The Tip of the Iceberg

Medical Misinformation and the Internet

Jeanne Reppert

December 3, 2020

0.1 Project Abstract

text to go here after completing the project

0.2 Introduction

As the Covid-19 pandemic has unfolded, the popularity of websites reputed for spreading misinformation has grown significantly. Some attribute this growth to the initial lack of information about the virus and the adjustments that medical professionals made to public policies as new information was revealed. Distrust has grown as public policies encroached on citizen's perceived freedoms. Anti-vaxxers, who have strong organizational structures and an existing following, harnessed public distrust to amplify the uncertainty surrounding Covid-19 related policies. Many leaders propagate conspiracy theories and use fear tactics to gain support. Key public health officials and philanthropists such as Bill and Melinda Gates have been targeted as villians.

Identifying and understanding online misinformation is essential for public health governance. In addition to identifying and flagging problematic websites, public health officials must directly refute conspiracy theories and specific scientific misinformation spread onlines. By understanding rhetorical patterns and themes utilized, a set of standards might be developed to assess websites and rate their reliability and trustworthiness.

This study investigates the website sponsored by Joseph Mercola (Mercola.com) and its relationship to health misinformation to understand the interaction between public health issues and the polarization surrounding Covid-19 and vaccinations. Mercola.com is known as one of the largest purveyors of "alternative" health information. His website has shared information through the publication of articles since 2008. Mercola earns his living by directing his followers to an online store where vitamins and other health related products are sold. According to Stephen Barrett of Quackwatch.org (https://quackwatch.org/11ind/mercola/), Mercola had a net worth in 2017 of more than $100 million.

Mercola's website states that he has more than a million subscribers. A typical email subscriber will recieve one daily newsletter with links to three articles. While two of those articles usually reference a short description about a supplement or health practice, the third article will often contain information that often would typically fall under the category of misinformation. Typically topics in this category would be common conspiracy theories, anti-vaccination information, and currently often debunked information pertaining to the Covid-19 virus.

0.3 Methods

In order to investigate the tactics Mercola uses to gather followers and spread health information (and misinformation), the full corpus of archived articles published by Mercola from 2008 to 2020 were scraped from his website. Utilizing RSelenium and rvest, 9306 URL's were identified for potential scraping. Some of the URL's identifed had broken links and were eliminated. Additionally, URL's that were not associated with health articles or with little text (either recipes, exercise illustrations or videos) were also omitted. Once these URL's were eliminated, the text, date, and title were scraped from a total of 8676 articles and organized into CSV files by year. Additionally, CrowdTangle was used to gather all available posts made on Mercola's group Facebook page and a history of tweets from Mercola were gathered. The data was investigated to note patterns and trends in the texts of the website articles and how those compared to the social media content.

0.4 Hypotheses

Websites that pass on medical misinformation have a tendancy to follow a patttern of communication. Most websites have some innocuous articles that pertain to health issues including articles about supplements, healthy eating, exercise, and health lifestyles. These articles may be atttractive to those who are mildly interested in health topic but not typically drawn to conspiracy theories that stretch the imagination more than the idea of trying a vitamin or other supplement. These types of articles may be a type of lure or "gateway drug" that draws the user in. Most websites press the viewer to share an email in order to view articles and once that email is captured, most websites will send a daily news digest that highlights additional articles. Thus over time, the user is exposed to a variety of topics including the more nefarious conspiracy theories that would be categorized as medical misinformation. One might expect then to observe in the text data, that during times when uncertainty is more prevalent due to a health crisis, these topics may skew more towards current events and during times when uncertainty in public health is less of a concern that topics will skew more towards articles about healthy lifestyles. One may also hypothesize that social media sites, for this type of misinformation, are more utilized to form connections to readers and to draw traffic to the website rather than to necessarily gain converts to the social media site itself. The real value in a tweet or a facebook post in this case is not necessarily to generate discussion or retweets but to gain clicks that will lead users to the website for the purpose of capturing their email and following up with additional emails and ultimatley product advertisements.

0.5 Dataset descriptions

For this study, due to the large volume of text involved with 13 years of articles, four years of articles were chosen for study. Each year represents a time frame when the United States was tasked with fighting a new virus that emerged from worldwide sources. In 2009, the H1N1 flu virus was considered to be at pandemic levels in the United States. The Ebola virus reached the shores of the United States in 2012 albeit in very small numbers, and in 2016 the Zika virus raised concerns around health professionals. In 2020, the SarsCov2 virus sparked the most serious pandemic throughout the United States since the 1918 Spanish flu pandemic.

Since Mercola's site was first launched in 2008, his 2009 article archive is relatively small with a total of 231 articles. By 2012 his collection for the year had grown to 639 years. In 2016, 793 articles were published and made available in the archived material and in 2020 (up to the beginning of November) the number of articles available was already up to 730 and is certain to grow over the next few weeks. According to the website, Mercola "fact checks" his archives which leads one to believe that he has a staff of ghost writers who are responsible for his website content.

The collection of Facebook archived posts are comprised of a csv file of 25,027 rows of data. Variables include number of page likes, number of comments, post likes, other post responses, shared links, videos, number of times videos are viewed, the post text and a score for overperformance. Most posts included range from mid-2009 to August 2019. His page has 1.78 million followers and 1.79 million likes. As of August 20, 2019, Mercola stopped posting on his Facebook page although posts by Mercola are still widely circulated on Facebook. He announced that he was leaving Facebook and includes a link to his website for more information. On his website, he encourages members to subscribe to his email, SMS messages and podcasts for daily information. Additionally, he encouraged his followers to find him on Parler. Mercola still maintains several Facebook groups in other languages. His Facebook page in Spanish has almost 1 million followers.

Twitter history for Mercola is more limited as Twitter limits the number of tweets that may be gathered to 3200. This smaller set for this reason, is comprised of tweets from July 4th, 2019 to November 13th, 2020. Variables include the tweet date, text, user description, total favorites, total followers, total friends, user location, user name, total statuses, tweet mentions, retweet mentions, tweeted urls's, url retweets and total retweets. Mercola has a total of 290.1K followers are his Twitter page. Mostly, Mercola tweets out his web page article links on a daily basis.

0.6 Tools and Libraries Used

A variety of r packages have been used to gather and clean the data collected. RSelenium and rvest have been utilized to scrape data from the website. Massmine was used to gather tweets and CrowdTangle was used to gather data about Facebook posts. Tidyverse, tidytext, widyr, tm, dplyr, readr and stringr were all utilized to process, clean and sort the data. Snowballc was used for stemming text and ggplot2, gridExtra, igraph, ggraph, wordcloud and plotly were all utilized to produce various visualizations. Sentiment analysis was performed utilized the syuzhet package. The package fs was used to load and manage files.

0.7 Visualization of Mercola Website Articles: Top Words for Each Year

Given that Mercola had significantly fewer articles in 2009 and the articles in his first couple of years were less verbose, the data for 2009 is quite different in number in comparison to the other three years. This needs to be kept in mind when investigating the results. In all four years, the words vitamin, health and food all appear to figure heavily in the articles surveyed. Risk, disease and research also appear as some of the top words for all four years. In 2009, 2012, and 2020, the word vaccine appears as one of the top three words in each year with nearly 800 mentions (out of 231 articles), but more than 4500 mentions (out of 639 articles.) In 2016, when 793 articles were published, the number of mentions of vaccines drops to around 1000, but rises again in 2020, when greater than 5000 mentions of vaccines are made throughout 730 articles. In 2009 and 2012, closely associated with the occurences of the word vaccine, the word mercury also occurs. In 2009 flu and its counterpart influenza is the third most mentioned word (the year when H1N1 peaked in the U.S.) However, neither Zika or Ebola are found in the top words in the years where this virus was an issue in the U.S. perhaps because not as many individuals werre directly impacted by these viruses. In 2020, in particular, the words covid being the second most frequently mentioned word with more than 4500 mentions alone. Associated words such as sars, coronvirus, masks, pandemic, and virus also are top words. While not found in the list of top words in 2020, cancer is also a top mentioned word in 2009, 2012, and 2016.

0.8 Visualization of Mercola Website Articles: Word Correlations with Top 9 Words

0.8.1 Top correlated words for Mercola Articles in 2009

Two primary topics appear to emerge when investigating word correlations for the Mercola 2009 dataset. The first topic appears to be centered around health related topics. For most of the top 9 words, correlated words are associated with major diseases, women's issues, vitamins, and food appear most frequently. Evidence of discussion regarding the H1N1 virus is also present and appears to center around the flu vaccine. Generally the correlated words to not appear to imply danger or any sort of ominous feeling.

0.8.2 Top Correlated Words for Mercola Articles in 2012

Top correlated words in this datset appear to be more varied. Primary topics still appear to revolve around health related topics with some attention to diet, exercise and nutrition. However, in this dataset there appears to be a greater emphasis on health malfunction particularly as it pertains to cancer, diabetes, and heart issues. The words associated with cancer, drug, disease, risk and vaccine all seem to connotate a more somber discussion with associated words like death, safe, chronic, and carcinogen included in the correlated word list. Interestingly, the vaccine list of correlated words contains two references to nvic which is the vaccine study organization that is funded by Mercola. This group focuses on descriptions and arguments about the "dangers" of vaccines. Additionally, two other organizations are mentioned: the FDA and CDC.

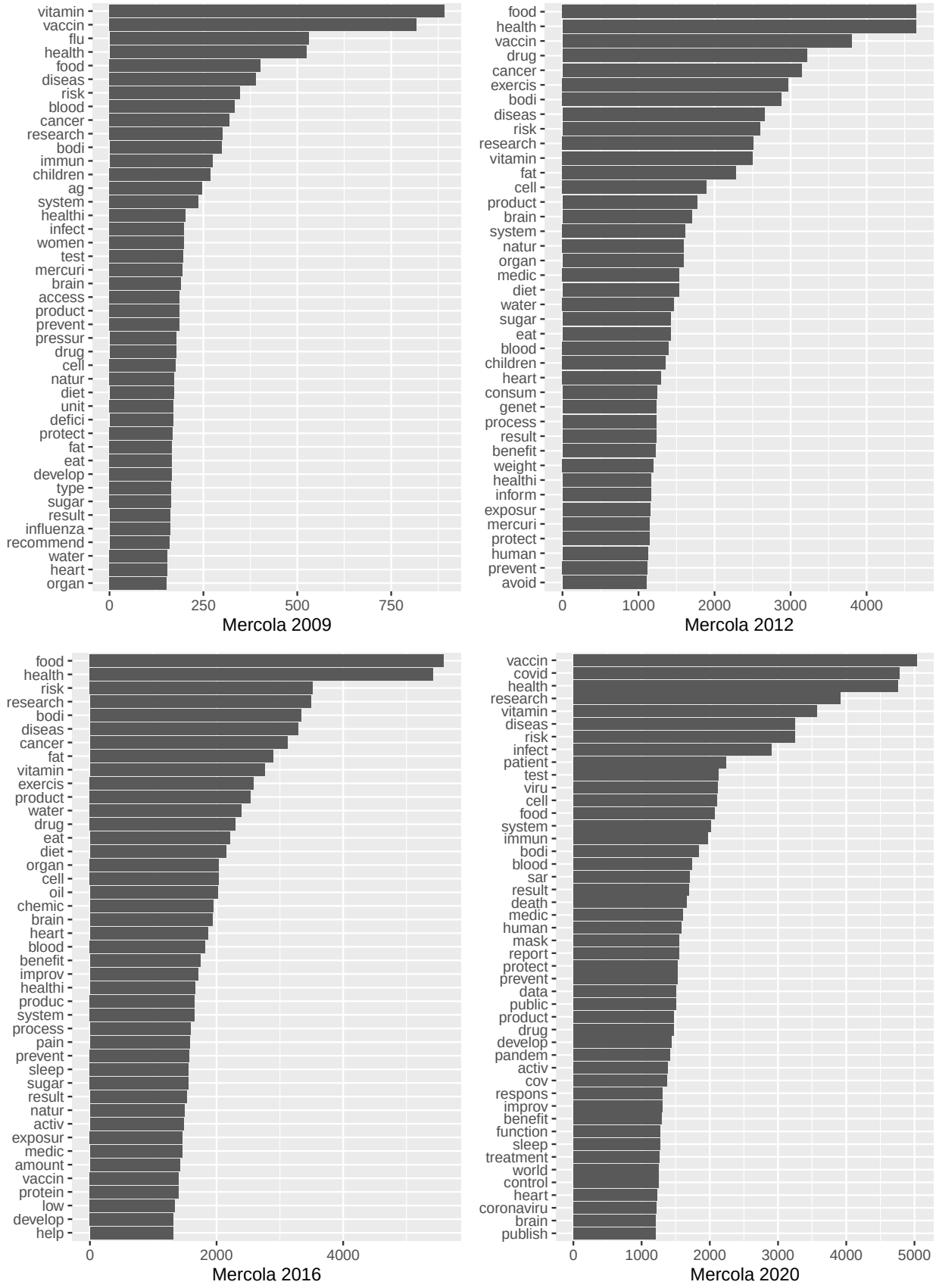


Figure 1: Top Words for Each Year

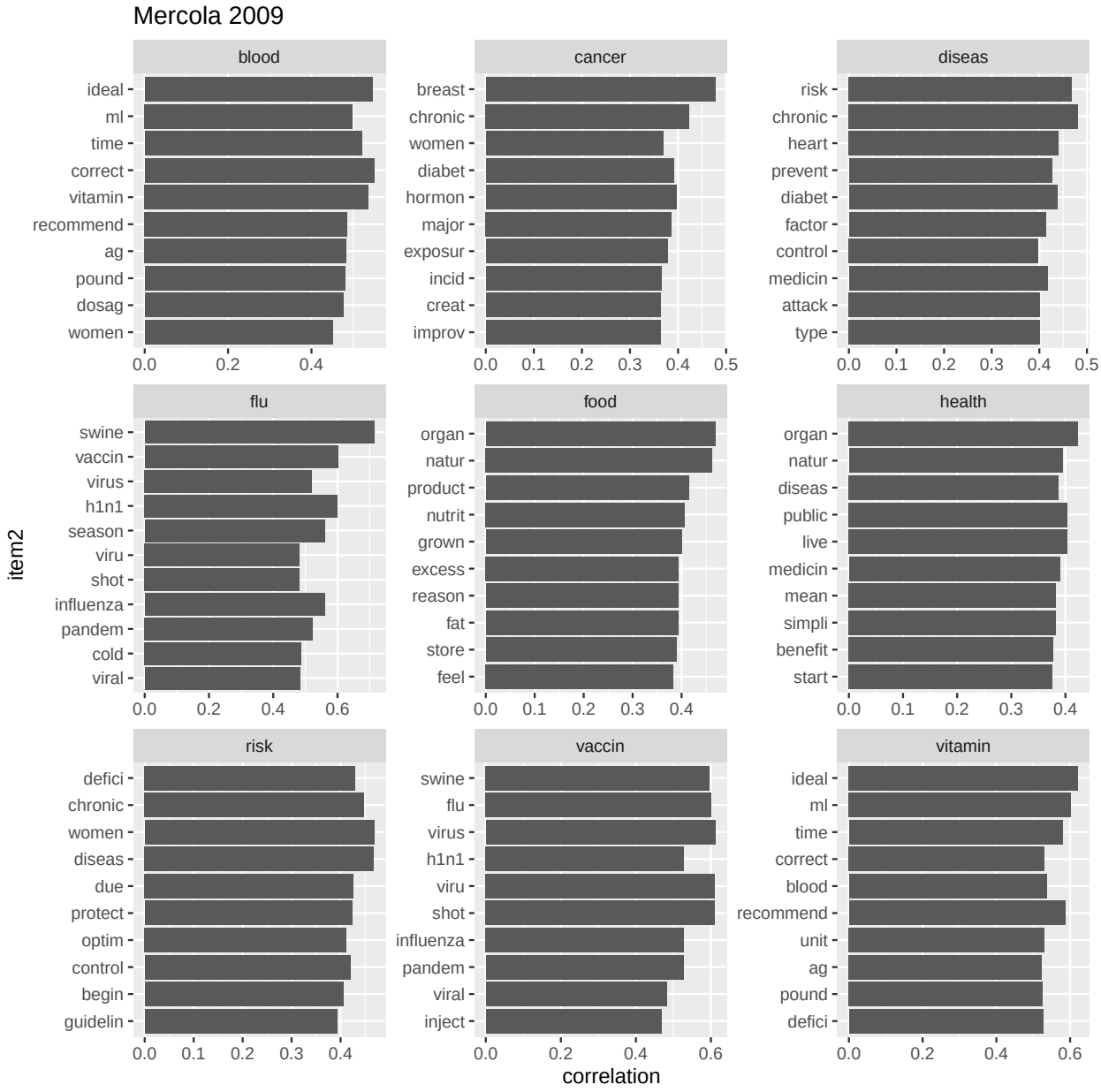


Figure 2: word correlations for top 9 words in 2009 Mercola articles

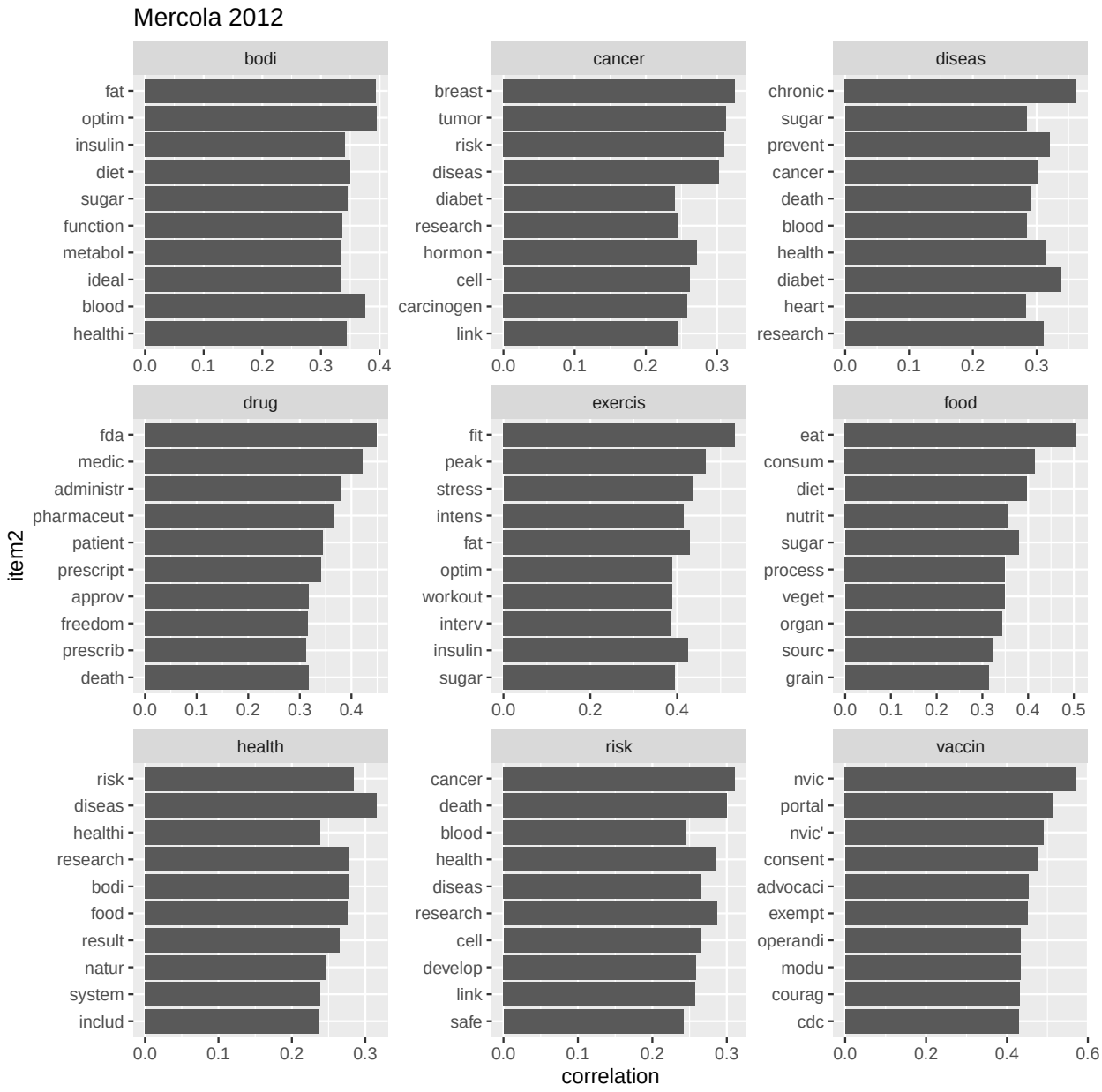


Figure 3: word correlations for top 9 words in 2012 Mercola articles

0.8.3 Top Correlated Words for Mercola Articles in 2016

Top correlated words in the 2016 articles again appear to move towards more light hearted subjects. Emphasis is placed on diet, nutrition, natural products, and more specific mentions of vitamins. Even categories for cancer, risk and disease appear to have less of a variety of words and the topics still seem to revolve around food intake.

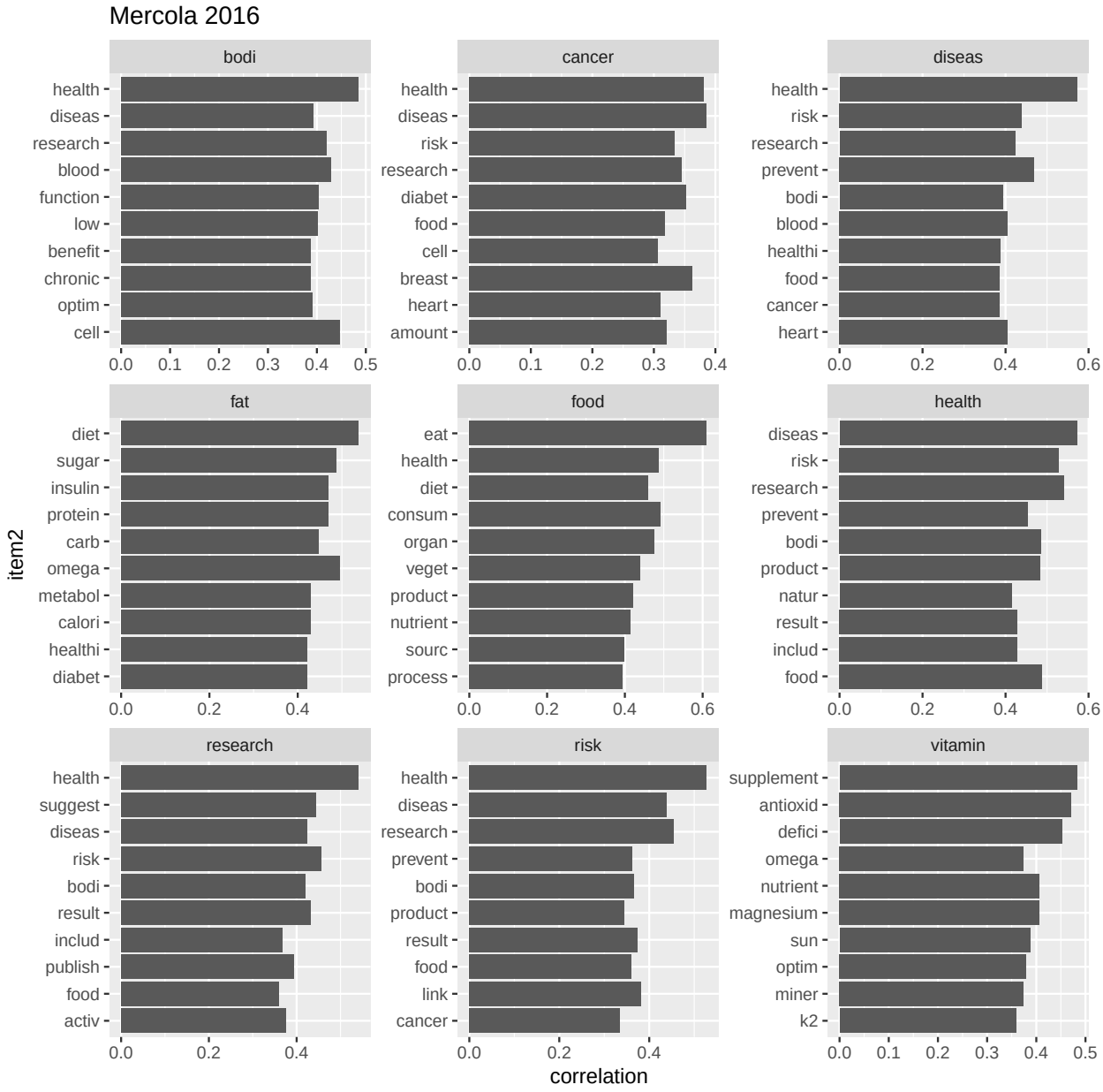


Figure 4: word correlations for top 9 words in 2016 Mercola articles

0.8.4 Top Correlated Words for Mercola Articles in 2020

Not surprisingly, 2020 has a unique list of top words in Mercola's articles and their correlated words connotate a far more serious discussion. The topics of the articles have almost fully departed from health, nutrition, and

exercise topics and have switched to discussions that involve aspects of the pandemic. The correlated words for vaccine include both references to the pandemic but also to Bill Gates, the government and mandates. In the only mention of nutrition in the correlated words for vitamin, specific vitamins that have been popularly touted for immune building effects against covid are listed (vitamin D and Zinc.) More ominous words such as respiratory, severe, mortal, and infection and more commonly mentioned

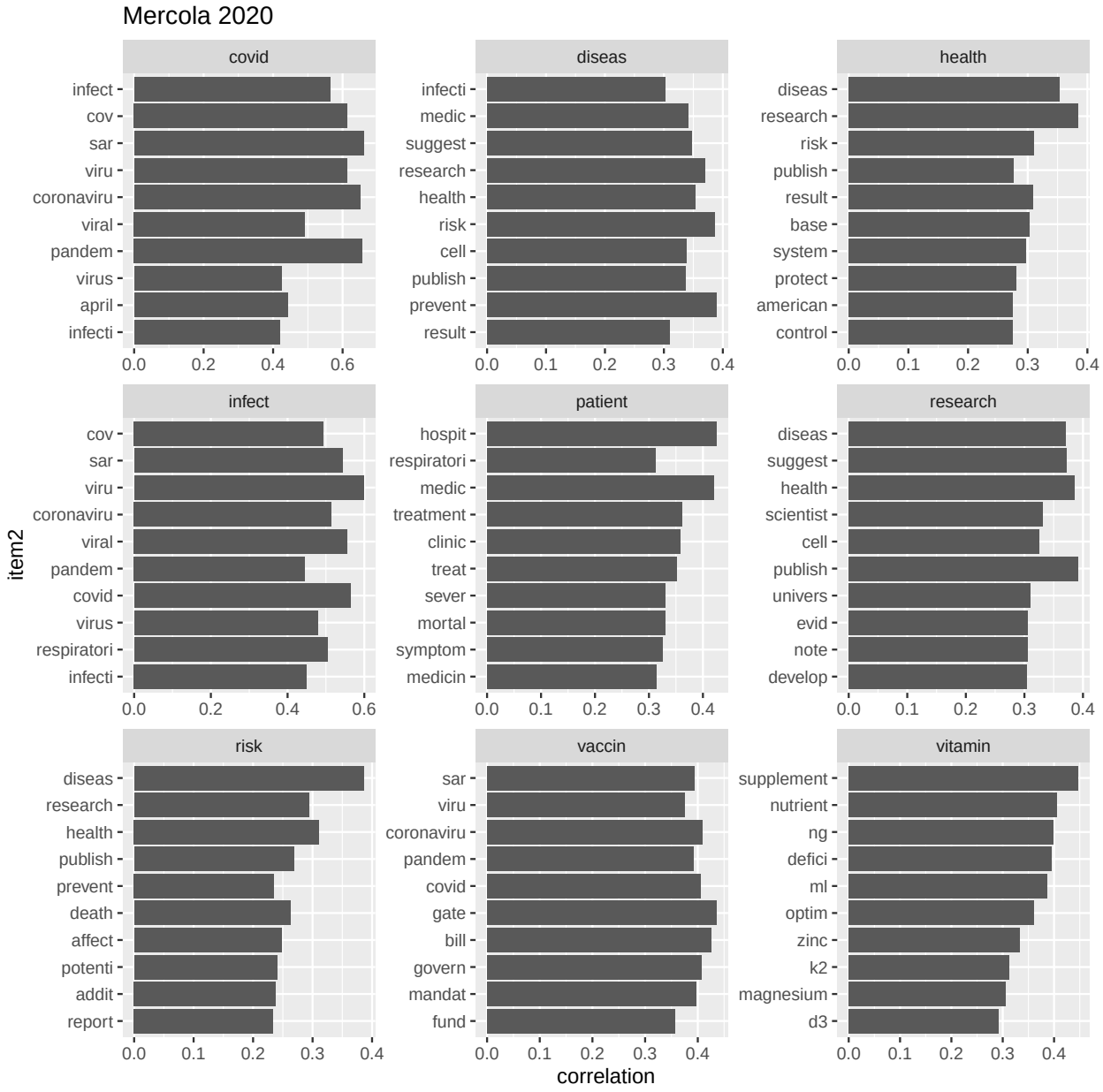


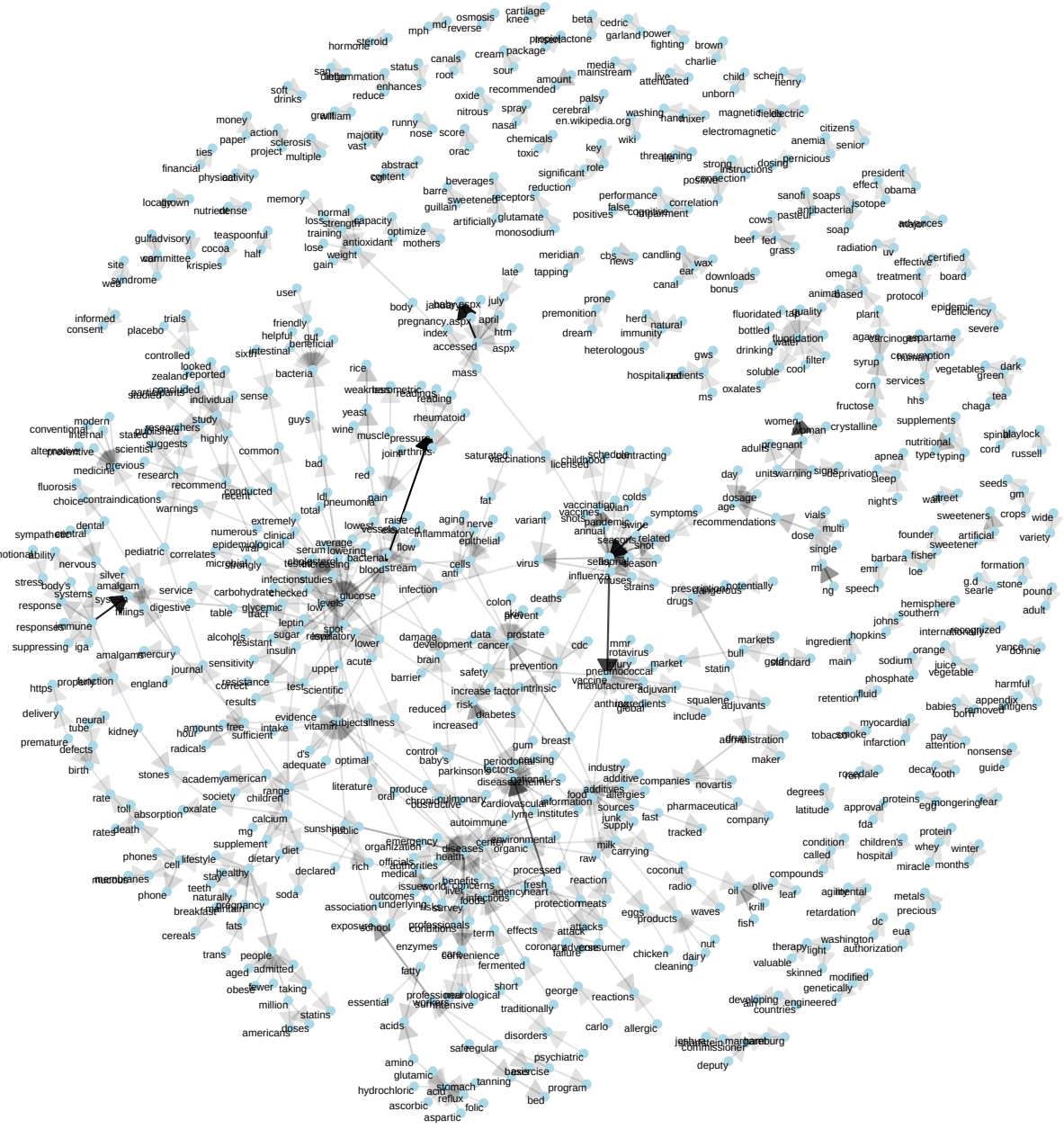
Figure 5: word correlations for top 9 words in 2020 Mercola articles

0.9 Visualization: Top Bigram Relationships

0.9.1 Bigram Relationships for 2009 Merola Articles

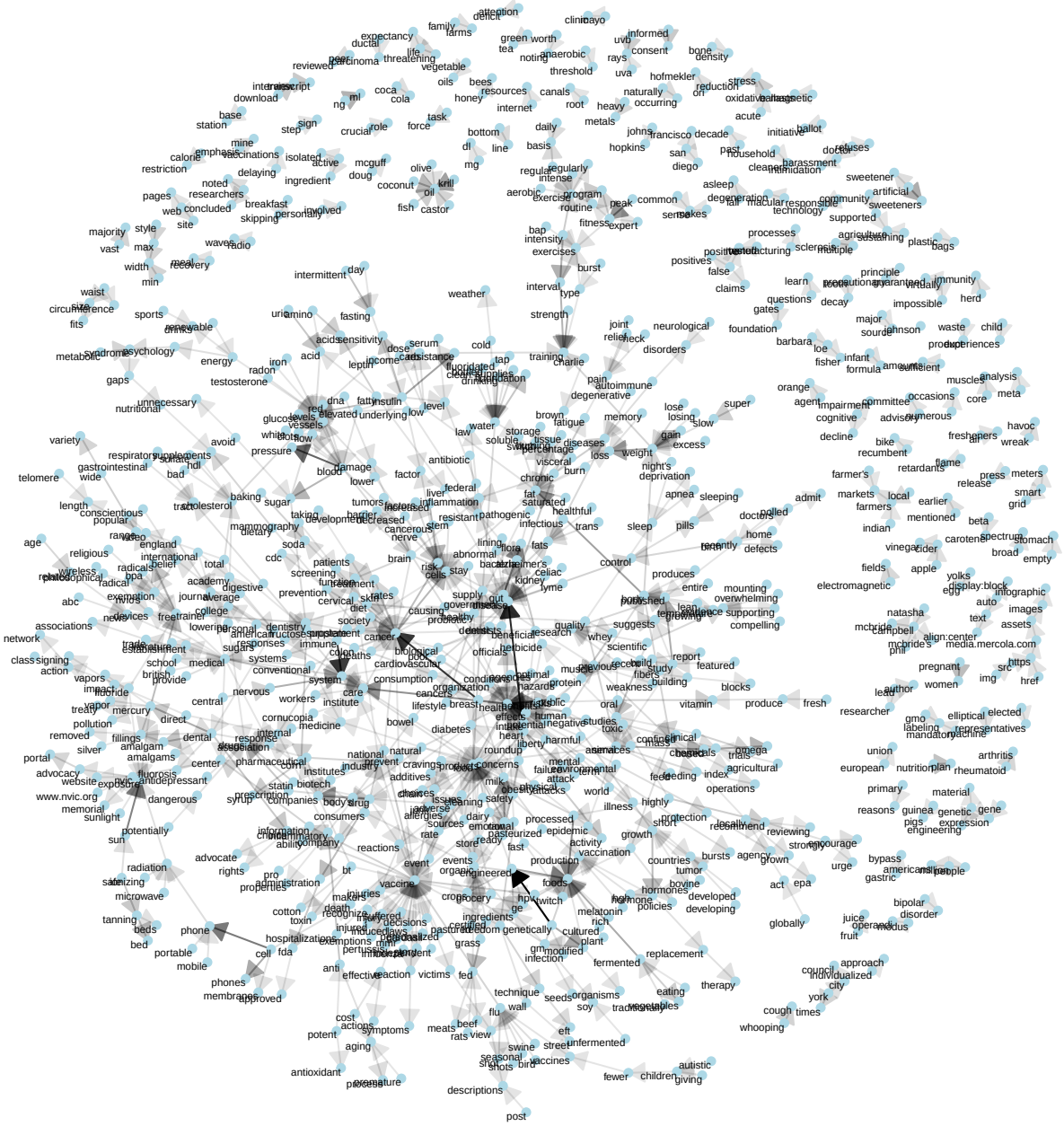
The bigram relationships for Mercola's 2009 articles closely mirror the observations made regarding the most frequent words and correlated words.

Mercola Bigrams 2009



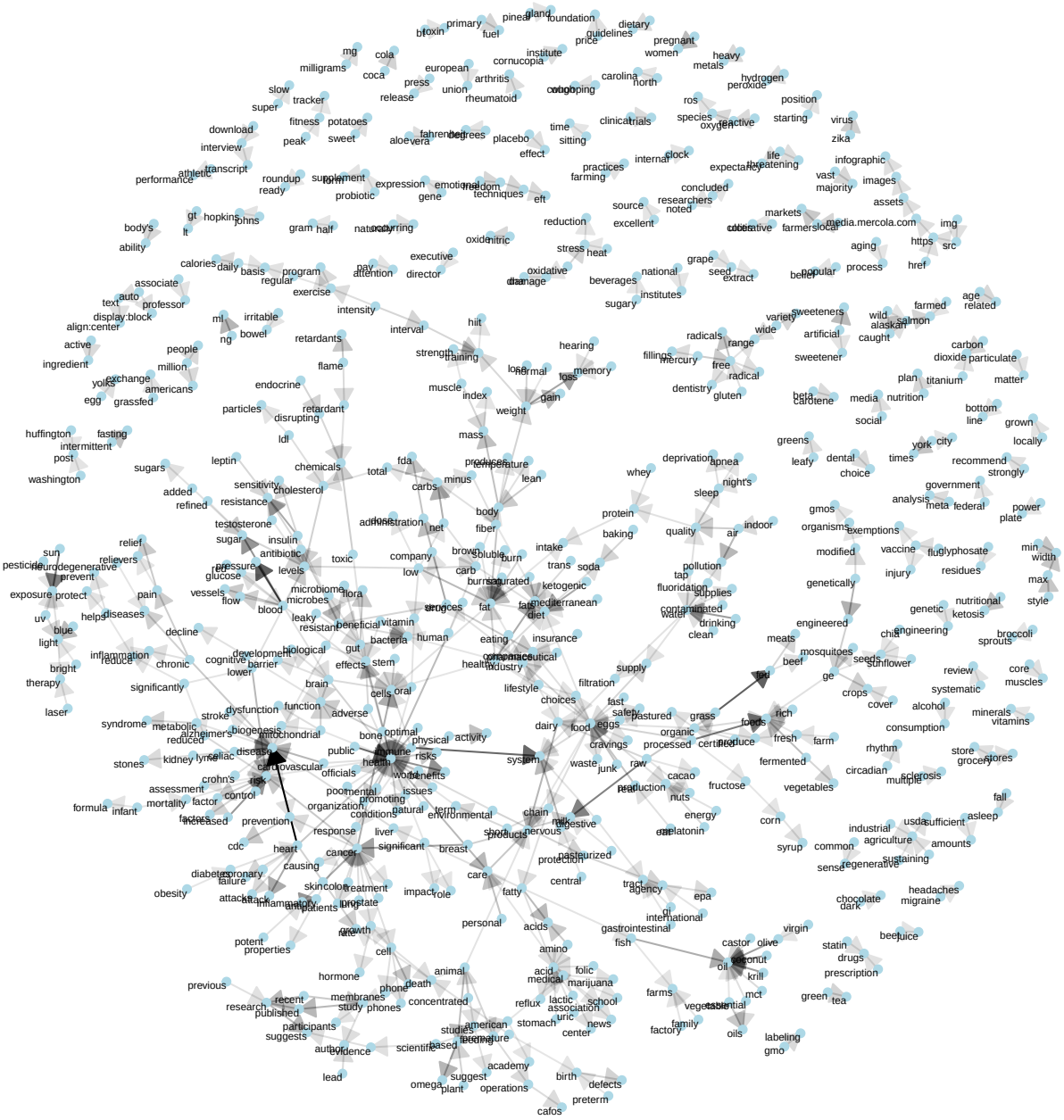
0.9.2 Bigram Relationships for 2012 Mercola Articles

Mercola Bigrams 2012



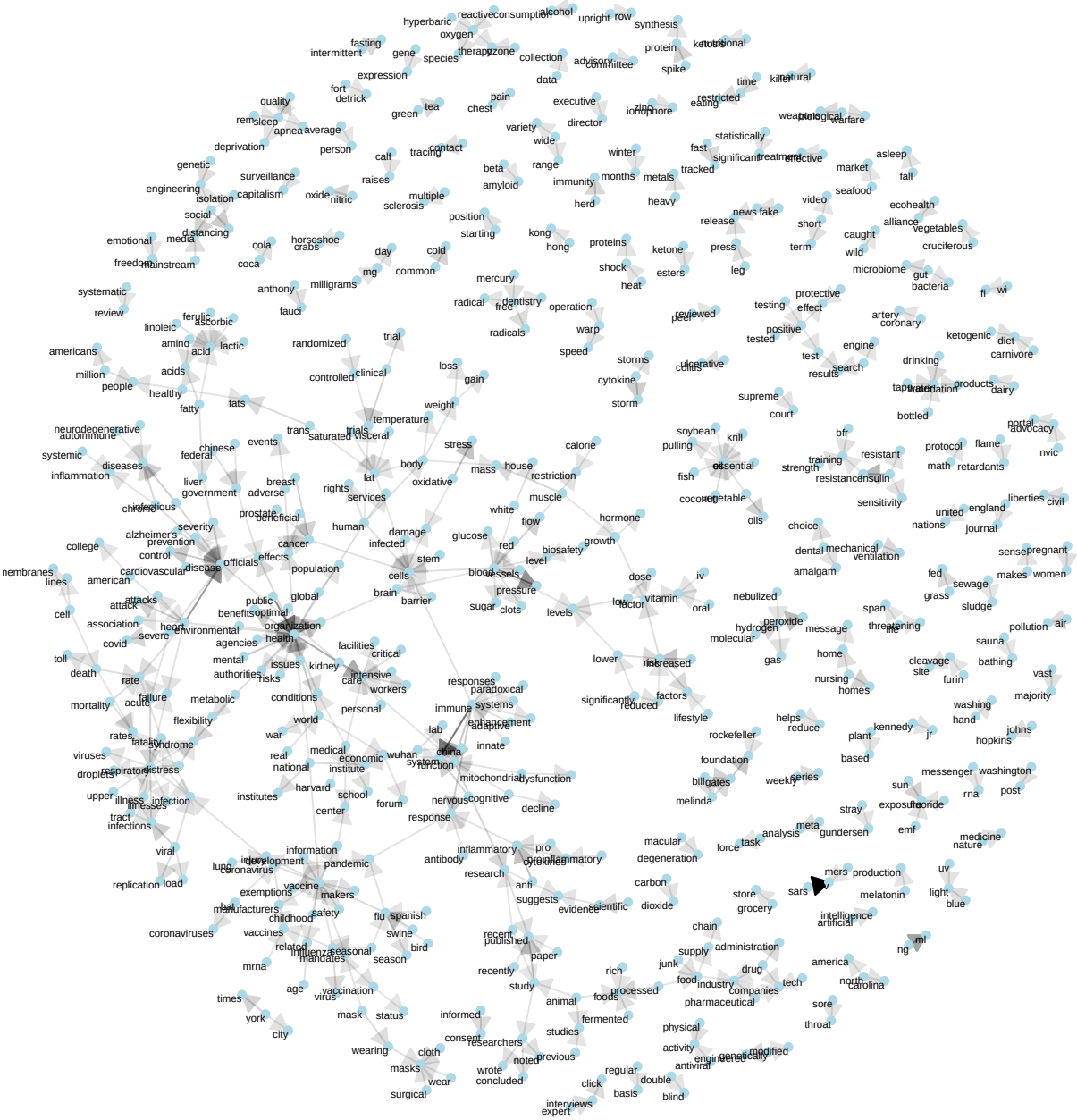
0.9.3 Bigram Relationships for 2016 Mercola Articles

Mercola Bigrams 2016



0.9.4 Bigram Relationships for 2016 Mercola Articles

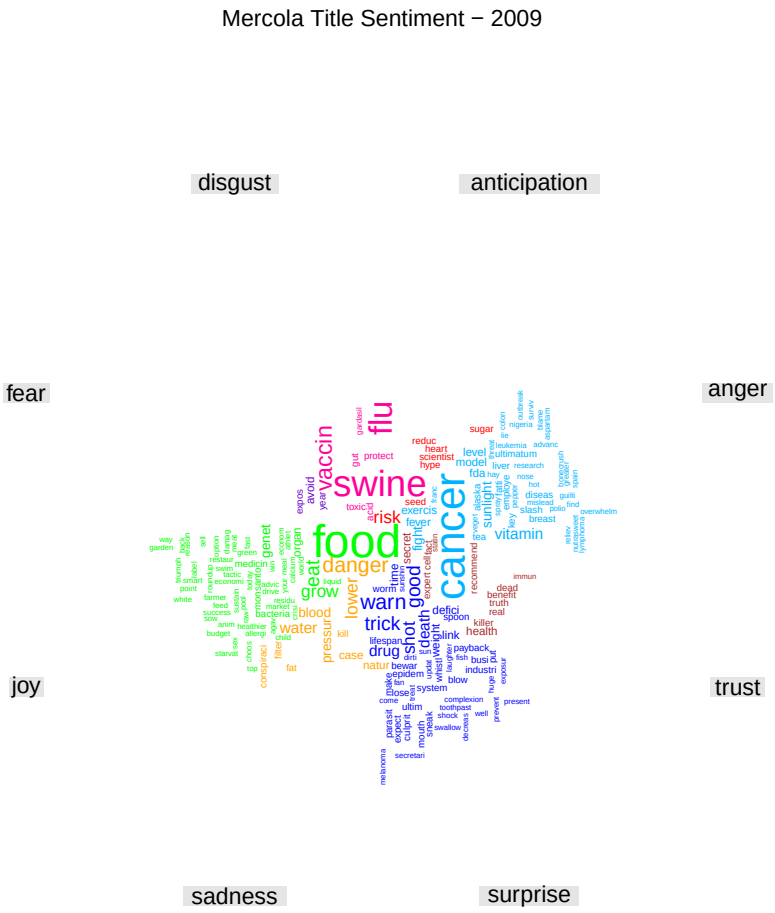
Mercola Bigrams 2020



0.10 Visualization: Word Cloud - Title Sentiment Analysis

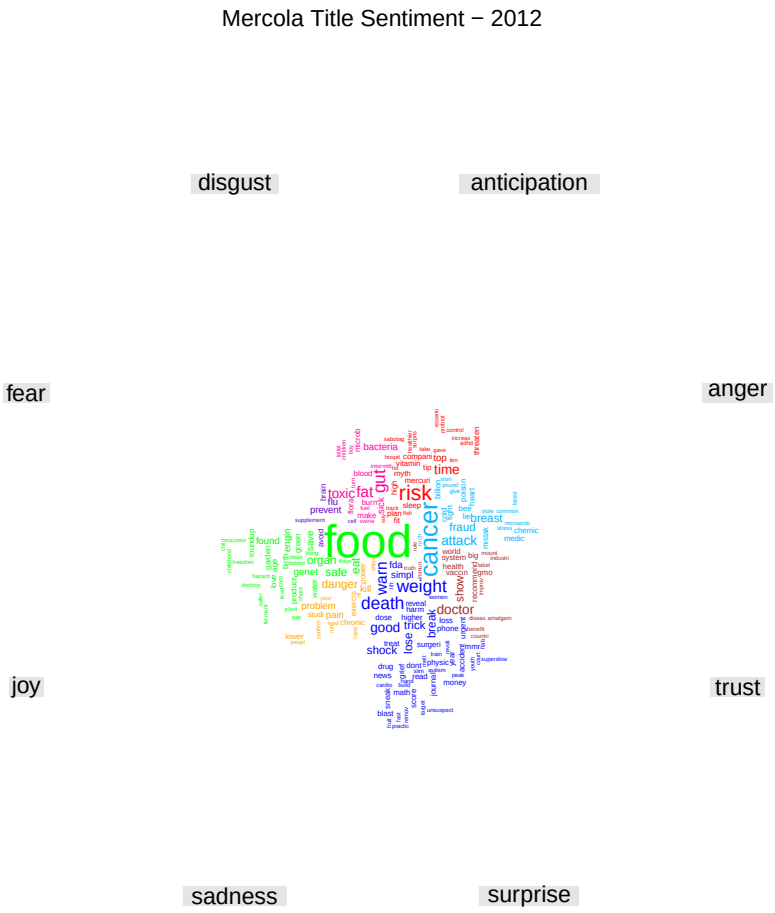
Another way to understand the content of the Mercola corpus is to investigate the titles of articles and their conveyed sentiment. This is of particular interest because these titles are what are displayed on social media posts and would be the attention grabbing material in daily emails and SMS texts.

0.10.1 Title Sentiment Analysis Word Cloud for 2009 Mercola Articles



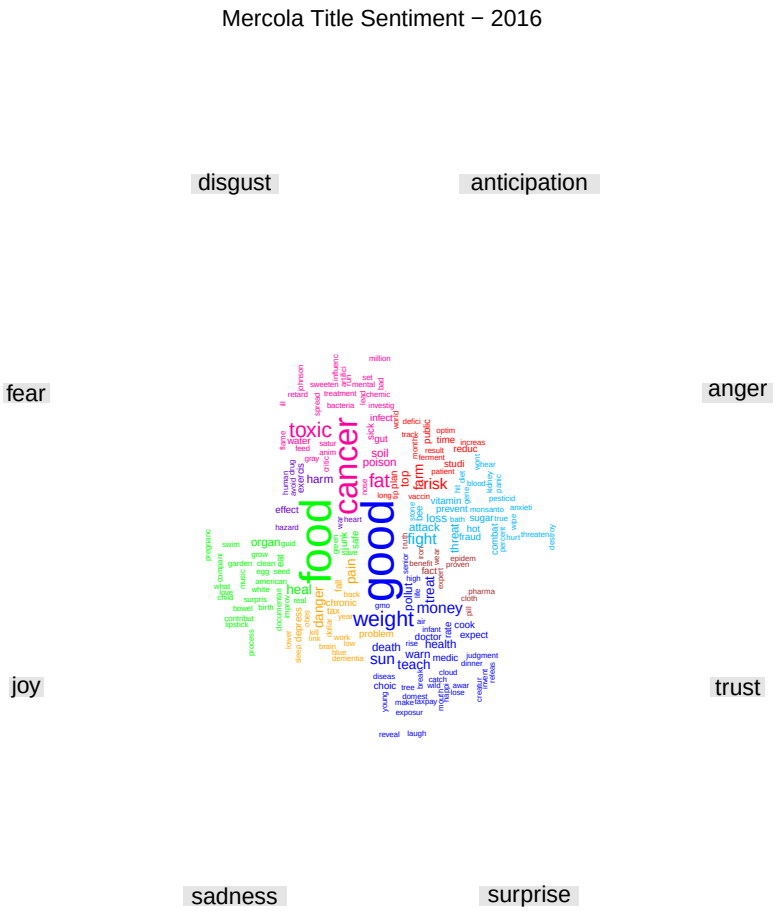

```
## NULL
```

0.10.2 Title Sentiment Analysis Word Cloud for 2012 Mercola Articles



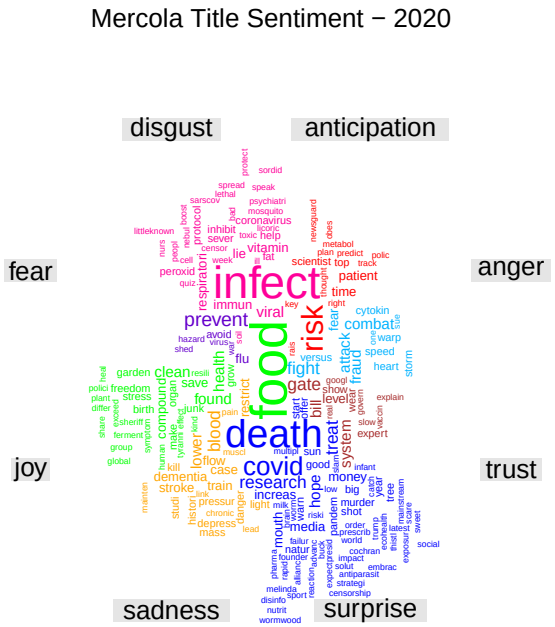

```
## NULL
```

0.10.3 Title Sentiment Analysis Word Cloud for 2016 Mercola Articles




```
## NULL
```

0.10.4 Title Sentiment Analysis Word Cloud for 2020 Mercola Articles




```
## NULL
```

0.11 Social Media Material - Facebook visualizations here

Why these 5? What did the comparison help understand about problem/issue? Understand about domains/locations/genres?

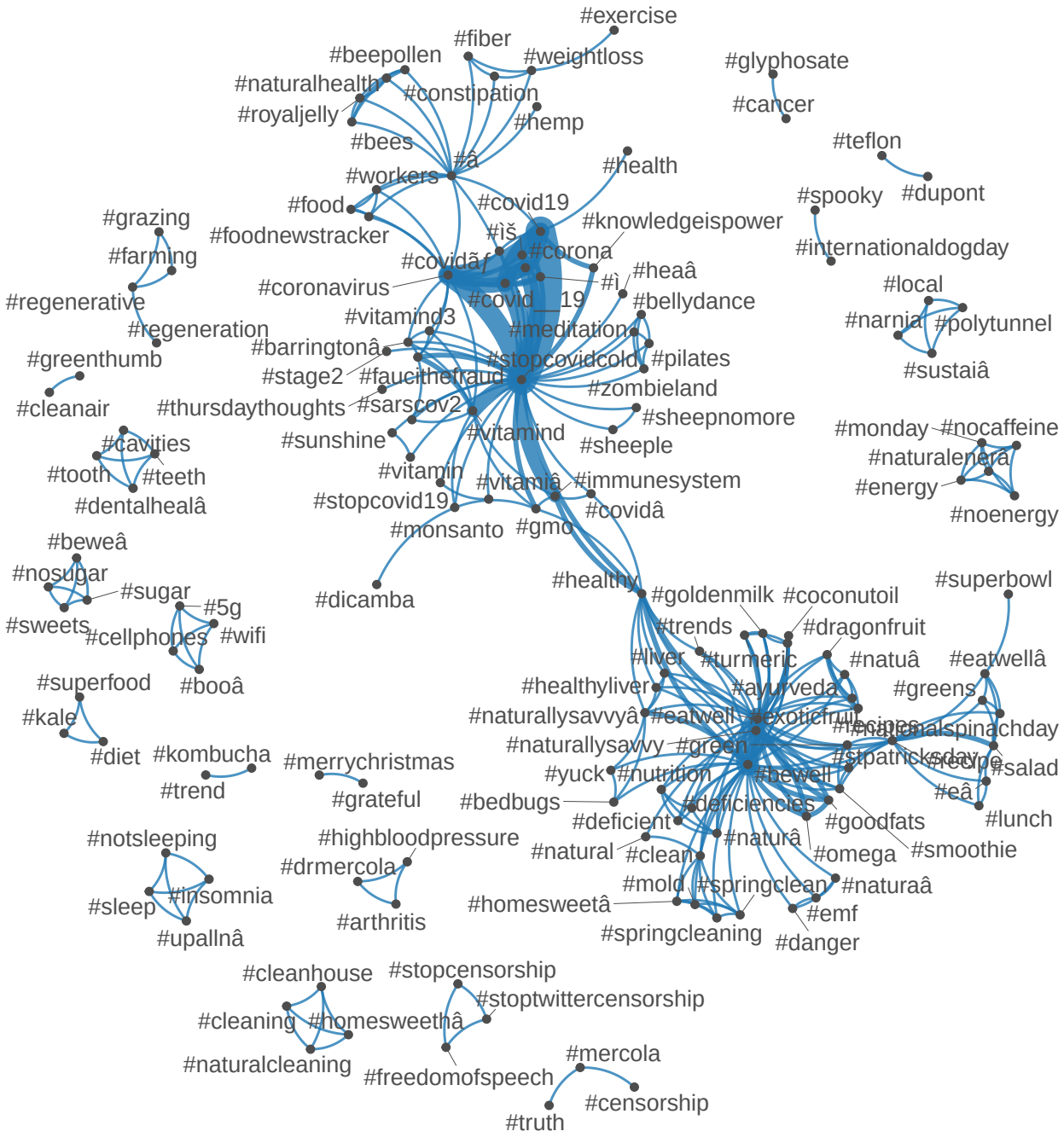
Sentiment of comments versus topics of articles Page views compared to topics or authors Activity within certain time periods compared to topics/authors/sentimeents Compare genres/site categories

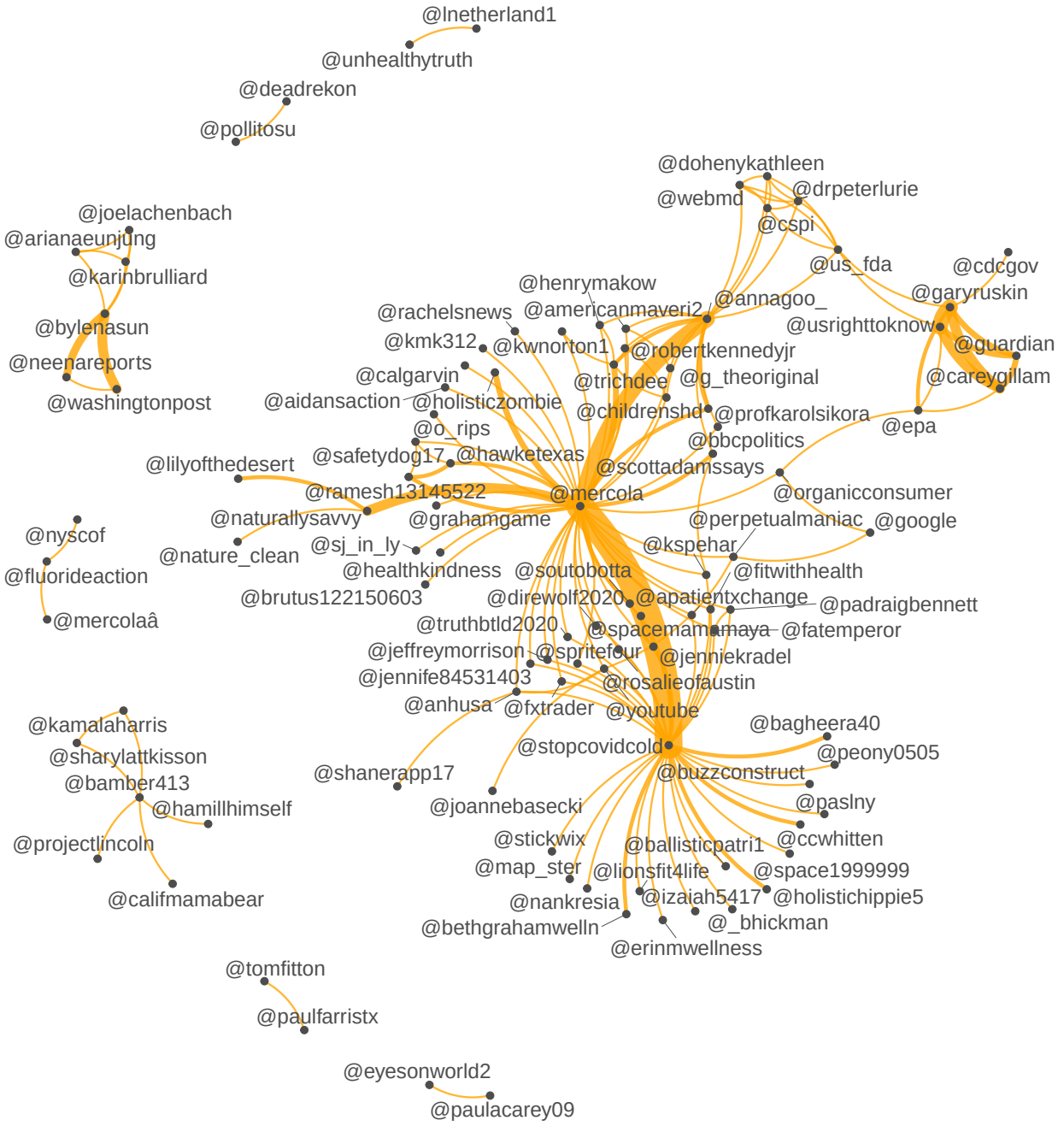
```
## # A tibble: 6 x 30
##   ï..Page.Name User.Name Facebook.Id Likes.at.Posting Created Type  Likes
##   <chr>        <chr>           <dbl> <chr>            <chr>   <chr> <int>
## 1 Dr. Joseph ~ doctor.h~     1.14e11 1794229          2020-1~ Link   1218
## 2 Dr. Joseph ~ doctor.h~     1.14e11 1820648          2019-0~ Photo   566
## 3 Dr. Joseph ~ doctor.h~     1.14e11 1820880          2019-0~ Link   1353
## 4 Dr. Joseph ~ doctor.h~     1.14e11 1820981          2019-0~ Link    412
## 5 Dr. Joseph ~ doctor.h~     1.14e11 1820981          2019-0~ Link    319
## 6 Dr. Joseph ~ doctor.h~     1.14e11 1820981          2019-0~ Link   1329
## # ... with 23 more variables: Comments <int>, Shares <int>, Love <int>,
## #   Wow <int>, Haha <int>, Sad <int>, Angry <int>, Care <int>,
```

```
## #   Video.Share.Status <chr>, Post.Views <int>, Total.Views <int>,
## #   Total.Views.For.All.Crossposts <int>, Video.Length <chr>, URL <chr>,
## #   Message <chr>, Link <chr>, Final.Link <chr>, Image.Text <chr>,
## #   Link.Text <chr>, Description <chr>, Sponsor.Id <dbl>, Sponsor.Name <chr>,
## #   Overperforming.Score <chr>
```

```
#head(Mercola_FB)
```

0.12 Twitter Data





0.13 Limitations/Next Steps

Vast number of sites with ongoing publishing. Some sites are much larger and more widely varied. Questions about validity of comparisons in this case. 2 sites dominate most of the infomration but how much do smaller websites contribute to misinformation. Intersection between political sites and other issues and this one. What role do these sites contribute to the corpus of misinformation? Expanding criterion for selecting sites would help understanding the issue.

0.14 Conclusion

Concluding Statement What were the big picture takeaways from your project? What have you learned?

0.15 References

References/libraries/Endnotes