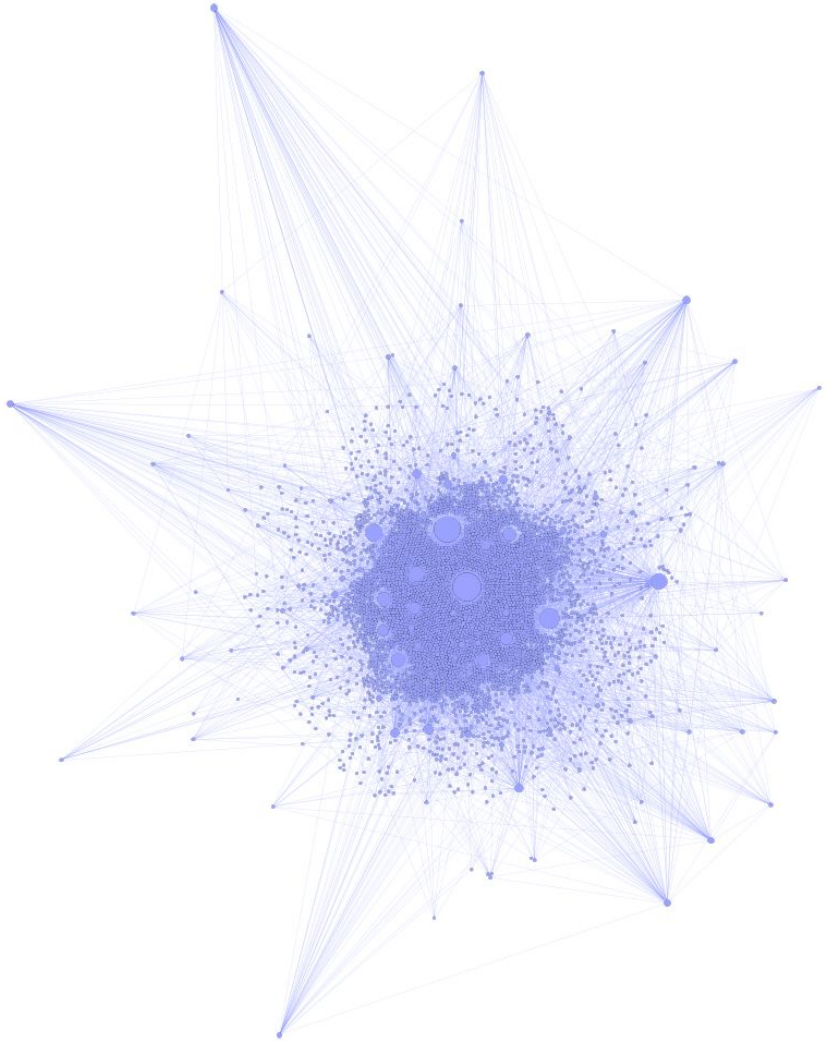
Brief Analysis & Visualization on Community Detection of Anti-Vaccine Web-Sphere in Feb 2015



By: Joshua Sinamo

Methods

Data Source

- Seed links : top 200 google links from February 2015 with keyword : anti AND (vacci OR vax)
- Crawler : issuecrawler.net
    - Depth = 2 (recursive depth from seeds)
    - Sampling method: snowball sampling

Visualization and Analytical Tools

- Gephi
- Python

<u>Graph visualization algorithm</u>

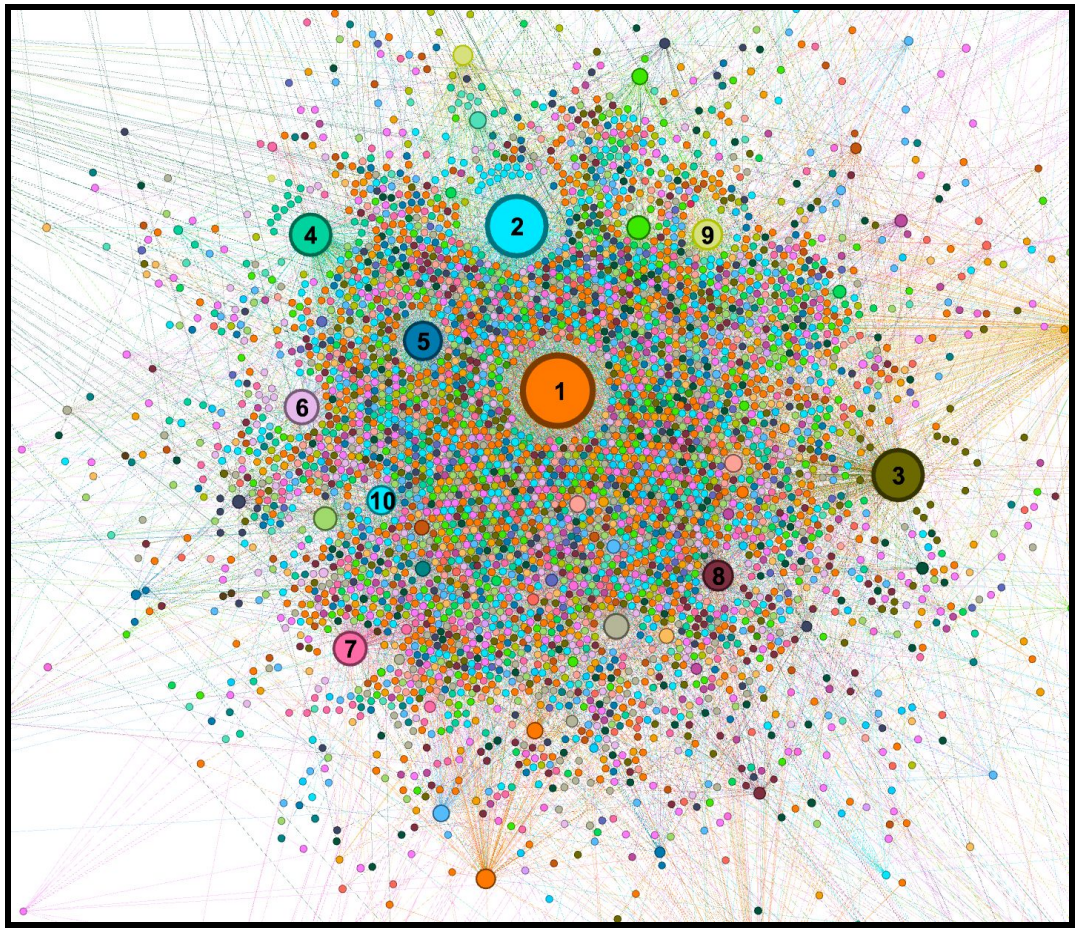
- Force Atlas 2 - https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0098679#pone.0098679-Newman
  Jacomy, Mathieu, et al. "ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software." *PLOS ONE*, Public Library of Science, 10 June 2014, journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0098679#pone.0098679-Newman2.
- Settings:
    - Dissuade Hubs : True
        - Hubs which attract less will be pushed to the edges
        - Subsequently, the more attractive / popular a node is, the more central it will be placed within the graph
        - Lin Log Mode = TRUE ; tightened distance between nodes of same cluster
    - Prevent Overlap = True
    - Clustering method: Louvain modularity

<u>Key points</u>

- February 2015 was chosen as the time-frame of our seed-links because between 2004-present, the aforementioned timestamp gained highest interest (check: https://trends.google.com/trends/explore?date=all&geo=US&q=anti%20vaccine)
- Best modularity score (louvain) = 0.684 with 35 community partitions. In this case, it is quite evident that there's no monopoly in information sources (people did not get information from a single source only) and multiple facets of discussion in the topic of vaccine hesitancy.

Measure: In-degree centrality

Fig.1 : Central view of the anti-vaccine online community with node sizes based on in-degree centrality coefficients



Rank	Site
1	http://www.theguardian.com/
2	http://www.washingtonpost.com/
3	http://www.nytimes.com/
4	http://www.vox.com/
5	http://www.motherjones.com/
6	http://www.snopes.com/
7	http://www.politico.com/
8	http://www.npr.org/
9	http://www.theatlantic.com/
10	http://www.cnn.com/

The size of the nodes in Fig.1 was determined by the number of in-degree centrality of that node. For this particular case, a high number of in-degree centrality means that there were a lot of hyperlinks pointing toward them. These hyperlinks oftentimes are citations to a news outlet or academic papers. Our top 10 sites with highest in-degree centrality coefficients are all news outlets.

Measure: Out-degree centrality

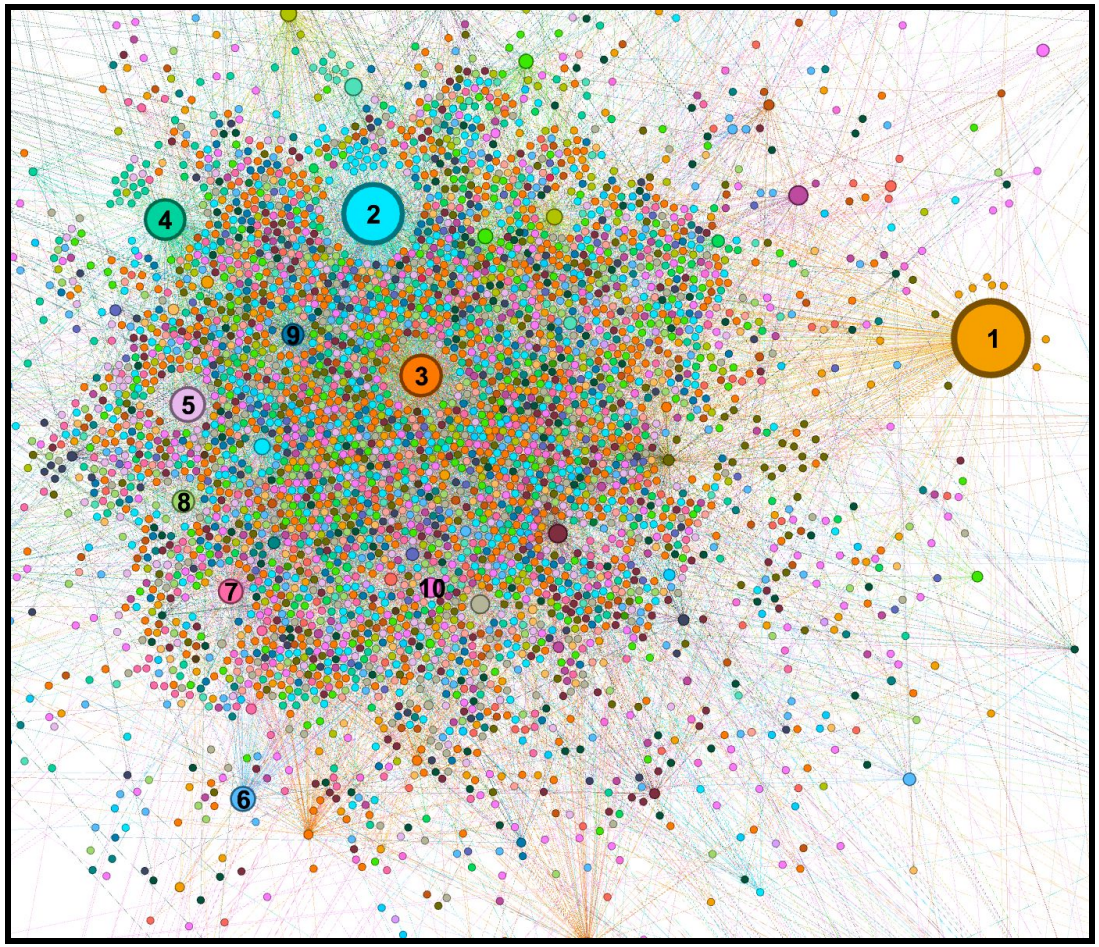
Fig. 2 : Central view of the anti-vaccine online community with node sizes based on out-degree centrality coefficients

Rank	Site
1	http://www.scienceblogs.com
2	http://www.caffeineandfairydust.com
3	http://reason.com
4	http://www.washingtonpost.com
5	http://www.wsj.com
6	http://www.huffpost.com
7	http://www.forbes.com
8	http://blogs.seattletimes.com
9	http://www.vox.com
10	http://www.usatoday.com

The size of the nodes in Fig.2 was determined by the number of out-degree centrality of that node. For this particular case, higher coefficient of out-degree centrality means that these sites referenced other sites more frequently than the other. In general, these hyperlinks oftentimes are news outlets (their adverts) and discussion forums. Our top 10 sites with highest out-degree centrality coefficients are all news outlets.

Measure: Betweenness centrality

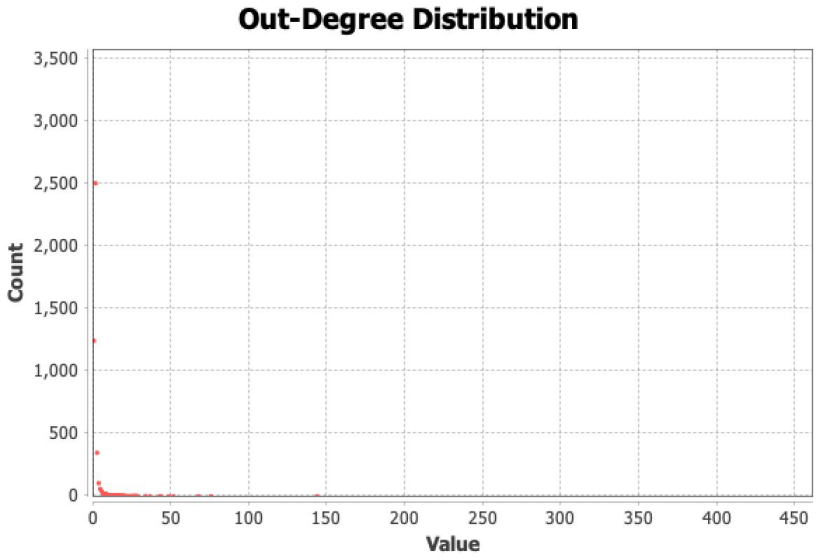
Fig. 3 : Central view of the anti-vaccine online community with node sizes based on betweenness centrality coefficients

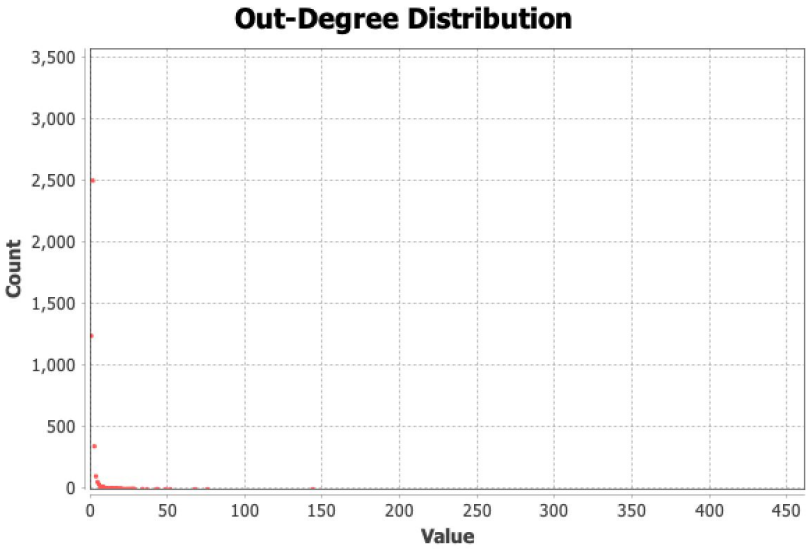


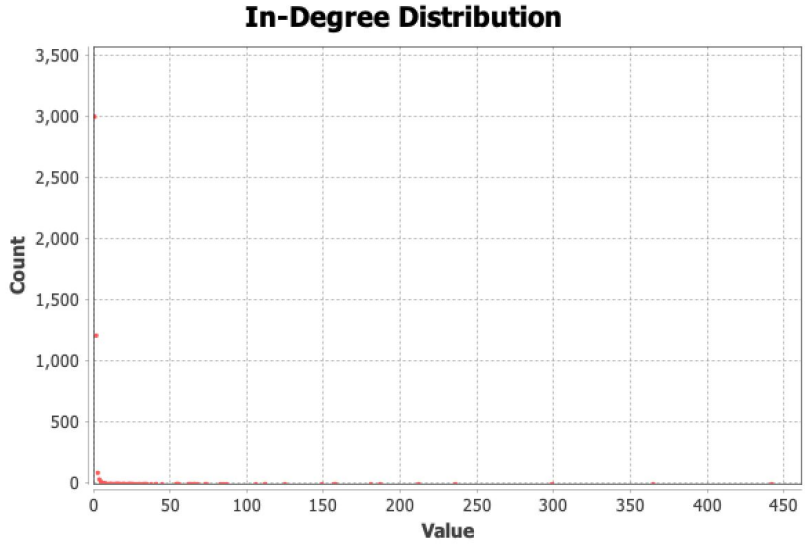
Rank	Site
1	http://www.scienceblogs.com
2	http://www.washingtonpost.com
3	http://www.theguardian.com
4	http://www.vox.com
5	http://www.snopes.com
6	http://www.forbes.com
7	http://www.politico.com
8	http://www.usatoday.com
9	http://www.motherjones.com
10	http://www.huffingtonpost.com

The size of the nodes in Fig.3 was determined by the coefficient of betweenness centrality of that node. For this particular case, higher coefficient of betweenness centrality means more control over the network as more information would pass through that node. In general, these hyperlinks oftentimes are news outlets (with their adverts) and discussion forums. Our top 10 sites with highest betweenness centrality coefficients are mainly news outlets, personal sites, and a scientific website as #1.

Distributions







All the centrality measures that we have reviewed so far follows the power law distribution. This means that the narrative of that community was potentially controlled by a handful of information outlets in particular news websites. Academic sources such as research papers are oftentimes not designed to be consumed by average Americans and evidently no academic portals ranked highly in any of the centrality measures.