Killer Robots, the Climate Scam, and Poisonous Vaccines

Understanding Scientific (Mis)Information on Twitter

Fiona Calcagnini, COL '23 ; Sabrina Kwenda, SFS '23; Rich Pihlstrom, COL '24

Georgetown University, Massive Data Institute (MDI), Center for Security and Emerging Technology (CSET)

Introduction

The objective of this research project is to understand how peer-reviewed, published, scientific articles are used (or misused) as evidence as part of misinformation conversation on social media. In our research, we consider three different fields of scientific information: artificial intelligence, climate change, and vaccines.

We began by developing an ethnography of scientific misinformation by computationally developing a list of topics that capture sub-categories within each respective field.

Then we determined the popularity of scientific articles using Altmetrics and collected tweets mentioning each article. Using these collected data, we determined the level of spread associated with these misinformation clusters.

Next, we refined the network analysis to better understand whether the spread of misinformation was propagated by particularly influential spreaders/tweeters.

Finally, we evaluated whether or not the spread dynamic of misinformation conversation could be modeled using a simple epidemiological model, where we model poor quality information spread as disease spread.

Ethnography

We used the "Digital Science Dimensions" dataset, which contains the DOI, title, abstract, year, and country of origin of published papers, to computationally develop topic-dictionaries of keywords for each scientific field using a semi-supervised topic-noise model. The dictionary distributions illustrate the popularity of different keywords across variables like year, country, and topic choice.

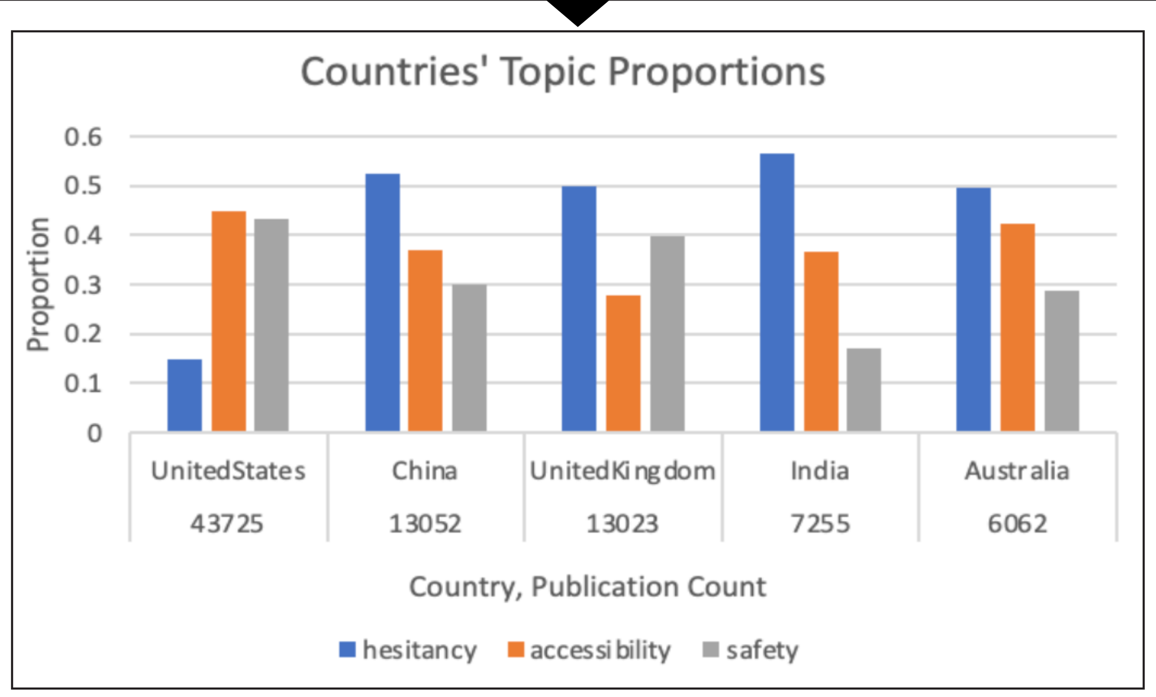

Figure 1. A nested bar chart illustrating the proportion of vaccine papers from five countries with at least one keyword from the hesitancy, accessibility, safety vaccine topics. This graph reveals how different countries focus on different aspects of vaccines. Countries like China and India focus on "hesitancy" while the US emphasizes "safety."

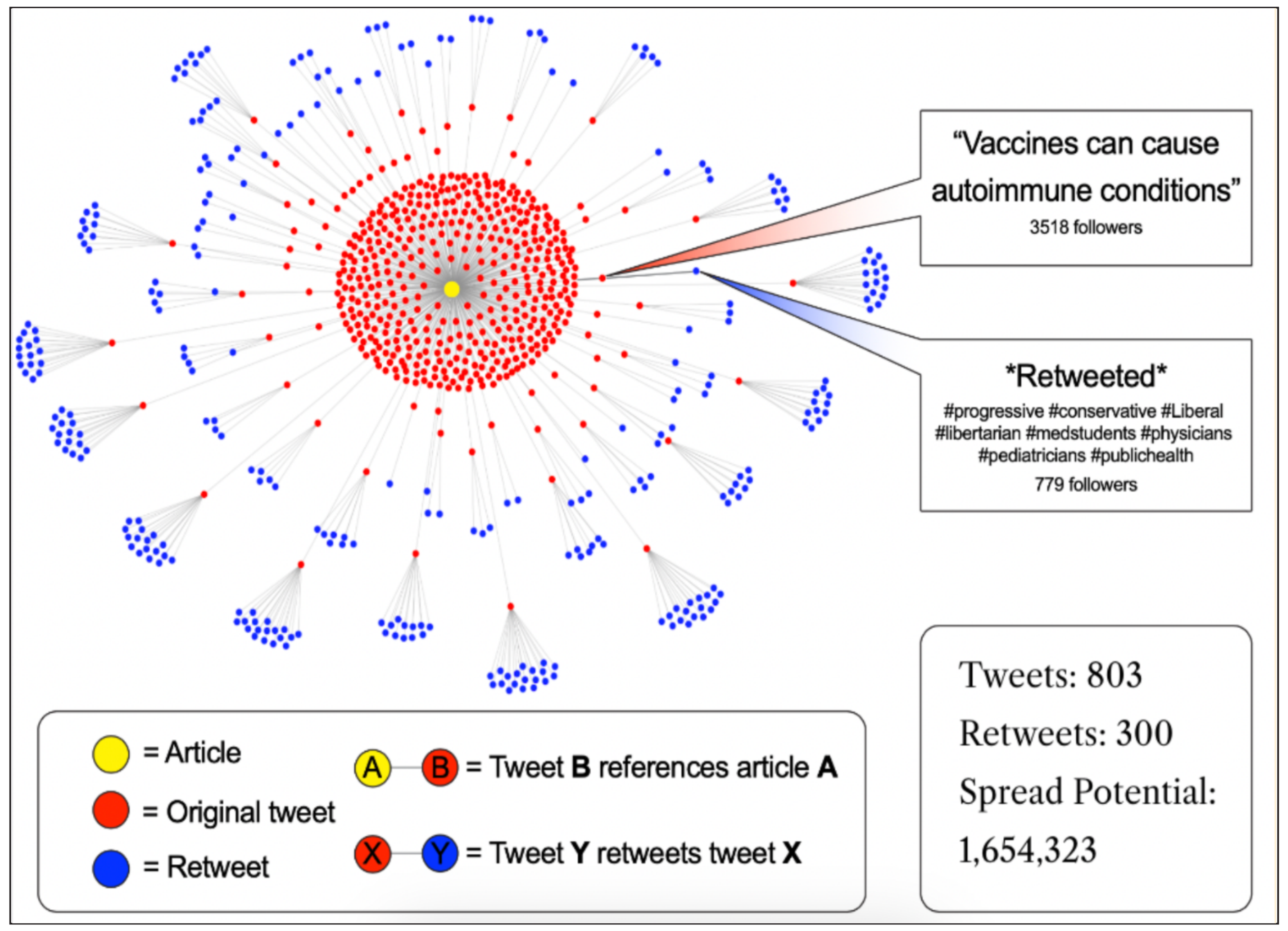

Figure 3. A network graph on an article titled "Vaccination and autoimmune diseases: is prevention of adverse health effects on the horizon?" The central yellow node represents the article, the red nodes represent original tweets that reference the article, and blue nodes represent retweets of the red node to which they are attached. We counted the number of followers for the tweets and retweets. These values reveal how quickly information can spread from just ~1100 tweets to over 1.6 million users.

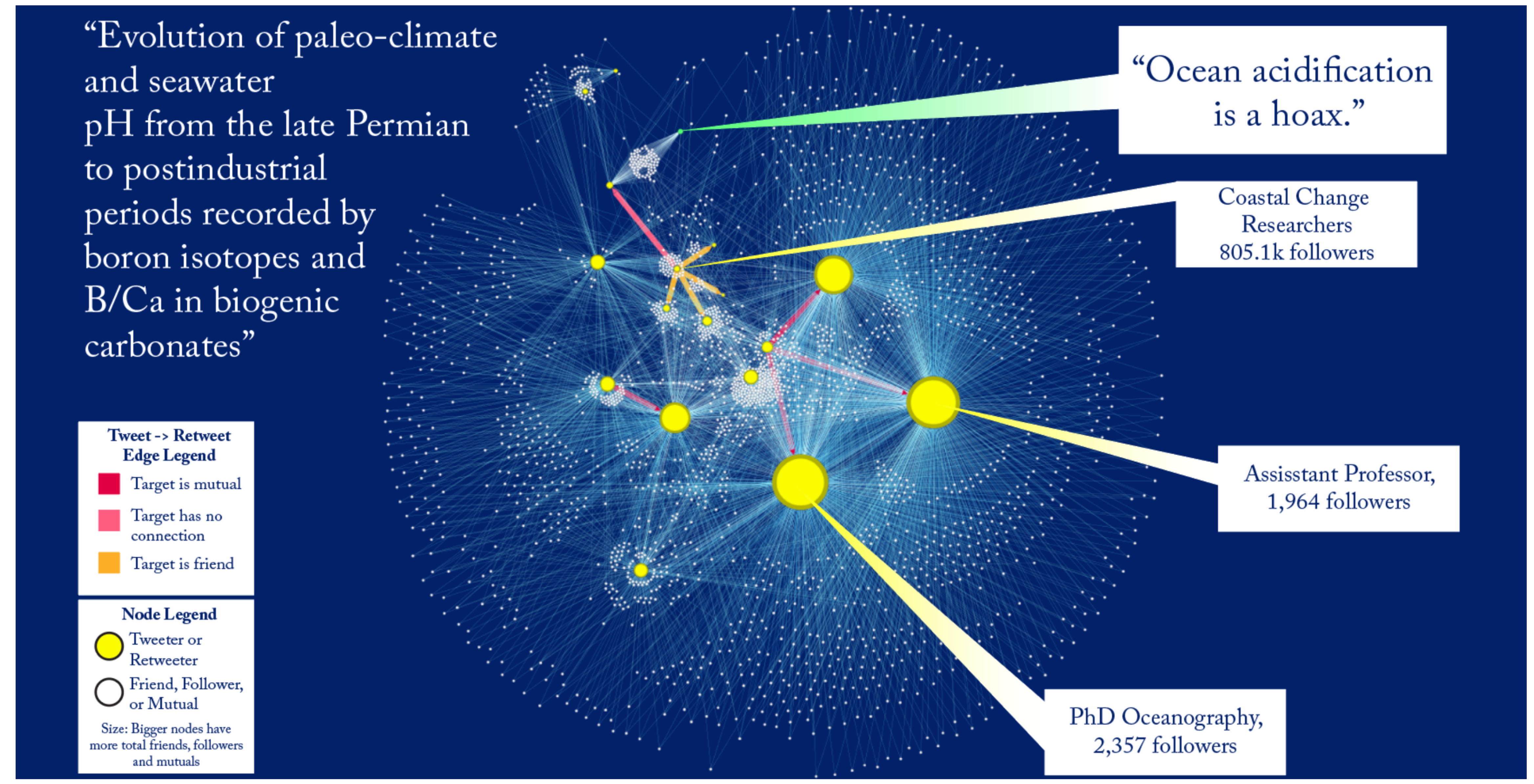

Figure 4. A network graph on a climate change article that shows the relationship between users posting and viewing related tweets. The yellow nodes represent users who posted (tweet or retweet) about the original article. The smaller white nodes represent the friends and followers of each poster. A colored edge between two yellow nodes depicts the relationship between a tweeter and a retweeter. This graph reveals that there are clearly influential nodes, in this case predominantly academics, and also that there is notable community overlap for different posts about the same article.

Popularity

We used the "Altmetrics" dataset, which contains ids for tweets associated with published papers, alongside Python's Tweepy Twitter library to collect information about original tweets related to articles in each scientific field. With this information we examined how tweeters posting about certain articles interact. For each domain, we identified scientific articles containing misinformation:

- Vaccines: articles that have been retracted were labeled as discussing misinformation.
- For climate, articles that were used as evidence online to support misinformed claims were marked as discussing misinformation.
- For AI, articles that rely on the misuse of models or methodology were flagged as discussing misinformation.

By categorizing articles as discussing misinformation versus not misinformation, we were then able to examine the distribution of traits like Altmetrics scores and tweet counts within each classification.

Misinformation Spread

From our gathered information about tweets related to scientific misinformation, we constructed network graphs in order to visualize the connections between both posts and users surrounding a certain article.

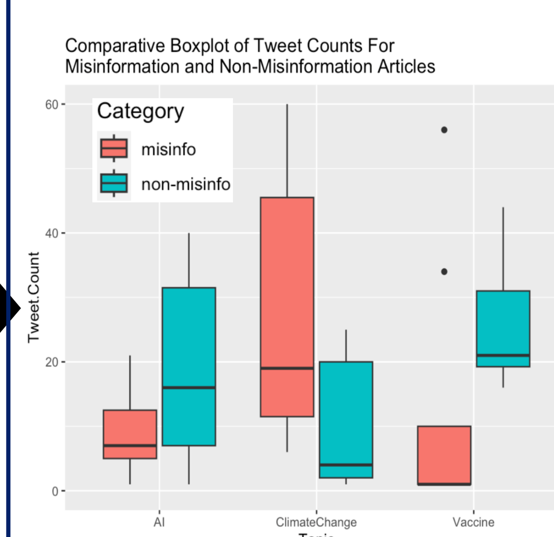

Figure 2. A comparative boxplot of the tweet counts for articles labeled as misinformation discussion and non-misinformation discussion for the different domains. This graph suggests that generally, the popularity of scientific misinformation does not surpass that of non-misinformation. The differences across domains may also be due to the different labeling methods used for each domain.

Epidemiological Model

In the last phase of our research, we considered an epidemiological model to see how well it could model the rate of spread of misinformation online. We looked at an SI (Susceptible-Infected) model where original tweeters are the initially-infected population and anybody who viewed the tweets or retweets are the final-infected population. We sought to narrow our infection rate to produce the same final-infected population that we saw in our Twitter data.

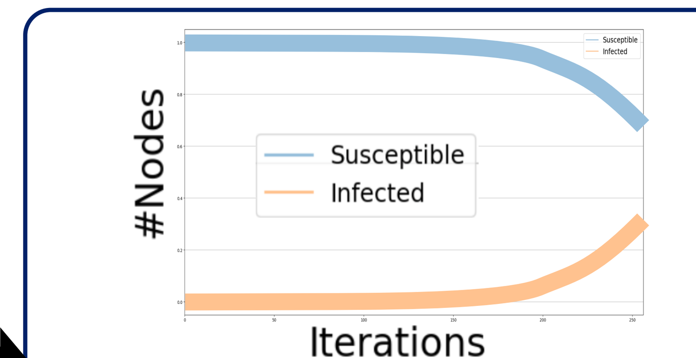

Figure 5. A line graph of susceptible and infected populations for a 24,196-user network, starting from a single infected node. The infection rate producing similar results to observed data is 0.001523. From an epidemiological perspective, the spread rate is higher than the rate using an SIR model for approximating flu spread, but we note that had we used the entire Twitter population for our model, it would be much lower.

Conclusion

Scientific articles are being used to both spread and refute misinformation in different fields. It is complicated to determine its influence. We used a combination of topic modeling and social media popularity to understand how quickly poor quality (or refuting of poor quality) information can spread. We find that the popularity of different misinformation themes is very dependent on the discipline and topic. Finally, spread on social media can be significant, but in general, most shared information is not spread heavily. The SI model was a good start to larger scale modeling of misinformation, but more sophisticated models should be explored in the future.

Advisers: Lisa Singh, Autumn Toney, Ken Kawintiranon

Thanks to our mentors, the Fritz Fellowship Program, CSET and MDI for their support!