiment group, thereby aggregating similar documents into a single semantic network, one for each vaccine sentiment (positive, negative, and neutral). We standardized node and edge labels to resolve lexical differences and grammatical dependencies across disparate sources. Details on semantic network annotation, construction, and analysis of vaccine sentiment networks are described in Appendix B.

2.4. Semantic network analysis

Our analysis of the positive, negative, and neutral sentiment networks was focused on the greatest connected component (or subgraph). We applied several measures of network analysis to the generated semantic networks in order to limit biased interpretation of selected network metrics [25] (Appendix B). Descriptive statistics included network size, density, and diameter, where network size is the total number of nodes (i.e., vaccine concepts); density measures the interconnectedness of nodes [28]; and diameter characterizes compactness of the network. We evaluated multiple measures of centrality which describes the importance, influence, or significance of concepts within the semantic network in various ways [29]; specific types include degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality [30].

Community detection algorithms [31] describe cohesive groups in the network [32], and clusters of important vaccine concepts were visualized by the network's maximum k-core (the maximal connected subgraph in which all nodes have degree of at least k)

[33]. We assessed differences in emphasis framing, which is the salience of certain story elements over others [34], for central concepts from networks of differing sentiment. Closeness vitality [49] measures how much the distances between all pairs of nodes change when a particular node is removed. This is an indicator of how much each node contributes to the overall structural cohesion of the network.

NetworkX [35] and iGraph [36] were used in network construction and analysis; visualizations were created in Gephi [37].

3. Results

3.1. Document characteristics

From the sample of webpages ($n = 50$), we coded 23 documents as having positive vaccine sentiment, 21 documents with negative vaccine sentiment, and 6 documents were classified as neutral. Table 1 summarizes document characteristics grouped by vaccine sentiment. Blog posts were the most shared document type overall, followed by news and “alternative news” for positive and negative sentiment articles respectively. Content of positive sentiment documents focused on specific childhood, adolescent, and adult vaccines, whereas negative sentiment documents focused primarily on childhood vaccines and vaccination in general.

3.2. Document text networks

Network properties of vaccine documents are summarized in Table 2. Negative sentiment documents ($n = 21$) formed the largest semantic networks with a mean network size of 90.9 concepts (nodes) per document, compared to smaller networks of positive sentiment ($n = 23$) and neutral sentiment documents ($n = 6$) with a mean of 51.3 and 43.8 concepts per document respectively.

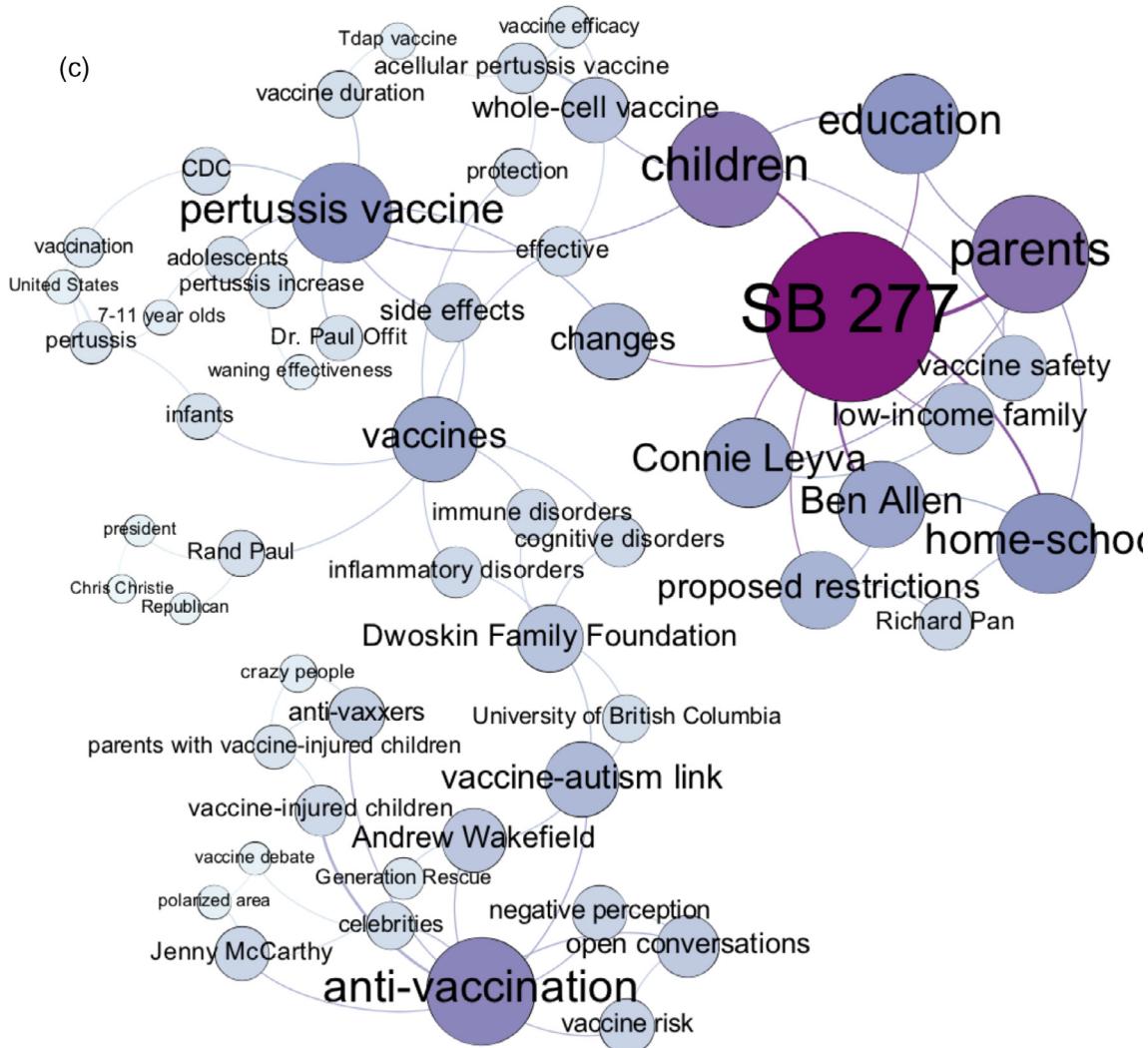


Fig. 1 (continued)

3.3. Vaccine sentiment networks

Document text networks were aggregated by vaccine sentiment to form 3 semantic networks representing positive, negative, and neutral sentiment. Network measures are summarized in Table 2. Network visualizations are in Appendix C.

In regards to the greatest component subgraph, size indicates the number of concepts in the network, whereas density describes interconnectedness of the concepts. The greatest component of the negative network was largest in size (1140 concepts) but less dense (0.0027) than the positive network (0.0061) also much smaller in size (585 concepts). Community detection analysis [31] identified 21 distinct communities within the positive network, 31 communities in the negative, and 10 communities in the neutral network. Compared to the original number of merged documents per sentiment network, the number of cohesive communities exceeded the number of original documents within the negative and neutral networks, whereas the positive network formed fewer communities than the original number of documents used in merging. Community findings and density measures for the positive network suggest a more cohesive and interconnected belief system among positive sentiment concepts compared to the larger, less-connected network of negative sentiment. Correspondingly, the average clustering coefficient (i.e., the tendency of nodes to form groups) and average node centrality for degree, betweenness,

closeness, and eigenvector centrality were higher for the positive network compared to the negative. Positive and negative networks exhibited structural similarities in regards to diameter (12 and 13, respectively) and average path length (4.5 and 4.8, respectively). Visualizations of maximum k-core subgraphs for each sentiment network highlight clusters of significant concepts in Fig. 1.

3.4. Central concepts

Fig. 2 plots significant concepts of each sentiment network by centrality measures for degree, betweenness, and closeness centrality (Appendix D). The most central concepts (greater than 2 standard deviations from the mean) ranked by eigenvector centrality are plotted in Fig. 3 and listed in Table 3.

Excluding expected nodes such as vaccines and vaccination, the most central concepts for the positive network included *parents*, *measles*, *children*, *SB 277*, *autism*, *community*, *religious groups*, *anti-vaccination*, *vaccine-autism link*, *HPV vaccine*, *meningococcal disease*, and *MMR vaccine*. Significant concepts within the negative sentiment network were *children*, *thimerosal*, *CDC*, *vaccine industry*, *mercury*, *autism*, *flu shots*, *mainstream media*, *doctors*, *SB 277*, *vaccine ingredients*, *mandatory vaccines*, and *pharmaceutical companies*. And the most central concepts of the neutral network were *SB 277*, *anti-vaccination*, *parents*, *children*, *pertussis vaccine*, *home-school*, *education*, *pertussis*, *vaccine-autism link*, *side effects*, *Dwoskin Family Foundation*.

Family Foundation, whole-cell vaccine, effective, acellular pertussis vaccine, and high-dose flu vaccine.

3.5. Dynamic visualizations

Dynamic, interactive visualizations and network data files from this study are available online ([Appendix E](#)).

4. Discussion

4.1. Semantic network analysis of vaccine sentiment

A long line of research in the psychology of memory and semantic processing has provided evidence for semantic network-like organization of internal representations and spreading activation as a process by which memories are activated and meaning is processed [53,54,50,51]. In this model, when an item in memory is activated, e.g., by a person reading about it or hearing about it, the activation spreads from that node in the person's internal semantic network to nearby nodes. Spreading activation is also hypothesized as the model for the automatic activation of attitudes [55].

From this perspective, closeness centrality is a useful metric to understand the organization of the vaccination semantic networks (though other centrality measures are quite similar in ranking, as the results show). Closeness centrality is a direct measure of which concepts are likely to be activated repeatedly in each of the semantic networks, even as different concepts are mentioned.

Many central concepts of the positive network were present in the negative network, but not vice versa. For example, while positive and neutral sentiment documents explicitly addressed the concept of *anti-vaccination*, negative sentiment articles did not. In regards to highly central concepts of the negative network, the positive network lacked any reference to the *vaccine industry* and *mainstream media*; *CDC* and *doctors* also held lesser significance in the context of positive vaccine sentiment.

Significant concepts within the positive network were related to health and medicine, such as *measles*, *autism*, *HPV vaccine*, *vaccine-autism link*, *meningococcal disease*, and *MMR vaccine*. In contrast, significant concepts of the negative network referred to organizational bodies such as *CDC*, *vaccine industry*, *doctors*, *mainstream media*, *pharmaceutical companies*, and *United States*. A notable contrast was the emergence of *parents* as the most central concept in the positive network, versus *children*, the most central node in the negative network.

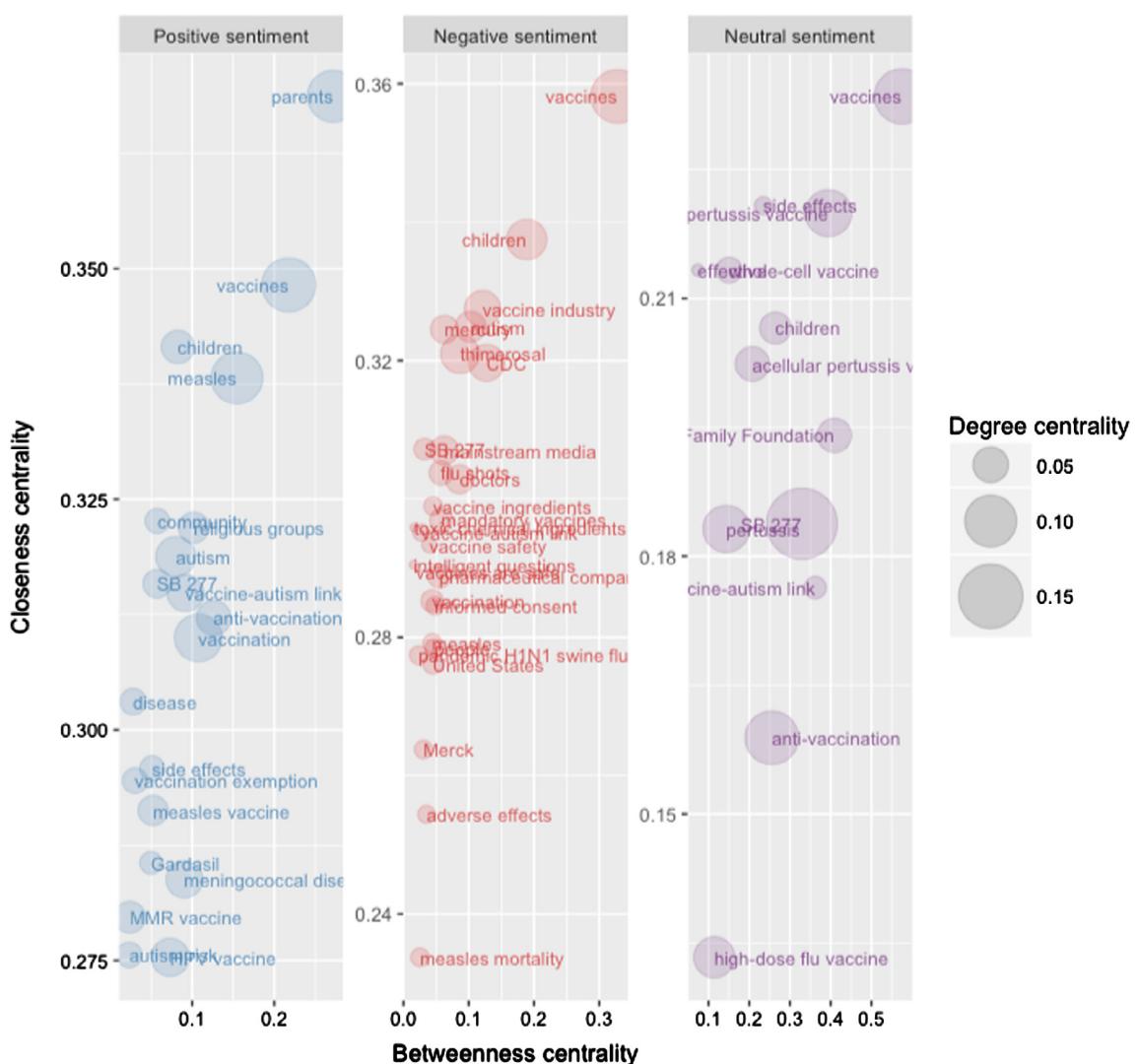


Fig. 2. Significant vaccine concepts by measures of degree centrality, betweenness centrality, and closeness centrality. The figure includes centrality measures for significant concepts from positive, negative, and neutral sentiment networks. Degree centrality (point size), betweenness centrality (x-axis), and closeness centrality (y-axis) are plotted.

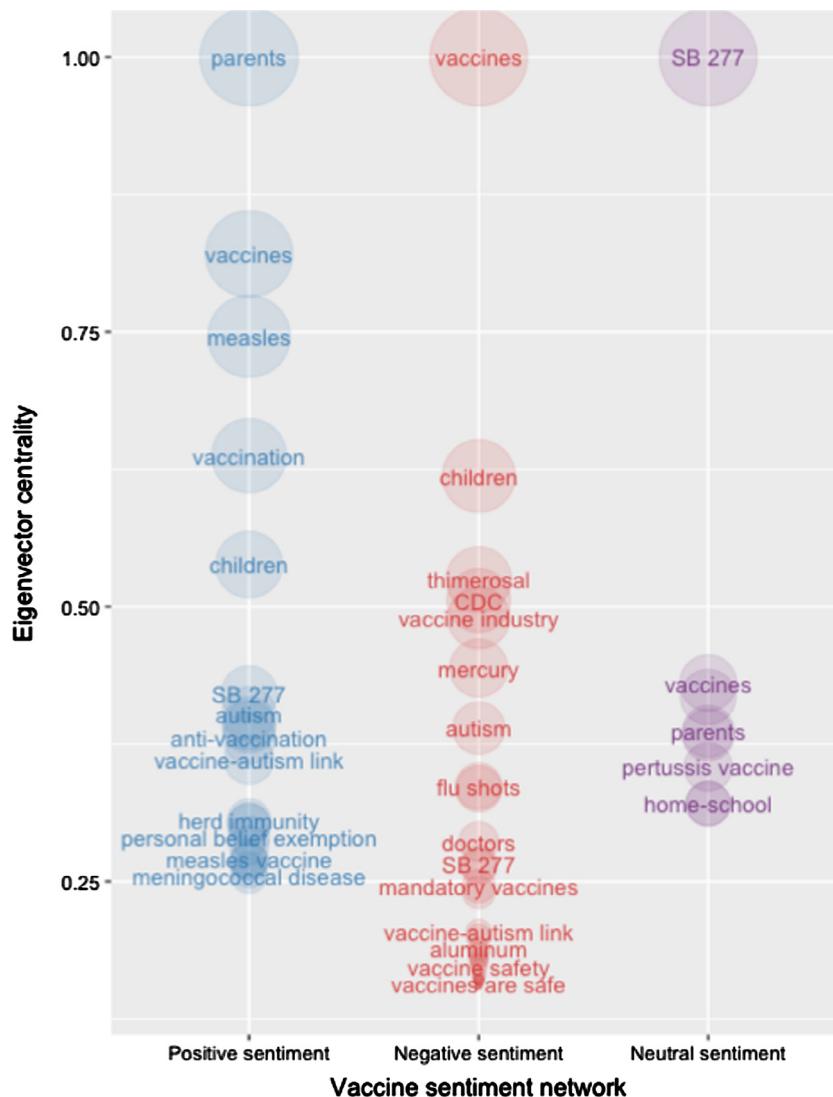


Fig. 3. Significant concepts ranked by eigenvector centrality. The figure plots the most central nodes by eigenvector centrality score for networks of positive, negative, and neutral vaccine sentiment.

Documents expressing positive and neutral vaccine sentiment were characterized by dense semantic networks with fewer concepts, compared to the semantic network of negative sentiment which presented a high number of vaccine concepts with low connectivity. Compared to the positive sentiment network, the negative sentiment network has more components, lower edge density, a larger diameter, and larger average path length (Table 2). Hence, positive sentiment documents indicated greater cohesiveness in vaccine-positive discourse compared to vaccine-negative documents which addressed a broad range of topics as potential contributors to vaccine hesitancy.

4.2. Message framing

Our study revealed sentiment-specific terminology used in framing positive and negative messages within vaccine communication. This included differences in term valence such as *required vaccines* versus *mandated vaccines* and *side effects* versus *adverse effects*, the selective targeting of *parents* versus *children*, and the overall presentation of evidence-based science versus social commentary related to issues of governance for the positive and negative vaccine sentiment networks, respectively.

Overall, the prevalence of negative vaccine sentiment was demonstrated through diverse messaging, framed around institutional distrust and skepticism towards the organizations that deliver scientific evidence of positive vaccine benefits. This is also shown by the list of top nodes for the closeness vitality measure for each network (Table D4), which is an indicator of the concepts which are responsible for providing structural cohesion to the semantic network [49]. Positive and negative vaccine articles largely differed in the framing of trust. Positive articles emphasized trust in vaccination by relying on scientific evidence as trusted authority. Negative articles framed trust issues not around vaccination science itself, but around the institutions that govern or finance matters of personal health. Neutral vaccine articles exemplified various sources of news coverage that expressed a mix of both positive and negative attitudes toward vaccines. Top news stories at the time of data collection included a new study debunking the vaccine-autism link and the passing of California Senate Bill 277 [38], which removed exemptions from school vaccination requirements. News coverage generally expressed positive vaccine sentiment, reporting official statements and statistics. In contrast, news coverage by negative vaccine articles additionally introduced a range of tangential topics, often proposing arguments through rhetorical questions and reframing official statistics.

Table 3

Top nodes by eigenvector centrality. The table lists the most central concepts of each sentiment network ranked by eigenvector centrality score (greater than 2 standard deviations from the mean).

Eigenvector centrality		
Positive sentiment network	Negative sentiment network	Neutral sentiment network
Mean = 0.0626	Mean = 0.0318	Mean = 0.0975
Std Dev = 0.0936	Std Dev = 0.06	Std Dev = 0.11
Parents	1	SB 277
Vaccines	0.8209	Vaccines
Measles	0.7458	Anti-vaccination
Vaccination	0.6373	Parents
Children	0.5382	Children
SB 277	0.4207	Pertussis vaccine
Autism	0.4025	Home-school
Community	0.3937	Education
Religious groups	0.3905	0.3863
Anti-vaccination	0.3802	Vaccines
Vaccine-autism link	0.3608	0.4177
Herd immunity	0.3058	Anti-vaccination
Vaccine refusal	0.3024	Parents
Vaccination exemption	0.3013	Children
Personal belief exemption	0.2909	0.3830
Disease	0.2829	Pertussis vaccine
Measles vaccine	0.2706	Home-school
Schools	0.2685	0.3209
HPV vaccine	0.2674	Education
Vaccine delay	0.2603	0.3206
Meningococcal disease	0.2551	
	Vaccines	1
	Children	SB 277
	Thimerosal	Vaccines
	CDC	0.4304
	Vaccine industry	Anti-vaccination
	Mercury	0.4177
	Autism	Parents
	Flu shots	Children
	Mainstream media	0.3863
	Doctors	Pertussis vaccine
	SB 277	Home-school
	Vaccine ingredients	0.3830
	Mandatory vaccines	0.3209
	Pharmaceutical companies	Education
	Vaccine-autism link	0.3206
	Toxic chemical ingredients	
	Aluminum	
	Vaccination	
	Monosodium glutamate	
	Hepatitis B vaccine	
	Vaccine-injured children	
	Vaccine safety	
	Evidence	
	Informed consent	
	Intelligent questions	
	Formaldehyde	
	Pregnant women	
	Pandemic H1N1 swine flu vaccine	
	Big pharma	
	Vaccines are safe	
	Quackery	
	Vaccine damage	
	SV40	
	Science	

4.3. Limitations

We assumed that popular vaccination information shared on Twitter is representative of prevalent vaccine sentiment, but may not reflect the broad spectrum of vaccine sentiment in the general population. Coding documents for neutral sentiment was difficult since documents presented a mix of both positive and negative attitudes, and not truly vaccine-neutral. Because health behaviors are founded upon a variety of beliefs and attitudes that change over time, vaccine sentiment categories are difficult to delineate since they do not exist as polarized groups.

While we attempted to resolve issues of meaning and context by manually transcribing implicit statements into explicit statements, reference resolution grew increasingly difficult across different documents. Consequently, there is potential inconsistency from the manual annotation of document text into network data, particularly when dealing with ambiguous language such as slang, hyperbole, and poetic devices. Despite these limitations, employing human interpretation of text greatly enhances qualitative aspects of data and is arguably more accurate than current NLP methods which lack explicit domain-specific knowledge or situational information [22]. Lastly, our analysis did not assess the qualitative relationships of connected concepts. Future studies incorporating edge data can provide detailed insight into the comparison of belief structures of varying vaccine sentiment.

Our study presents only a broad overview of general network measures. Greater depth into specific metrics, such as community detection analysis, can provide useful insight and should be addressed in future studies.

4.4. Implications for public health and vaccine communication

The SAGE WG on Vaccine Hesitancy [26] states that communication is a tool to address vaccine sentiment rather than a determinant of hesitancy. However, poor communication can undermine vaccine acceptance in any setting [39]. Our study lends itself to the development of effective communication strategies for target populations by identifying specific factors that influence vaccine hesitancy—an integral component of every immunization program [39].

Semantic network analysis of vaccine sentiment in online social media can enhance our understanding of the scope and variability of attitudes and beliefs toward vaccination. Our findings emphasize the need to improve the framing and messaging of public health communication, that not only highlights the vaccine benefits, but also addresses specific issues related to vaccine hesitancy and institutional distrust. Enhancing public trust in relevant scientific institutions and engaging in efficient public health communication is critical in improving vaccine confidence and vaccination coverage [40].

4.5. Conclusion

We discussed findings from a novel framework that uses semantic network analysis as an efficient and effective way to analyze vaccine sentiment. This study adds to a growing body of vaccine hesitancy research by investigating emerging topics and the various discourse surrounding current vaccine perspectives. Findings related to significant concepts, the structure of its relations,

and semantic qualities can better inform targeted vaccine communication strategies and enhance effectiveness of public health efforts to increase vaccine confidence.

Funding

This study is supported by NIH/NIGMS R01GM109718, NSF/NRT 1545362, and NSF IBSS Grant SMA-1520359. The funding sources had no role in study design; collection, analysis, and interpretation of data; writing of the paper; or the decision to submit it for publication.

Conflicts of interest

None.

Appendix A. ChatterGrabber parameters, search terms, and summary of results

ChatterGrabber search terms were selected through an iterative process involving manual selection and testing of data retrieval as detailed in [27].

[A1] Description of ChatterGrabber parameters.

Location	United States
Tweet data	Text, ID, time posted, retweet count, favorite count
User data	Screen name, language
Media data	Url, display Url

[A2] ChatterGrabber search terms.

Conditions	Qualifiers	Exclusions
Vaccine	Autism	Bullshit
Vaccinat	Autistic	Penn & teller
Vaccine	Conspiracy	Penn and teller
Vacinate	Gave my	Enter the kingdom of heaven
MMR	Gave me	Heroin
Antivac	Oprah	Eye of a needle
	Aspergers	Thread
	Poison	Molds
	Jenny mccarthy	Record
	Kristin cavallari	Efficacy
	Conspiracy	Shoot up
	Mercury	Needle exchange
	Aluminum	Morphine
	Truther	Knit
	Bravo	Crochet
	Anti	Fracking
	Manufacturers	Insulin
	Have known	Malware
	Vaccine choice	Pincushion
	Your child	Addict
	Your right	Fuel
	Cancer	Needlework
	Fertility	Felt
	Constitution	Caffeine
	Risks	Scaling
	Dangerous	Space

[A3] Twitter data via ChatterGrabber.

	<i>n</i>
Total number of collected tweets	26,389
Number of unique urls	8416
Number of unique domains	2372
Number of web articles selected for analysis	50

Appendix B. Network methods

B1. Network annotation and construction

To create document networks, article text was manually transcribed into structured belief statements, or relevant information extracted from natural language text. Similar to methods of information extraction used by the *Knowledge Vault* project [41], document text was formatted as *triples*, in which (*subject*, *predicate*, *object*) correspond to (*node*, *edge*, *node*) in the network. For example, the sentence “*Vaccines prevent communicable diseases*” is represented by (*vaccines*, *prevent*, *communicable diseases*). Three researchers initially annotated a subset of 10 documents to gauge inter-annotator variability in transcribing article documents into network datasets. All co-references were resolved and the original text was adhered to as much as possible. Discordant results were resolved through consensus in order to maintain standard formatting of network data. Final network datasets were synthesized by standardizing terminology, resolving grammatical dependencies and lexical differences in the semantic network.

The resulting standards for network vocabulary were based on term frequency. For example, synonymous nodes labeled “*communicable diseases*”, “*infectious diseases*”, and “*contagious diseases*”, we applied the most commonly used term across same-sentiment documents (in this case “*infectious diseases*”) to replace labels of all semantically equivalent nodes.

B2. Definitions of network measures

Network size is the total number of nodes or vaccine-related concepts. Density measures the interconnectedness of nodes, calculated as the proportion of existing edges (or relations between concepts) over all possible edges in the network [42]. Diameter characterizes the compactness of the network, measured as the longest path of all shortest paths across all node pairs.

Degree centrality characterizes how connected a node is to other nodes in the network, measured by its number of connections (and normalized by the total number of network connections) [43]. Betweenness centrality measures the frequency of a given node on the shortest paths to all other pairs of connected nodes, representing the probability of a concept to be involved in connecting two other concepts in the semantic network [43,44]. Closeness centrality measures closeness, calculating the sum of the shortest paths between a node to all other nodes in the network [43]. Nodes with smaller path lengths have higher closeness centrality and are interpreted to be more important concepts than nodes with longer paths [45]. Lastly, eigenvector centrality provides a more complex measure of node influence by assigning relative scores to all concepts in the network, based on the number and quality of its relationships; a concept is significant to the extent that it is connected to other significant concepts [46].

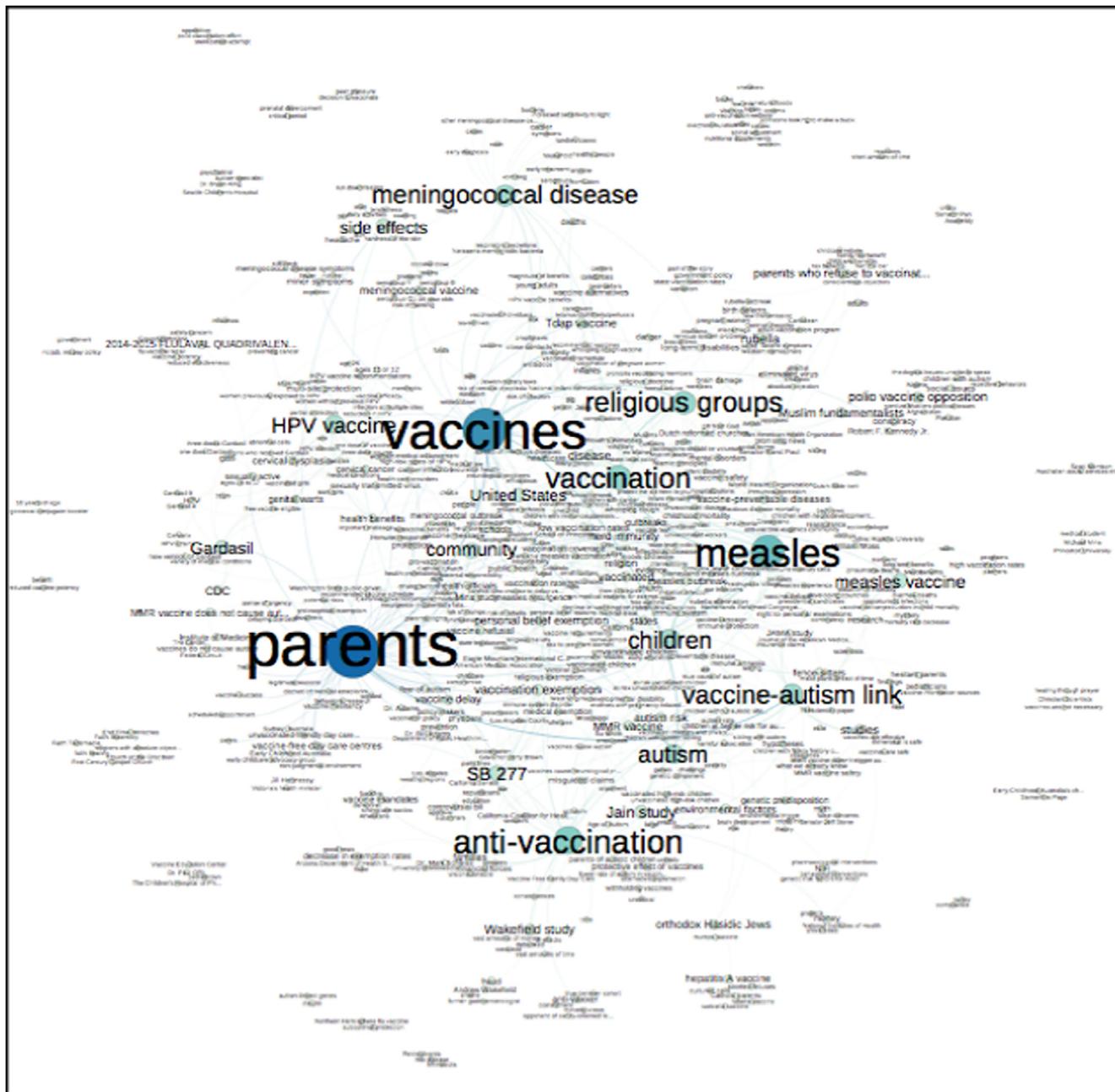
Community detection using the Newman-Girvan algorithm detects communities by consecutively removing each edge with

the highest edge betweenness from the graph [31]. Edge-betweenness refers to the number of shortest paths from one node to another that traverse through that edge. Cohesive groups in the network are measured by modularity, in which a good partition has more intra-community edges than expected at random; modularity values other than zero represent deviations from randomness [32].

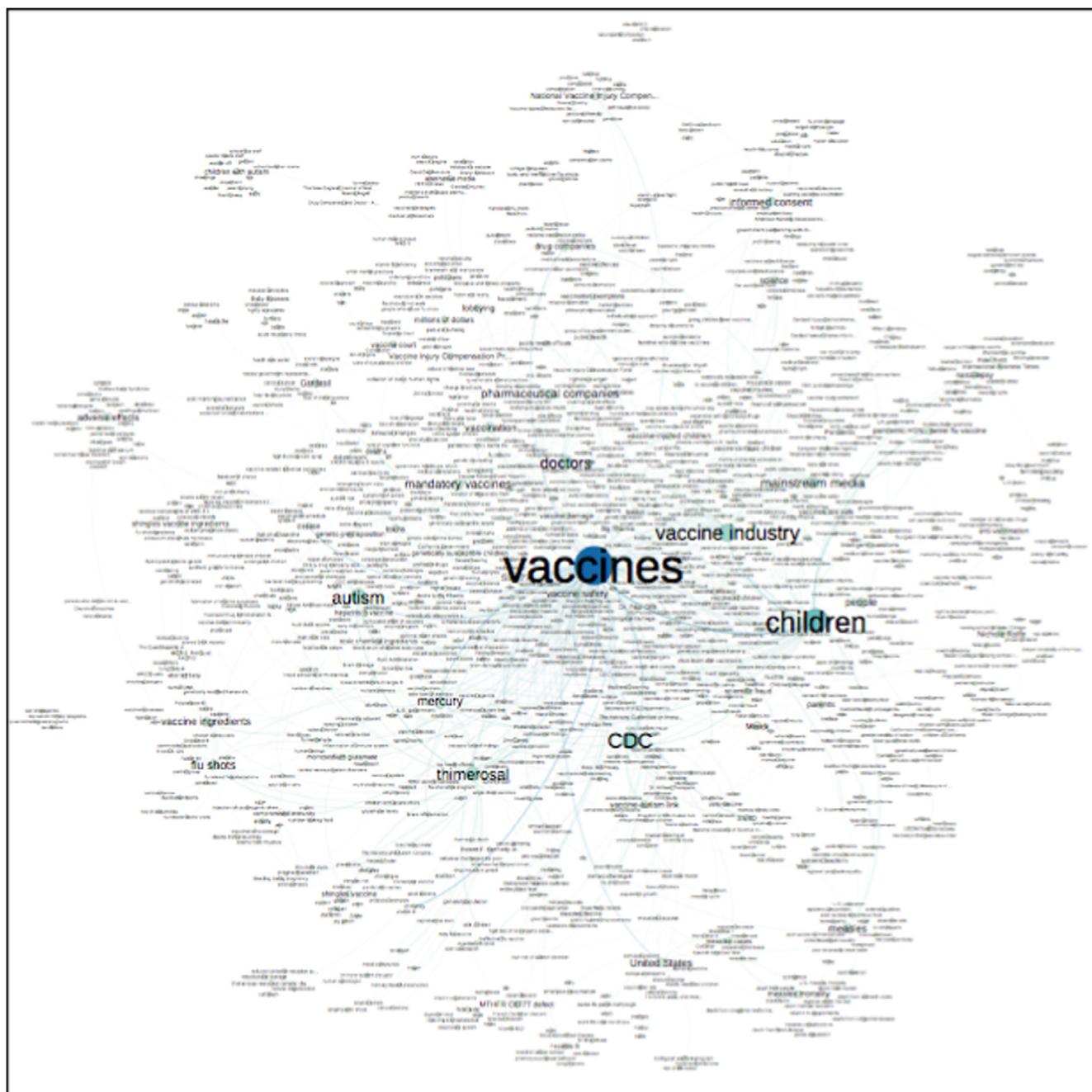
Appendix C. Network visualizations

[C1–C3]: Full semantic networks of vaccine sentiment. Visualizations for full semantic networks of [C1] positive vaccine sentiment, [C2] negative vaccine sentiment, and [C3] neutral vaccine sentiment. Node size represents betweenness centrality.

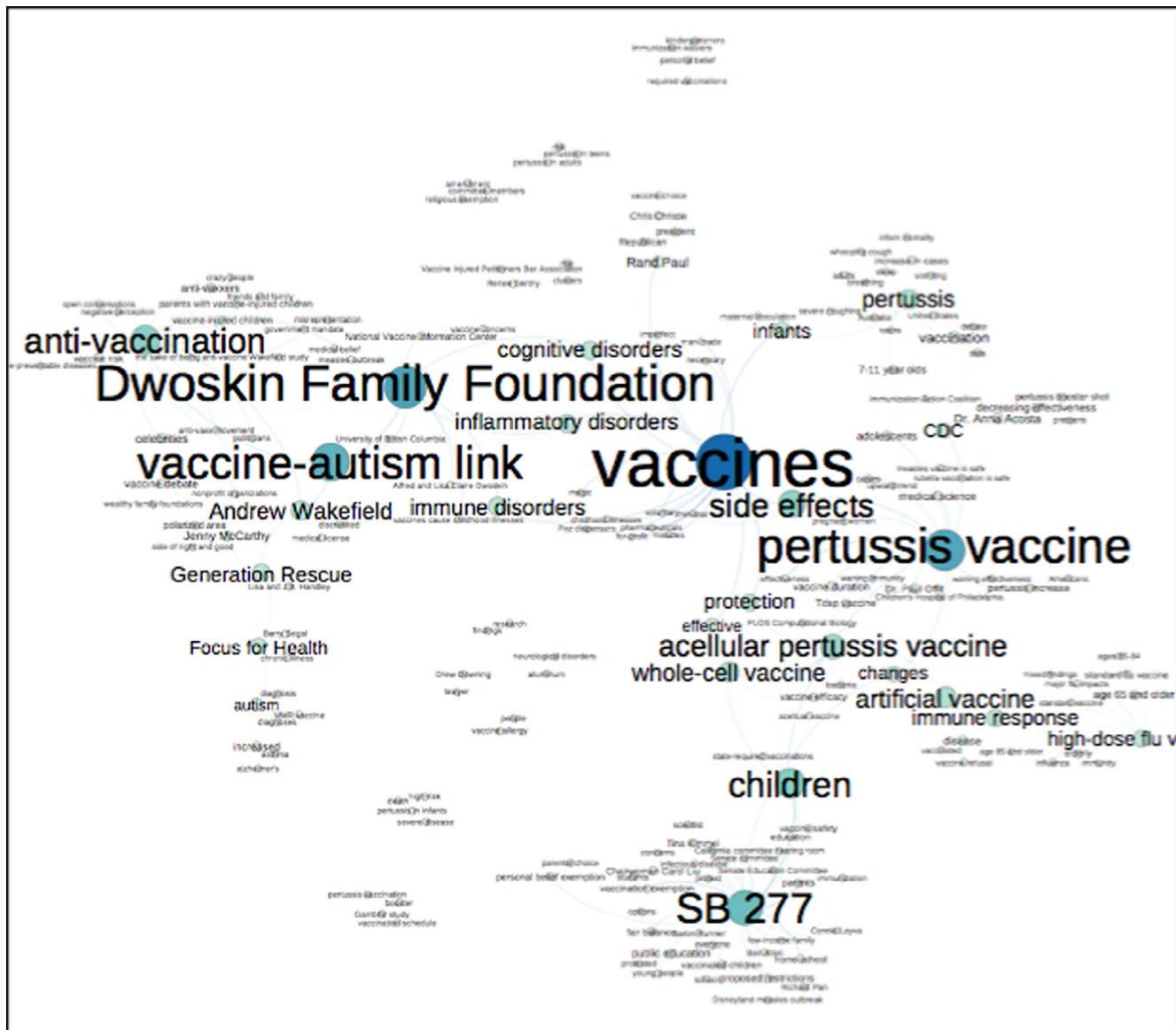
[C1] Full semantic network of positive vaccine sentiment



[C2] Full semantic network of negative vaccine sentiment

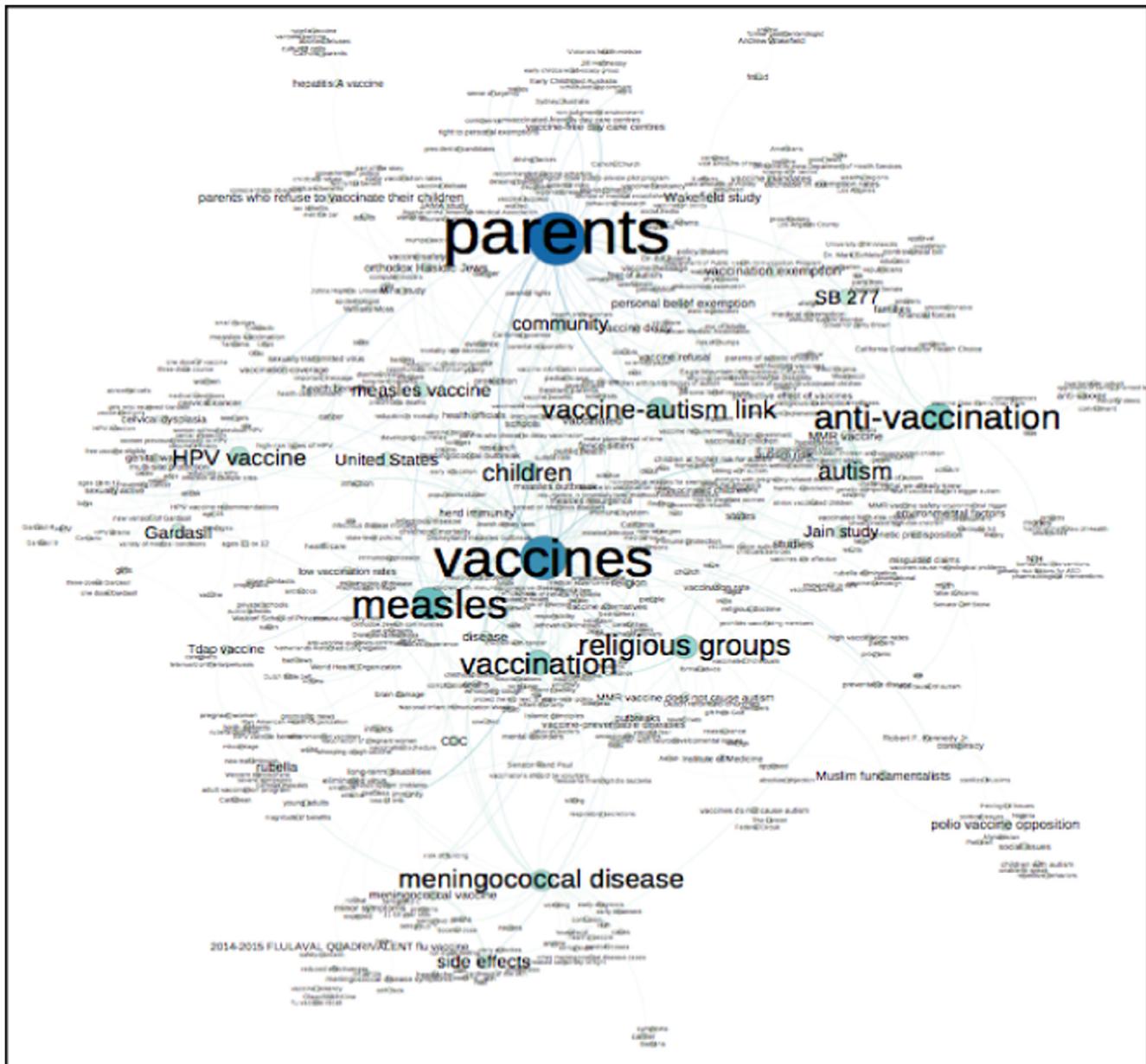


[C3] Full semantic network of neutral vaccine sentiment

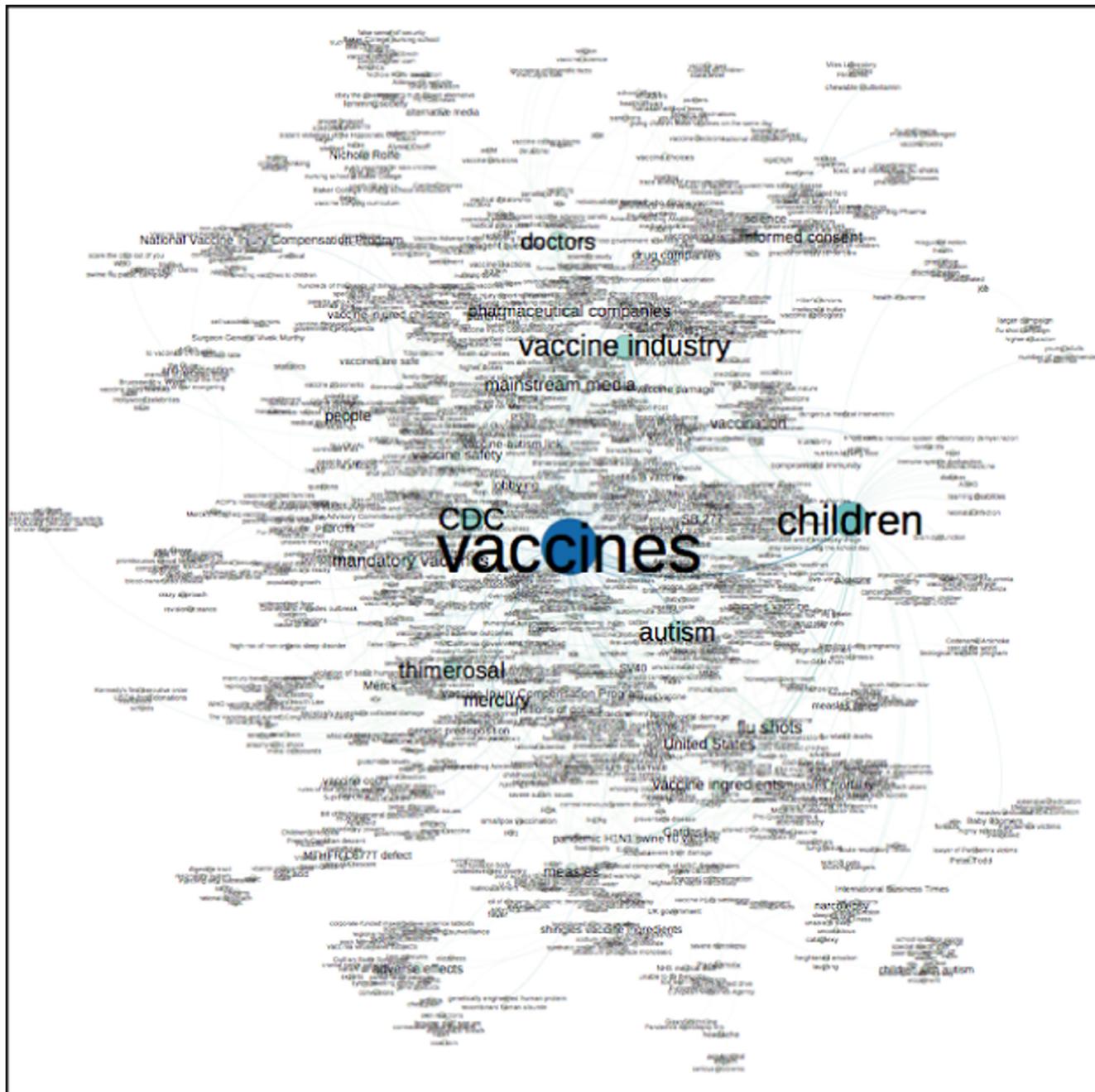


[C4–C6]: Greatest component subgraph of vaccine sentiment networks. Visualizations of the greatest component subgraph for networks of [C4] positive vaccine sentiment, [C5] negative vaccine sentiment, and [C6] neutral vaccine sentiment, where increasing node size represents greater betweenness centrality.

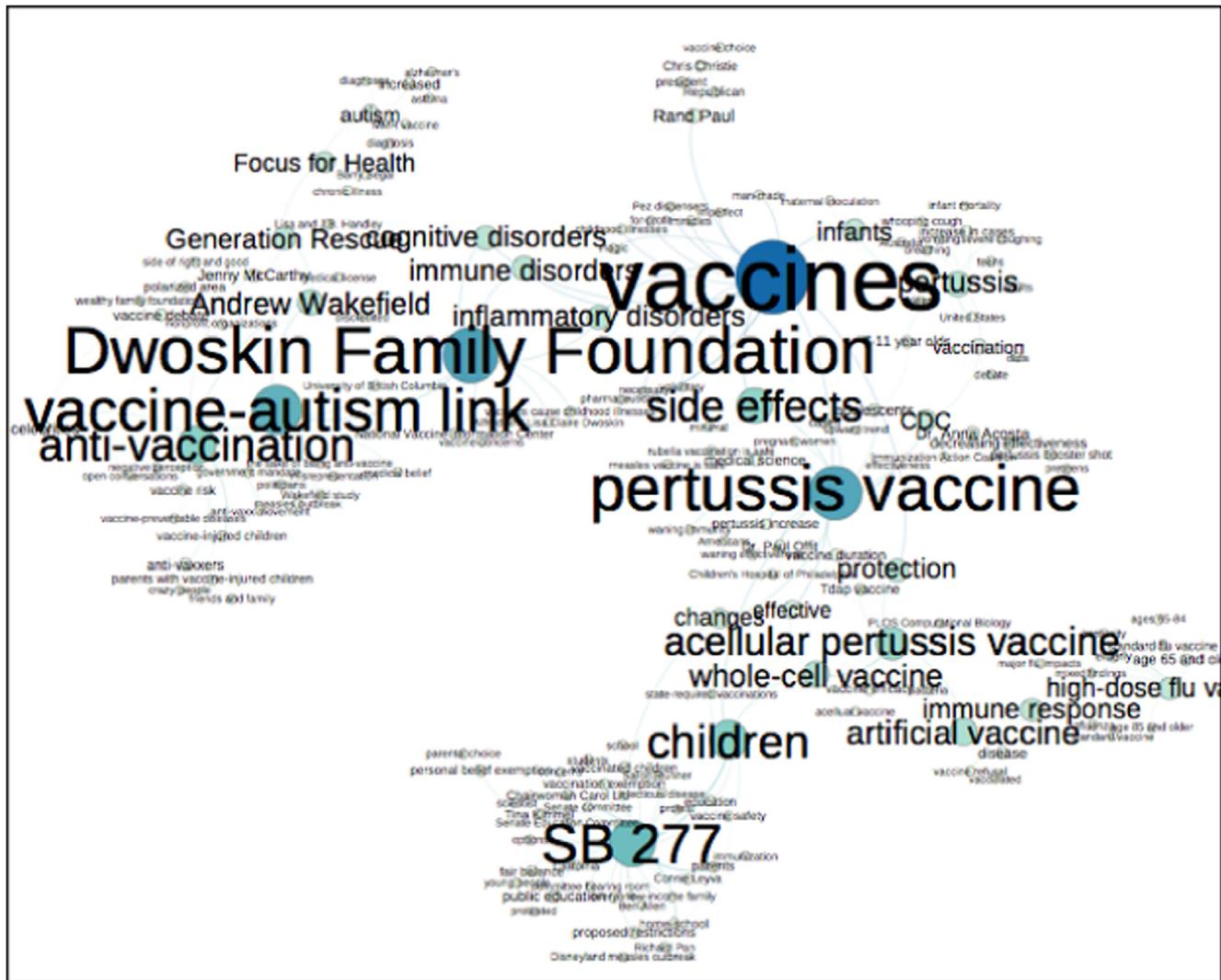
[C4] Greatest component subgraph of the positive sentiment network



[C5] Greatest component subgraph of the negative sentiment network



[C6] Greatest component subgraph of the neutral sentiment network



Appendix D. Degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality

[D1–D3]: Significant vaccine concepts by centrality. Centrality characterizes the importance, influence, or power of vaccine-related concepts in the semantic network. The table lists measures for the most central concepts (greater than 2 standard deviations from the network mean) by degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality for [D1] positive sentiment, [D2] negative sentiment, and [D3] neutral sentiment networks.

[D1] Most central nodes and centrality measures for the positive sentiment network.

Positive vaccine sentiment network						
Degree centrality		Betweenness centrality		Closeness centrality		Eigenvector centrality
	Mean = 0.0061 Std Dev = 0.0107		Mean = 0.006 Std Dev = 0.0203		Mean = 0.2292 Std Dev = 0.038	Mean = 0.0626 Std Dev = 0.0936
Vaccines	0.1079	Parents	0.2718	Parents	0.3687	Parents
Parents	0.0993	Vaccines	0.2176	Vaccines	0.3482	Vaccines
Measles	0.0993	Measles	0.1546	Children	0.3415	Measles
Vaccination	0.0856	Anti-vaccination	0.1261	Measles	0.3382	Vaccination
Autism	0.0616	Religious groups	0.1018	Community	0.3227	Children
HPV vaccine	0.0565	Vaccine-autism link	0.0917	Religious groups	0.3219	SB 277
Vaccine-autism link	0.0531	Meningococcal disease	0.0905	Autism	0.3188	Autism
Meningococcal disease	0.0531	Children	0.0825	SB 277	0.3158	Community
Anti-vaccination	0.0479	Autism	0.0799	Vaccine-autism link	0.3148	Religious groups
Children	0.0445	HPV vaccine	0.0732	Anti-vaccination	0.3121	Anti-vaccination
MMR vaccine	0.0411	Community	0.0574	Vaccination	0.3100	Vaccine-autism link
Religious groups	0.0394	SB 277	0.0571			Herd immunity
Measles vaccine	0.0377	Measles vaccine	0.0523			Vaccine refusal
SB 277	0.0342	Side effects	0.0510			Vaccination exemption
Disease	0.0308	Gardasil	0.0496			Personal belief exemption
Vaccination exemption	0.0291					Disease
Autism risk	0.0291					Measles vaccine
						Schools
						HPV vaccine
						Vaccine delay
						Meningococcal disease

[D2] Most central nodes and centrality measures for the negative sentiment network.

Negative vaccine sentiment network						
Degree centrality		Betweenness centrality		Closeness centrality		Eigenvector centrality
	Mean = 0.0027 Std Dev = 0.0058		Mean = 0.0033 Std Dev = 0.0148		Mean = 0.2161 Std Dev = 0.0365	Mean = 0.0318 Std Dev = 0.06
Vaccines	0.1054	Vaccines	0.3280	Vaccines	0.3582	Vaccines
Children	0.0623	Children	0.1889	Children	0.3375	Children
Thimerosal	0.0588	CDC	0.1274	Vaccine industry	0.3275	Thimerosal
CDC	0.0527	Vaccine industry	0.1213	Autism	0.3249	CDC
Vaccine industry	0.0518	Autism	0.1028	Mercury	0.3245	Vaccine industry
Autism	0.0386	Thimerosal	0.0869	Thimerosal	0.3209	Mercury
Doctors	0.0351	Doctors	0.0863	CDC	0.3197	Autism
Mainstream media	0.0351	Mercury	0.0629	SB 277	0.3072	Flu shots
Mercury	0.0334	Mainstream media	0.0624	Mainstream media	0.3070	Mainstream media
Flu shots	0.0263	Mandatory vaccines	0.0583	Flu shots	0.3037	Doctors
Pharmaceutical companies	0.0263	Flu shots	0.0576	Doctors	0.3028	SB 277
Mandatory vaccines	0.0255	Pharmaceutical companies	0.0552	Vaccine ingredients	0.2990	Vaccine ingredients
Vaccination	0.0237	Informed consent	0.0485	Mandatory vaccines	0.2969	Mandatory vaccines
SB 277	0.0228	People	0.0474	Toxic chemical ingredients	0.2958	Pharmaceutical companies
United States	0.0202	Vaccine ingredients	0.0453	Vaccine-autism link	0.2952	Vaccine-autism link
Measles	0.0193	United States	0.0449	Vaccine safety	0.2933	Toxic chemical

Appendix D (continued)

Negative vaccine sentiment network						
Degree centrality	Betweenness centrality		Closeness centrality		Eigenvector centrality	
Mean = 0.0027	Mean = 0.0033		Mean = 0.2161		Mean = 0.0318	
Std Dev = 0.0058	Std Dev = 0.0148		Std Dev = 0.0365		Std Dev = 0.06	
Vaccine ingredients	0.0184	Measles	0.0444	Intelligent questions	0.2905	ingredients
Informed consent	0.0184	Vaccination	0.0438	Vaccines are safe	0.2895	Aluminum
People	0.0184	Vaccine safety	0.0399			0.1889
Pandemic H1N1 swine flu vaccine	0.0184	Adverse effects	0.0354			0.1853
Merck	0.0184					Monosodium glutamate
Measles mortality	0.0184					Hepatitis B vaccine
						0.1793
						Vaccine-injured children
						Vaccine safety
						Evidence
						Informed consent
						Intelligent questions
						Formaldehyde
						Pregnant women
						Pandemic H1N1 swine flu vaccine
						Big Pharma
						Vaccines are safe
						Quackery
						Vaccine damage
						SV40
						Science
						0.1591
						0.1565
						0.1552
						0.1547
						0.1545
						0.1531

[D3] Most central nodes and centrality measures for the neutral vaccine network.

Neutral vaccine sentiment network						
Degree centrality	Betweenness centrality		Closeness centrality		Eigenvector centrality	
Mean = 0.0149	Mean = 0.0342		Mean = 0.1533		Mean = 0.0975	
Std Dev = 0.0204	Std Dev = 0.0839		Std Dev = 0.0296		Std Dev = 0.11	
SB 277	0.1824	Vaccines	0.5749	Vaccines	0.2335	SB 277
Vaccines	0.1118	Dwoskin Family Foundation	0.4092	Side effects	0.2208	Vaccines
Anti-vaccination	0.1059	Pertussis vaccine	0.3947	Pertussis vaccine	0.2199	Anti-vaccination
Pertussis vaccine	0.0824	Vaccine-autism link	0.3620	Whole-cell vaccine	0.2133	Parents
Pertussis	0.0824	SB 277	0.3294	Effective	0.2133	Children
High-dose flu vaccine	0.0647	Children	0.2643			Pertussis vaccine
		Anti-vaccination	0.2554			Home-school
		Side effects	0.2347			Education
		Acellular pertussis vaccine	0.2077			0.3540
						0.3209
						0.3206

[D4] Top ranked nodes by closeness vitality for the three networks.

Closeness vitality						
Negative sentiment network	Neutral sentiment network			Positive sentiment network		
Mean = 19148.407	Mean = 7029.871			Mean = 8449.754		
Std Dev = 24052.786	Std Dev = 16597.291			Std Dev = 9778.734		
Thimerosal	239154	Vaccines	127564	Meningococcal disease	79948	
MTHFR C677T defect	222220	Dwoskin family foundation	109972	Vaccination	77396	
Millions of dollars	210944	Vaccine-autism link	100468	Polio vaccine opposition	74438	
Children with autism	201122	SB 277	49768	Wakefield study	64018	
Measles mortality	179468	Acellular pertussis vaccine	48048	HPV vaccine	63748	
Vaccine court	172456	Artificial vaccine	43430	Vaccines	61934	

(continued on next page)

Appendix D (continued)

Closeness vitality					
Negative sentiment network		Neutral sentiment network		Positive sentiment network	
Mean = 19148.407		Mean = 7029.871		Mean = 8449.754	
Std Dev = 24052.786		Std Dev = 16597.291		Std Dev = 9778.734	
National vaccine injury compensation program	168948	Anti-vaccination	41638	Autism	61016
Anti-vaccination	145200	Generation Rescue	37594	Orthodox Hasidic Jews	55846
Measles	141736	immune response	34424	Measles	47038
Adverse effects	141140	Focus for Health	32640	Hepatitis A vaccine	44804

Appendix E. Data files

Data files and dynamic web-based interactive visualizations of semantic networks can be accessed online at: <http://dx.doi.org/10.1016/j.vaccine.2017.05.052>.

References

- [1] Zipprich J, Winter K, Hacker J, Xia D, Watt J, Harriman K, et al. Measles outbreak—California, December 2014–February 2015. *MMWR Morb Mortal Wkly Rep* 2015;64:153–4.
- [2] Peretti-Watel P, Patrick P-W, Ward JK, Schulz WS, Pierre V, Larson HJ. Vaccine Hesitancy: Clarifying a Theoretical Framework for an Ambiguous Notion. *PLoS Curr* 2015. <http://dx.doi.org/10.1371/currents.outbreaks.6844c80ff9f5b273f34c91f71b7fc289>.
- [3] Kang GJ, Culp RK, Abbas KM. Facilitators and barriers of parental attitudes and beliefs toward school-located influenza vaccination in the United States: Systematic review. *Vaccine* 2017;35:1987–95. <http://dx.doi.org/10.1016/j.vaccine.2017.03.014>.
- [4] Dubé E, Gagnon D, MacDonald NE. SAGE Working Group on Vaccine Hesitancy. Strategies intended to address vaccine hesitancy: Review of published reviews. *Vaccine* 2015;33:4191–203.
- [5] Betsch C, Böhm R. Detrimental effects of introducing partial compulsory vaccination: experimental evidence. *Eur J Public Health* 2015. <http://dx.doi.org/10.1093/ejph/rkv154>.
- [6] Betsch C, Böhm R, Chapman GB. Using behavioral insights to increase vaccination policy effectiveness. *Policy Insights from the Behavioral and Brain Sciences* 2015;2:61–73.
- [7] Larson HJ, Cooper LZ, Eskola J, Katz SL, Ratzan S. Addressing the vaccine confidence gap. *Lancet* 2011;378:526–35.
- [8] Garrett C, Wilson R, O'Leary M, Eckersberger E, Larson HJ. SAGE Working Group on Vaccine Hesitancy. Strategies for addressing vaccine hesitancy – A systematic review. *Vaccine* 2015;33:4180–90.
- [9] Hausman B. Vaccination and the Public in the 21st Century 2016.
- [10] Lawrence HY, Hausman BL, Dannenberg CJ. Reframing Medicine's Publics: The Local as a Public of Vaccine Refusal. *J Med Humanit* 2014;35:111–29.
- [11] Luke DA, Harris JK. Network analysis in public health: history, methods, and applications. *Annu Rev Public Health* 2007;28:69–93.
- [12] Salathé M, Bonhoeffer S. The effect of opinion clustering on disease outbreaks. *J R Soc Interface* 2008;5:1505–8.
- [13] Salathé M, Bengtsson L, Bodnar TJ, Brewer DD, Brownstein JS, Buckee C, et al. Digital epidemiology. *PLoS Comput Biol* 2012;8:e1002616.
- [14] Dredze M, Broniatowski DA, Smith MC, Hilyard KM. Understanding Vaccine Refusal. *Am J Prev Med* 2016;50:550–2.
- [15] Salathé M, Khandelwal S. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. *PLoS Comput Biol* 2011;7:e1002199.
- [16] Salathé M, Vu DQ, Khandelwal S, Hunter DR. The dynamics of health behavior sentiments on a large online social network. *EPJ Data Sci* 2013;2:4.
- [17] Gloer P, Diesner J. Semantic Social Networks. *Encyclopedia of Social Network Analysis and Mining*: Springer; 2014. p. 1654–9.
- [18] Eklund PW, Haemmerlé O. Conceptual structures: knowledge visualization and reasoning. In: 16th International conference on conceptual structures, ICCS; 2008. p. 7–11.
- [19] Doerfel ML. What constitutes semantic network analysis? A comparison of research and methodologies. *Connect*; 1998.
- [20] Ruiz JB, Barnett GA. Exploring the presentation of HPV information online: A semantic network analysis of websites. *Vaccine* 2015;33:3354–9.
- [21] Ruiz JB, Bell RA. Understanding vaccination resistance: vaccine search term selection bias and the valence of retrieved information. *Vaccine* 2014;32:5776–80.
- [22] Drieger P. Semantic network analysis as a method for visual text analytics. *Procedia - Soc Behav Sci* 2013;79:4–17.
- [23] Poland CM, Brunson EK. The need for a multi-disciplinary perspective on vaccine hesitancy and acceptance. *Vaccine* 2015;33:277–9.
- [24] Larson H, Leask J, Aggett S, Sevdalis N, Thomson A. A multidisciplinary research agenda for understanding vaccine-related decisions. *Vaccines* 2013;1:293–304.
- [25] Thovex C, Trichet F. Semantic social networks analysis. *Soc Netw Anal Min* 2012;3:35–49.
- [26] Revised report of the SAGE working group on vaccine hesitancy; 2014. http://www.who.int/immunization/sage/meetings/2014/october/SAGE_working_group_revised_report_vaccine_hesitancy.pdf?ua=1 (accessed April 25, 2017).
- [27] Schiltz JT, Lewis B, Eubank S. ChatterGrabber: A Lightweight Easy to Use Social Media Surveillance Toolkit. *Online J Public Health Inform* 2015;7. doi:10.5210/ojphi.v7i1.5717.
- [28] Wasserman S, Faust K. *Social network analysis: methods and applications*. Cambridge University Press; 1994.
- [29] Van Steen M. *Graph theory and complex networks*. di.unipi.it; 2010.
- [30] Freeman LC. Centrality in social networks conceptual clarification. *Soc Networks* 1978;1:215–39.
- [31] Newman MEJ, Girvan M. Finding and evaluating community structure in networks. *Phys Rev E Stat Nonlin Soft Matter Phys* 2004;69:026113.
- [32] Fortunato S. Community detection in graphs. *Phys Rep* 2010/2; 486: 75–174.
- [33] Batagelj V, Zaversnik M. An O(m) Algorithm for cores decomposition of networks. *arXiv [cs.DS]* 2003.
- [34] Burscher B, Vliegenthart R, de Vreese CH. Frames beyond words. *Soc Sci Comput Rev* 2016;34:530–45.
- [35] Schult DA, Swart P. Exploring network structure, dynamics, and function using NetworkX, permalink.lanl.gov; 2008.
- [36] Csardi G, Nepusz T. The igraph software package for complex network research. *Int J Complex Syst* 2006;1695:1–9.
- [37] Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. aaai.org; 2009.
- [38] An act to amend Sections 120325, 120335, 120370, and 120375 of, to add Section 120338 to, and to repeal Section 120365 of, the Health and Safety Code, relating to public health. n.d.
- [39] Goldstein S, MacDonald NE, Guirguis S. SAGE Working Group on Vaccine Hesitancy Health communication and vaccine hesitancy. *Vaccine* 2015;33:4212–4.
- [40] Miton H, Mercier H. Cognitive obstacles to pro-vaccination beliefs. *Trends Cogn Sci* 2015;19:633–6.
- [41] Dong X, Gabrilovich E, Heitz G, Horn W, Lao N, Murphy K, et al. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining, New York, NY, USA: ACM; 2014. p. 601–10.
- [42] Wasserman S, Faust K. *Social network analysis: methods and applications*. Cambridge University Press; 1994.
- [43] Freeman LC. Centrality in social networks conceptual clarification. *Soc Networks* 1978; 1: 215–39.
- [44] Wasserman S, Faust K. *Social network analysis: methods and applications*. Cambridge University Press; 1994.
- [45] Raad E, Chebir R. *Socio-graph representations, concepts, data, and analysis*. In: Alhaji PR, Rokne PJ, editors. *Encyclopedia of social network analysis and mining*. Springer: New York; 2014. p. 1936–46.
- [46] Gephi Heymann S. In: *Encyclopedia of social network analysis and mining*. New York: Springer; 2014. p. 612–25.
- [47] Koschützki D, Lehmann KA, Peeters L, Richter S, Tenfelde-Podehl D, Zlotowski O. Centrality Indices. In: Brandes U, Erlebach T, editors. *Network Analysis*. Springer-Verlag; 2005. p. 16–61.
- [48] Anderson JR. A spreading activation theory of memory. *J Verbal Learning Verbal Behavior* 1983;22:261–95.
- [49] Dell GS. A spreading-activation theory of retrieval in sentence production. *Psychol Rev* 1986;93:283–321.
- [50] Collins AM, Quillian MR. Retrieval time from semantic memory. *J Verbal Learning Verbal Behavior* 1969;8:240–7.
- [51] Collins AM, Loftus EF. A spreading-activation theory of semantic processing. *Psychol Rev* 1975;82:407–28.
- [52] Fazio RH, Sanbonmatsu DM, Powell MC, Kardes FR. On the automatic activation of attitudes. *J Personality Social Psychol* 1986;50:229–38.