Semantic Network Analysis of COVID-19 Vaccine Related Text from Reddit

Chad A. Melton[1,3], Jintae Bae[2], Olufunto A. Olusanya[3], Jon Hael Brenas[4], Eun Kyong Shin[2], Arash Shaban-Nejad[1,3]

[1]The Bredesen Center for Interdisciplinary Research and Graduate Education, University of Tennessee, Knoxville, USA

[2]Korea University, Seoul, South Korea

[3]Center for Biomedical Informatics, Department of Pediatrics, College of Medicine, University of Tennessee Health Science Center, Memphis TN, USA

[4]Sanger Institute, Cambridge, UK

chadmeltone@gmail.com, pureskw@korea.ac.kr, oolusan1@uthsc.edu, jb52@sanger.ac.uk, eunshin@korea.ac.kr,ashabann@uthsc.edu

Abstract. Vaccinations are critical and effective in resolving the current pandemic. With the highly transmissible and deadly SARS-CoV-2 virus (COVID-19), a delay in acceptance, or refusal of vaccines despite the availability of vaccine services poses a significant public health threat. Moreover, vaccine-related hesitancy, mis/disinformation, and anti-vaccination discourse are hindering the rapid uptake of the COVID-19 vaccine. It is urgent to examine how anti-vaccine sentiment and behavior spread online to influence vaccine acceptance. Therefore, this study aimed to investigate the COVID-19 vaccine hesitancy diffusion networks in an online Reddit community within the initial phase of the COVID-19 pandemic. We also sought to assess the anti-vaccine discourse evolution in language content and style. Overall, our study findings could help facilitate and promote efficient messaging strategies/campaigns to improve vaccination rates.

Keywords. Semantic Network Analysis, COVID-19 Vaccines, Misinformation, Online social media, Reddit

1. Introduction

The severe acute respiratory syndrome coronavirus 2, or SARS-COV-2 is responsible for the Coronavirus Disease 2019 (COVID-19) which has profoundly impacted the globe causing significant morbidity and mortality [1]. An unprecedented and accelerated effort was made to develop the COVID-19 vaccine since long-term pandemic containment and recovery were contingent upon the collective acceptance and uptake of the vaccine. Moreover, the COVID-19 vaccine is proven to be safe and effective at preventing life-threatening COVID-19 infections, hospitalizations, and deaths [2].

However, despite the many benefits, vaccine mis/disinformation, myths, false conspiracy theories, rumors, skepticism, and mistrust propagated through online digital platforms have driven vaccine hesitancy and compromised the effectiveness of scientific-based evidence on vaccines. As a result, vaccine hesitancy is classified by the World Health Organization to be among the 10 threats to global health [3].

Digital platforms provide the heuristic locus to fully comprehend the public discourse and sentiments regarding the COVID-19 pandemic [4], [5], [6], [7], [8], [9], [10], [11], [12]. Specifically, online digital data offers access to real-time which facilitates the collection, management, retrieval, mining, and interpretation of information to trace trends as well as gauge public sentiments, networks, and behaviors influencing vaccine acceptance. Therefore, in order to develop efficient messaging strategies and intervention campaigns to address the information crises and improve vaccination rates, it is essential to examine how multiple underlying factors ranging from personal, religious, political, social, etc. within the online interactive digital space influence vaccination decision making. However, a few studies have sought to understand the impacts of social networks and interactions on epidemic-related discourse within online digital platforms. Herein, our study objective was to examine how the public's anti-vaccine sentiment and behavior influenced vaccine acceptance via an online social media platform. This work is part of an ongoing study to explore vaccine-related content in social media with a focus on identifying, characterizing, and combating misinformation/disinformation.

2. Data

We collected approximately 300,000 posts and comments from 12 online communities or *subreddits* from the social media platform, Reddit. Subreddit members have the option to post links, images, videos, and text which community users can *upvote* or *downvote* based on their sentiment of the post and/or leave comments. The upvote/downvote system within Reddit is intended to increase the quality of the posts to minimize non-relevant material from a community. These communities often have rules that members must follow or risk the deletion of a post or a community ban. Though these community rules have the potential to create echo chambers, subreddits are typically less segregated than Facebook groups [13]. In our dataset, a few subreddits actively remove posts containing misinformation (e.g., r/CovidVaccine) but must rely on other users to report the occurrence to a moderator while other communities are more open to discussions involving misinformation and do not remove posts unless considered dangerous and detected/reported by Reddit officials. In some occurrences, entire subreddits have been removed due to rampant misinformation/disinformation, harassment, and threats of violence (r/NoNewNormal). To avoid this echo chamber effect, we attempted to choose an ideologically heterogeneous collection of subreddits (r/CovidVaccinated, r/Coronavirus, r/CovidVaccine, r/conspiracy_commons, r/COVID19_support, r/COVID19, r/AntiVaxxers, r/VACCINES, r/conspiracytheories, r/China_Flu, r/COVID19positive, r/conspiracy) with a total of 5.1 million. Data were

cleaned and queried for posts related to COVID-19 vaccines/vaccination. Our final amalgamated dataset consisted of 31432 posts/comments authored by 20429 users between January 1, 2020, and Oct 31, 2021. Finally, these posts/comments received a total of approximately 1.26 million votes, indicating a high degree of community interaction. The majority of subreddits were similar in posting frequency for the first months of our timeframe. Posting in several subreddits rapidly increased over time with the greatest increase occurring in April 2021 as a result of more widespread vaccine availability (e.g., Figure 1).

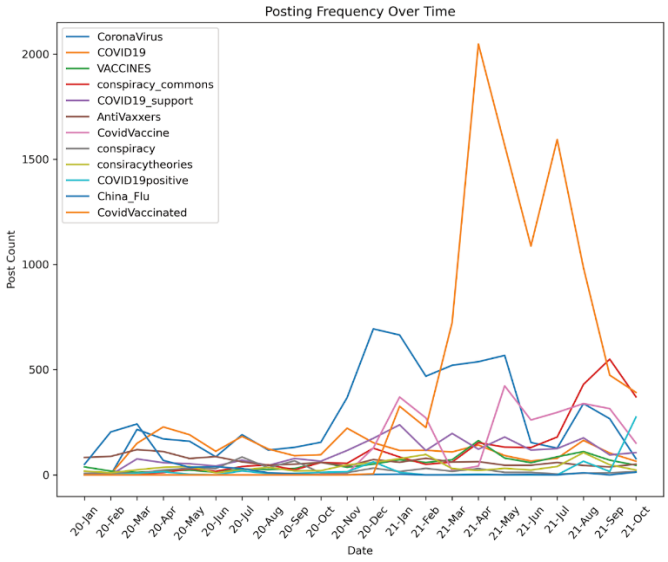

Figure 1: Reddit Posting-Frequency Over Time. Subreddits are symbolized by lines of varying color.

3. Methods

Our approach can unveil the unbounded network structure and facilitate vaccine hesitancy discourse online. To accomplish this task, we conducted a temporal semantic network analysis to observe graph evolution over time. Network analysis [14], [15]] uses a graphical representation of nodes and edges to provide insight into data that may not be observable upon the surface. In computational semantic network analysis, *nodes* represent *tokenized* words while *Edges* represent a connectedness between nodes. Our multistep analysis was conducted with the *Python* library, *Networkx* [16], [17]. After removing *stop words* (e.g., determiners, conjunctions, and prepositions*)*, lemmatizing and normalizing the corpus (i.e., converting a word to its base form), a word co-occurrence network was created with *nx.Graph()*. Due to the *hairball* effect that occurred while including each word in a monthly corpus, networks included in this document were limited to less than or equal to 160 nodes by setting an appropriate weight threshold determined by iterating through the corpus. After removing isolates, *betweenness*

centrality (BC) networks were created for each month in our dataset. Essentially, BC displays the importance of a node (i.e., word) based on calculating the number of times a node is included in the shortest route between other nodes [18]. Ego networks were centered around the word "vaccine" and built for each month. To detect the tendency for clustering to occur within our data, transitivity was calculated for the complete corpus as well as for the scaled-down versions with limited node counts. Graph density and networks/subreddit statistics were collected. Lastly, networks were also visually inspected to verify coherence in our interpretation of the results.

4. Results

Overall, the *Giant* network node and edge count generally increased with time, exhibiting fluctuations in certain months. The *Vaccine Ego (VE)* network behaved similarly although the node and edge count were much less due to the nature of an ego network. Transitivity values for the complete corpus or *Giant* network decreased over time and ranged between 0.13 (Dec 2020) and 0.21 (Jan 2020) with an average value of 0.15. Greater node clustering occurred within this downtrend during Aug 2020, Oct 2020, April 2021, and July 2021.

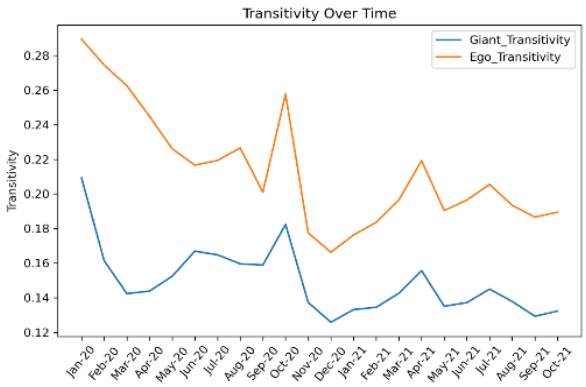

Figure 2. Transitivity Over Time. The orange line represents the *Vaccine* Ego network and the blue line represents the *Giant* network transitivity.

The *VE* net exhibited similar characteristics with increased clustering in June 2020, Oct 2020, April 2021, and July 2021. Density for both *Giant* and *VE* networks tended to ebb and flow from month to month but decreased over time as well due to an increasing post quantity (e.g., Figure 2 and Figure 3).

For the betweenness centrality networks, centrality ranged from 0.21 for the node *immunity* to 0.892 for the node *vaccine*. Variations of the word "vaccine" (i.e., vaccine, vaccines) appeared in the top 10 highest BC values throughout each month in our data set. Furthermore, we observe changes in centrality values related to nodes that represent some terms common in COVID-19 misinformation. For example, in September 2020 (e.g., Figure 4), the nodes *vitamin* and *d* are connected to a few terms related to

COVID19. Centrality for the nodes was calculated to be 0.16 and 0.20 respectively and amongst the top five for the month.

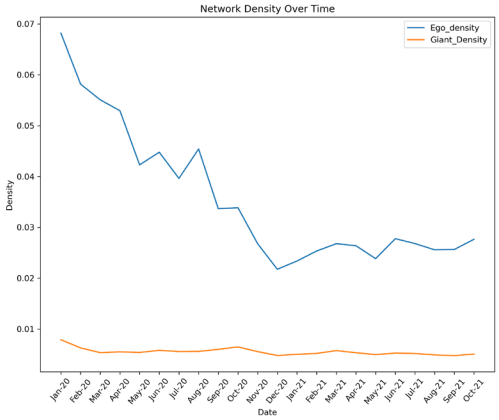

Figure 3. Network Density Over Time. The blue line represents the *Vaccine* Ego network density and the orange line represents the *Giant* network density.

As the occurrence of these two nodes diffuses over time, the centrality values diminished substantially to *vitamin* (0.0004725 and rank 513/5894) and *d* (0.00055 and rank 461/5894) in October 2020.

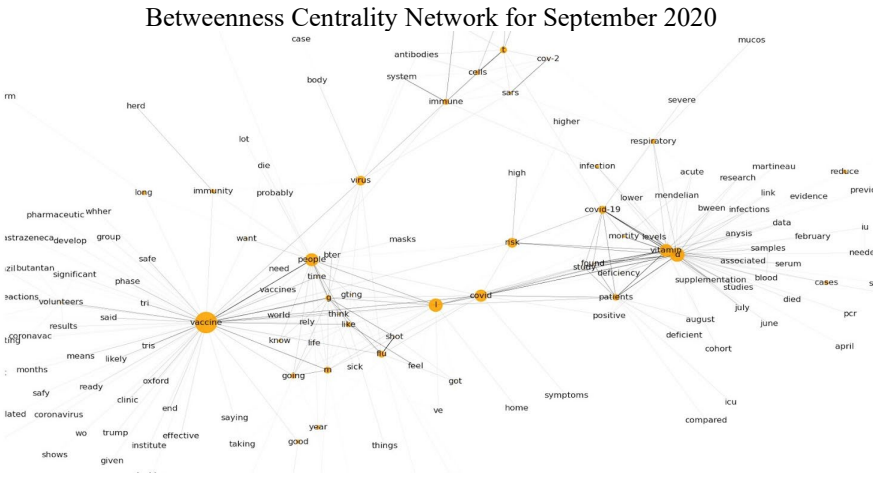

Figure 4. Betweenness Centrality for September 2020. Nodes are indicated by orange circles and edges are indicated lines. Node size is reflective of the weight and edge thickness indicative of interconnectedness between nodes. Note the vaccine centered cluster and the vitamin and d centered clusters are mainly connected by the Covid node in between the two clusters.

However, *d* (0.001193 and rank 253/9301) and *vitamin* (0.00168 and rank 177/9301) both increase but increase again in December 2020. Nodes indicating vaccine hesitancy were also observable in our networks and represented by example keywords such as *scared* or *worried*. However, visual inspection of some comments revealed the intent of vaccination even though the user experienced anxiety related to the vaccine. For example, a comment from April 2021 said, "I am going to get my shot today! Half excited, half scared. Not scared from like conspiracy theory stuff lol, but I have had systemic allergic reactions before, so yeah a little nervous there."

The ego network focused on the node *vaccine* provides interesting insight into vaccine discussion directly related to COVID-19 related terms. In February 2020, the *vaccine* is significantly connected to the nodes *flu, people*, and *g (*i.e., G-protein). The node *coronavirus* is present in this network. However, the connection to the central node is so infrequent that edges are not visible (see Figure 5a). Moving six months forward to August 2020, the flu node is still present but is far less prevalent as before, as terms related to COVID-19 (e.g., sars, covid, virus have begun to dominate online discussion (see Figure 5b). The network for October 2020 continues to display connectedness with terms related to COVID-19 along with other terms now typical of this pandemic (i.e., masks, and deaths). This month also begins to display other terms related to vaccine testing and manufacturers (i.e., Moderna, placebo, and efficacy) (see Figure 5c). The greatest increase in node/edge quantity and second largest spike for transitivity occurs in April 2021. With the successful rollout of COVID-19 vaccines, a multitude of terms appears related to vaccination, types of vaccines, most vaccine manufacturers, as well as a wide variety of side-effects ranging in severity (see Figure 5d).

5. Discussion

As we expected, changes in our networks reflect the dynamic conditions and events that have occurred since the first COVID-19 cases were detected. Semantic as well as network structural changes are observable in the *Giant* and *VE* networks in several significant shifts as COVID-19 spread throughout the world. For example, in early Jan 2020, a small number of nodes are visible representing words associated with COVID 19 but the *Giant* network does not display interconnectedness between the node *Vaccine* and other nodes representing the COVID19. These occurrences rapidly increase as infection rates climb and online discussion shifts towards vaccines for COVID-19. Nodes tend to reflect conversation regarding side effects (e.g., fever, sore arm, body aches, etc.) as vaccines become more readily available. In our data set, the large increase in posts from r/CovidVaccinated in April 2021 contributes to vaccine side-effects interconnectedness and appearance as well. Moreover, this occurrence was also expected based on previous topic modeling studies (Melton et al 2021). Unfortunately, nodes representing misinformation keywords become more apparent as the interconnectedness with the node "Vaccines" increases in conjunction with COVID-19 keywords (e.g., Vitamin D, autism, Bill Gates, Big Pharma) in several months. Visual analysis of the raw text data suggested a wide range of vaccine hesitancy behaviors as well. These behaviors included hesitancy due to fear of vaccine side effects, feelings of

"threatened freedoms", false expertise, ignorance of how vaccines work, "big pharma motivated pandemic conspiracy, anti-vaccination beliefs, and many others (see https://github.com/Cheltone/W3PHIAI-2022 for supplementary materials including plots, data, and tables).

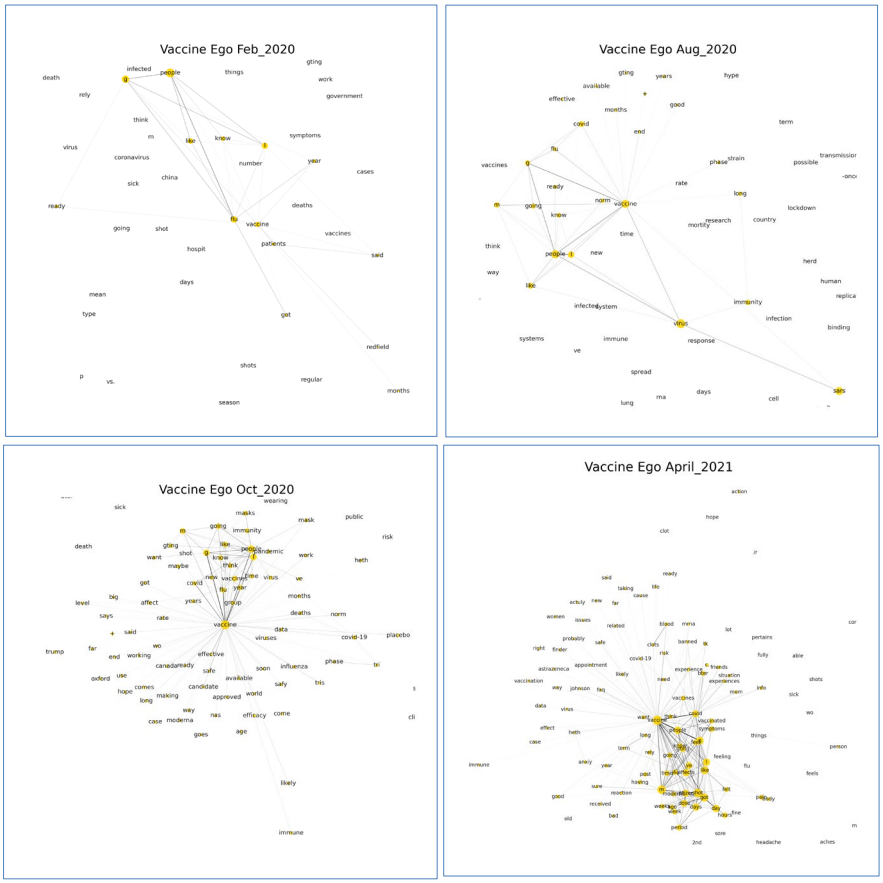

Figure 5. a (top left), b (top right), c (bottom left), d (bottom right): *Vaccine Ego* Network for February 2020, August 2020, October 2020, and April 2021. Nodes are indicated by gold circles and edges are indicated lines. Node size is reflective of the weight and edge thickness indicative of interconnectedness between nodes.

Our study has some limitations. Though great care was taken to create an unbiased data set, the possibility for some potential biases still exists, including selection bias from our choice of subreddits. Obtaining purely non-biased data is a challenging aspect of many scientific domains, and especially important with sentiment analysis, topic modeling, and semantic network modeling because unsupervised learning methods may cluster topics in noncoherent ways. Results of unsupervised classification methods are often challenging to evaluate for similar reasons. Conducting a manual semantic analysis [19] of a sample of our dataset could offer further insights

into discourse occurring in social media. Moreover, utilizing graph mining algorithms or comparison with topic modeling (Latent Dirichlet Allocation) techniques could bolster our results. Lastly, it is conceivable that the significant increase in posts in r/CovidVaccinated that occurred in April 2021 could have overwhelmed the network structure with nodes concerning vaccine side effects in graphs from April 2021 through October 2021. A comparison of the data set without r/CovidVaccinated could reveal discussions involving misinformation/disinformation in the other subreddits.

6. Conclusion

We conducted a betweenness centrality analysis of the Vaccine ego networks using approximately 31,000 comments/posts harvested from 12 subreddits. Our analysis found significant mentioning of COVID-19 and COVID-19 vaccine misinformation/disinformation, along with other vaccine-hesitant content. Ongoing work by our team is focusing on exploring other measures (i.e., semantic centrality, degree centrality, eigenvector centrality, PageRank) as well as tracing the diffusion on nodes specific to misinformation throughout our data set. Future work will also explore the evolution of semantic networks along with user activities. Because users from different thought communities contribute differently to semantic networks, it is crucial to understand both users and their activities. Tracing activity logs of Reddit users along with their posted contents, we expect to detect the diffusion pathways and their associated focal actors. We also plan to work on a user-subreddit bipartite network to examine the dynamics of vaccine discourse throughout a wider community level. These next steps will ultimately guide the development of a precision digital intervention tool that can target misinformation.

References

1- Ortiz-Prado, E., Simbaña-Rivera, K., Gómez-Barreno, L., Rubio-Neira, M., Guaman, L. P., Kyriakidis, N. C., ... & López-Cortés, A. (2020). Clinical, molecular, and epidemiological characterization of the SARS-CoV-2 virus and the Coronavirus Disease 2019 (COVID-19), a comprehensive literature review. *Diagnostic microbiology and infectious disease*, *98*(1), 115094.

2- Rosenberg ES, Holtgrave DR, Dorabawila V, Conroy M, Greene D, Lutterloh E, Backenson B, Hoefer D, Morne J, Bauer U, Zucker HA. New COVID-19 cases and hospitalizations among adults, by vaccination status—New York, May 3–July 25, 2021. Morbidity and Mortality Weekly Report. 2021 Sep 17;70(37):1306.

3- World Health Organization (WHO) . (2019) Ten threats to global health in 2019. https://www.who.int/news-room/spotlight/ten-threats-to-global-health-in-2019

4- Broniatowski, D. A., Paul, M. J., & Dredze, M. (2013). National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. PloS one, 8(12), e83672.

5- Brownstein, J., & Freifeld, C. (2007). HealthMap: the development of automated real-time internet surveillance for epidemic intelligence. Euro Surveill, 12(11), E071129.

6- Culotta, A. (2010). Towards detecting influenza epidemics by analyzing Twitter messages. Paper presented at the Proceedings of the first workshop on social media analytics.

7- Fung, I. C.-H., Tse, Z. T. H., & Fu, K.-W. (2015). The use of social media in public health

surveillance. Western Pacific Surveillance and Response Journal, 6(2), 3-6.
8- Gu, H., Chen, B., Zhu, H., Jiang, T., Wang, X., Chen, L., . . . Jiang, J. (2014). Importance of Internet surveillance in public health emergency control and prevention: evidence from a digital epidemiologic study during avian influenza A H7N9 outbreaks. Journal of medical Internet research, 16(1), e20.
9- Mollema, L., Harmsen, I. A., Broekhuizen, E., Clijnk, R., De Melker, H., Paulussen, T., Das, E. (2015). Disease detection or public opinion reflection? Content analysis of tweets, other social media, and online newspapers during the measles outbreak in The Netherlands in 2013. Journal of medical Internet research, 17(5).
10- Salathé, M., Freifeld, C. C., Mekaru, S. R., Tomasulo, A. F., & Brownstein, J. S. (2013). Influenza A (H7N9) and the importance of digital epidemiology. The New England journal of medicine, 369(5), 401.
11- Shin, E. K., & Shaban-Nejad, A. (2017). Public Health Intelligence and the Internet: Current State of the Art. In A. Shaban-Nejad, J. S. Brownstein, & D. L. Buckeridge (Eds.), Public Health Intelligence and the Internet (pp. 1-17). Cham: Springer International Publishing.
12- Zhang, E. X., Yang, Y., Di Shang, R., Simons, J. J. P., Quek, B. K., Yin, X. F., . . . Ling, V. R. Y. (2015). Leveraging social networking sites for disease surveillance and public sensing: the case of the 2013 avian influenza A (H7N9) outbreak in China. Western Pacific Surveillance and Response Journal, 6(2), 66-72.
13- Cinelli, M., Morales, G. D. F., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, *118*(9).
14- Melton C, Olusanya OA, Shaban-Nejad A. Network Analysis of COVID-19 Vaccine Misinformation on Social Media. Stud Health Technol Inform. 2021 Nov 18;287:165-166. doi: 10.3233/SHTI210839. PMID: 34795104.
15- Shin EK, Shaban-Nejad A. Applied Network Science for Relational Chronic Disease Surveillance. Stud Health Technol Inform. 2019 Jul 4;262:336-339. doi: 10.3233/SHTI190087. PMID: 31349336.
16- Developers, NetworkX. "NetworkX documentation." (2012).
17- Srinivasa-Desikan, B. (2018). Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras. Packt Publishing Ltd.
18- Linton, C. (1977). Freeman. A set of measures of centrality based on betweenness. *Sociometry*, *40*(1), 35-41.
19- Brien S, Naderi N, Shaban-Nejad A, Mondor L, Buckeridge DL. Vaccine Attitude Surveillance using Semantic Analysis: Constructing a Semantically Annotated Corpus. *WWW (Companion Volume) 2013,* Rio de Janeiro, Brazil, 13-17 May 2013, ACM Press, pp. 683-686. Doi:10.1145/2487788.2488023.