Helen Chung
Hasan Liou
Cindy Huynh
DSC 180B WI21

Analyzing the the Diffusions of Various Forms of Misinformation on Reddit

In September 2020, Gallup surveyed over a thousand US adults regarding social media and the spread of misinformation online. They had found that nearly three-quarters (74%) were very concerned about the spread of misinformation online, which seems to be relatively stable across party lines. They also estimated that 61% of news content on social media to contain at least some misinformation.[1] With the increased discourse that is readily taking place on social media, misinformation has started to grow, and sometimes even thrive within these platforms. Whether it is as simple as the sighting of an alien or something more complicated like the accusation of a stolen election, the effects of misinformation online can be destructive. We are just beginning to see examples of these kinds of consequences and their lasting effects.

We hope to look at the spread of misinformation online, specifically on Reddit. We hope to compare two different kinds of misinformation: one about political information as well as about urban myth legends, while comparing that with how scientific information may spread on this platform. We hope to analyze the diffusion of these kinds of content on Reddit, looking into their respective spread patterns.

Data Generation and EDA

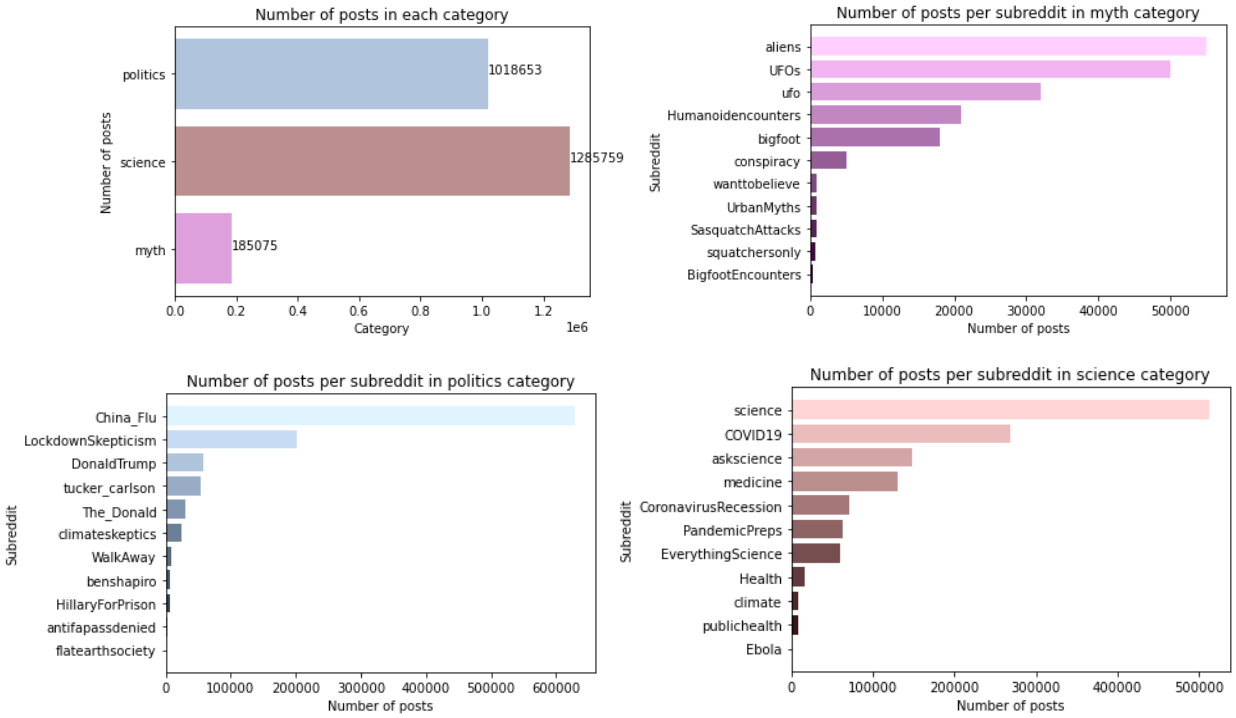
To obtain data, we used Pushshift Reddit API. Initially, we had plans to use PRAW, the API provided by Reddit themselves; however, we found that it did not have all the features we were looking for. We decided on the Pushshift Reddit API because it allowed us to download data for posts and comments that were either deleted or archived in addition to the posts that are currently live on Reddit. This allowed us to get more data and create less biased data since posts in the misinformation subreddits get deleted more often than in the science subreddits. We also preferred the Pushshift Reddit API due to the fact that it had a higher rate limit.

Using the Pushshift API we gathered data about users and their posts across subreddits relating to scientific, urban myths, and misinformation. We collected data of the username of the poster, the date of the post, and the subreddit of the post. Using this information, we can calculate user polarity and further analyze the existence of echo chambers.

We collected data from 33 subreddits (11 from each category) which is about 2.4 million comments. In the Myth category, we gathered comments from a wide variety of subreddits such

[1] Jones, Jeffrey M. *Most Americans Would Believe Social Media Misinformation Warnings*. 13 Oct. 2020, knightfoundation.org/articles/most-americans-would-believe-social-media-misinformation-warnings/.

as UFOs, conspiracy, BigfootEncounters. For the science category, we collected comments from subreddits such as publichealth, askscience, and COVID19. For the political category, we collected data also from a wide range of topics including flatearthsociety, The_Donald, and LockdownSkepticism.



As you can see in the charts above, there were about about 1 million rows collected for the politics category, about 1.2 million comments collected from the science category, and only about 185 thousand rows from the myth category. The data is heavily skewed towards the politics and science categories but that is because the subreddits from these categories are much more popular compared to the ones from myth.

For each of these 2.4 million rows, we collected data on the author, time, id, and subreddit of each comment. The author is defined by the username of the poster. The time created is the time of when the comment was posted in UTC time. ID is the unique identifier of the comment which can be used to rehydrate the comments at a later time if we so choose to. Finally, we also kept track of the subreddit in which the comment was collected from.

Methodology

Our analysis begins by selecting a set of subreddits that can reasonably be labelled as "factual," "politically misinformative," and "myth." Such subreddits include r/science (factual), r/The_Donald (politically misinformative), and r/Bigfoot (myth). We collect usernames from comments made between March 2020 and June 2020, at the start of the pandemic.

Each user will then be rated with a "user polarity", a set of three metrics representing how often they browse each category of subreddit. Each metric is the percentage of comments a user has made in a certain category of subreddit. For instance, a user who frequents and comments in mostly factual pages may have a polarity of {science: 99%, political_misinformation: 1%, myth: 0%}. User polarity is calculated by the following formula:

$$Myth\ polarity\ =\ \frac{\#\ of\ Posts\ in\ Myth\ Subreddits}{Total\ \#\ of\ Posts}$$

$$Science\ polarity\ =\ \frac{\#\ of\ Posts\ in\ Science\ Subreddits}{Total\ \#\ of\ Posts}$$

$$Political\ polarity\ =\ \frac{\#\ of\ Posts\ in\ Political\ Subreddits}{Total\ \#\ of\ Posts}$$

For each user, all three subreddit polarities will add up to 1, and will tell us how "loyal" a user is to a certain category. The closer the polarity is to 1, the more loyal a user.

To analyze variations in echo chambers, we will be studying how likely users are to participate in and consume either factual, politically misinformative, or mythological information. Observing the tendencies of any one subreddit's user base will shed light on its users' general consumption patterns. Aggregated usernames and their associated subreddits, therefore, will provide the basis for this analysis.

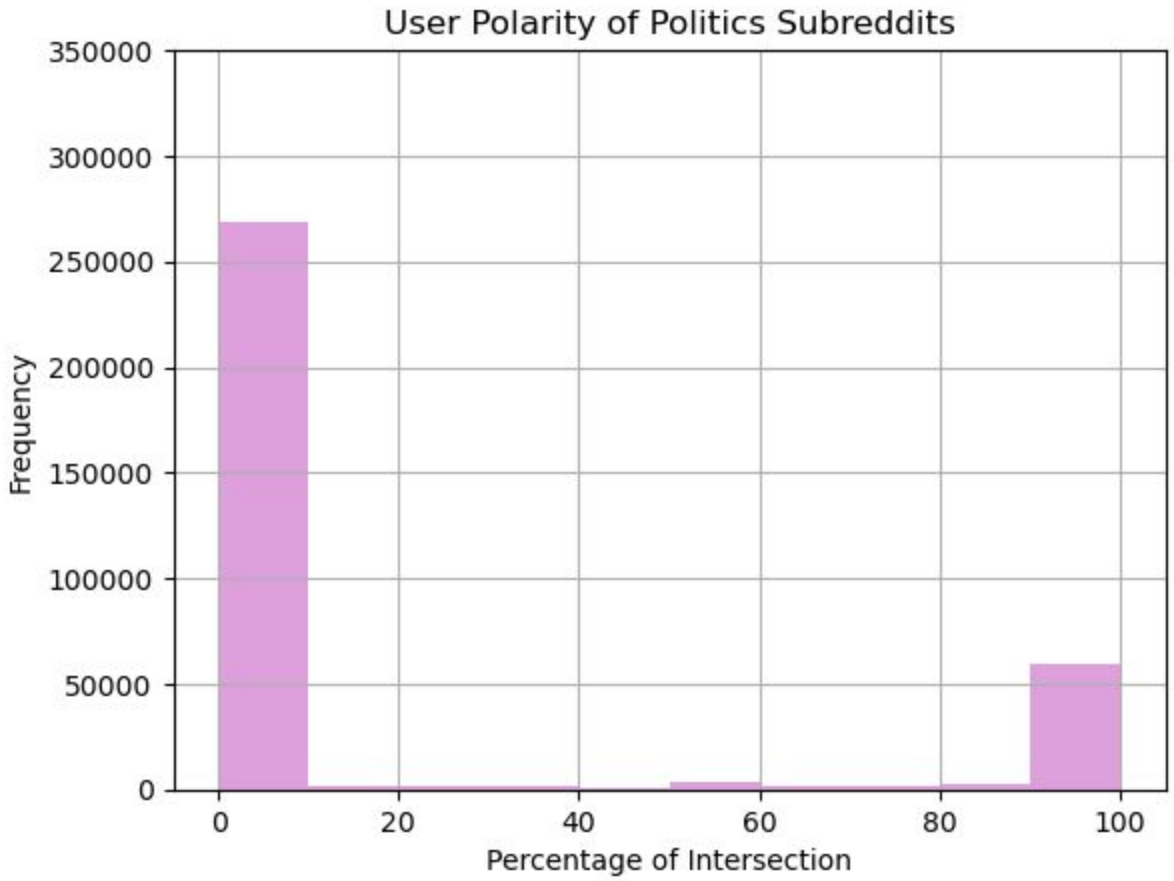


Fig 1. Histogram of the political polarity of users

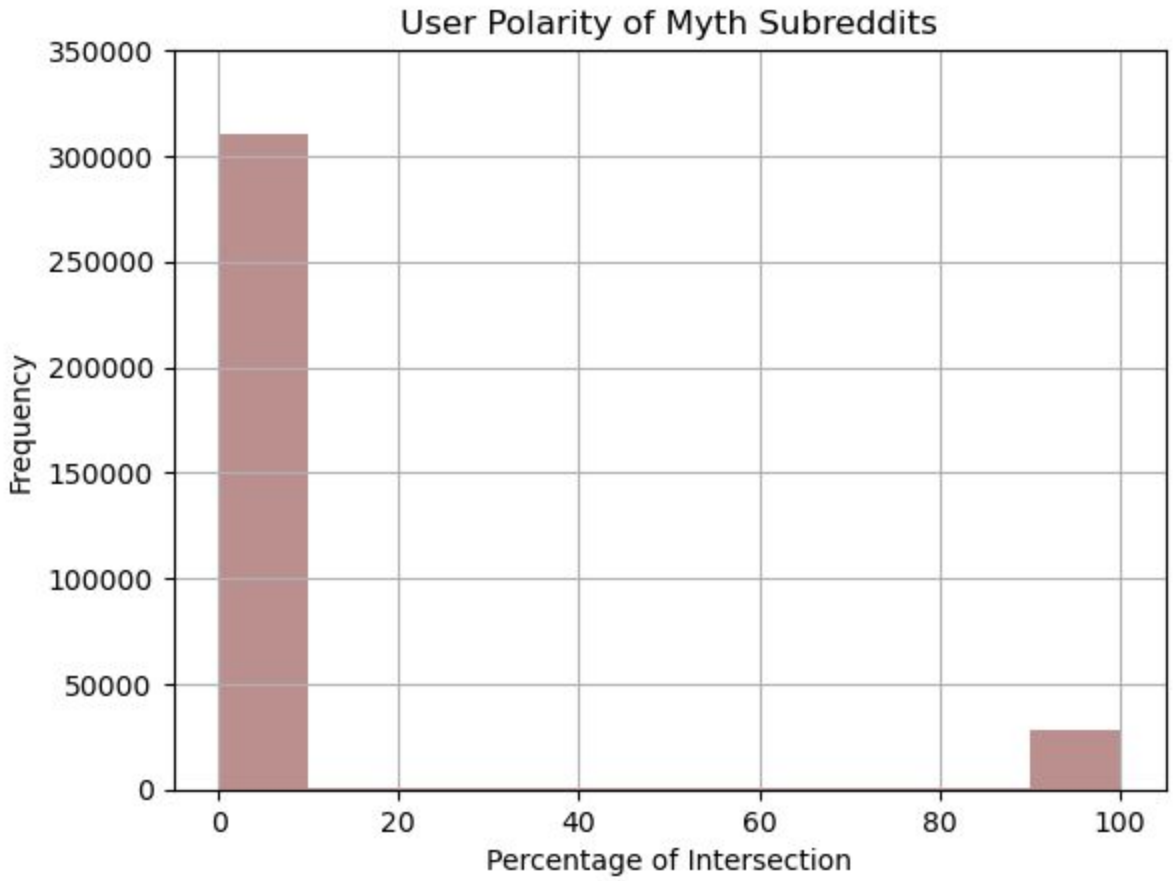


Fig 2. Histogram of the myth polarity of users

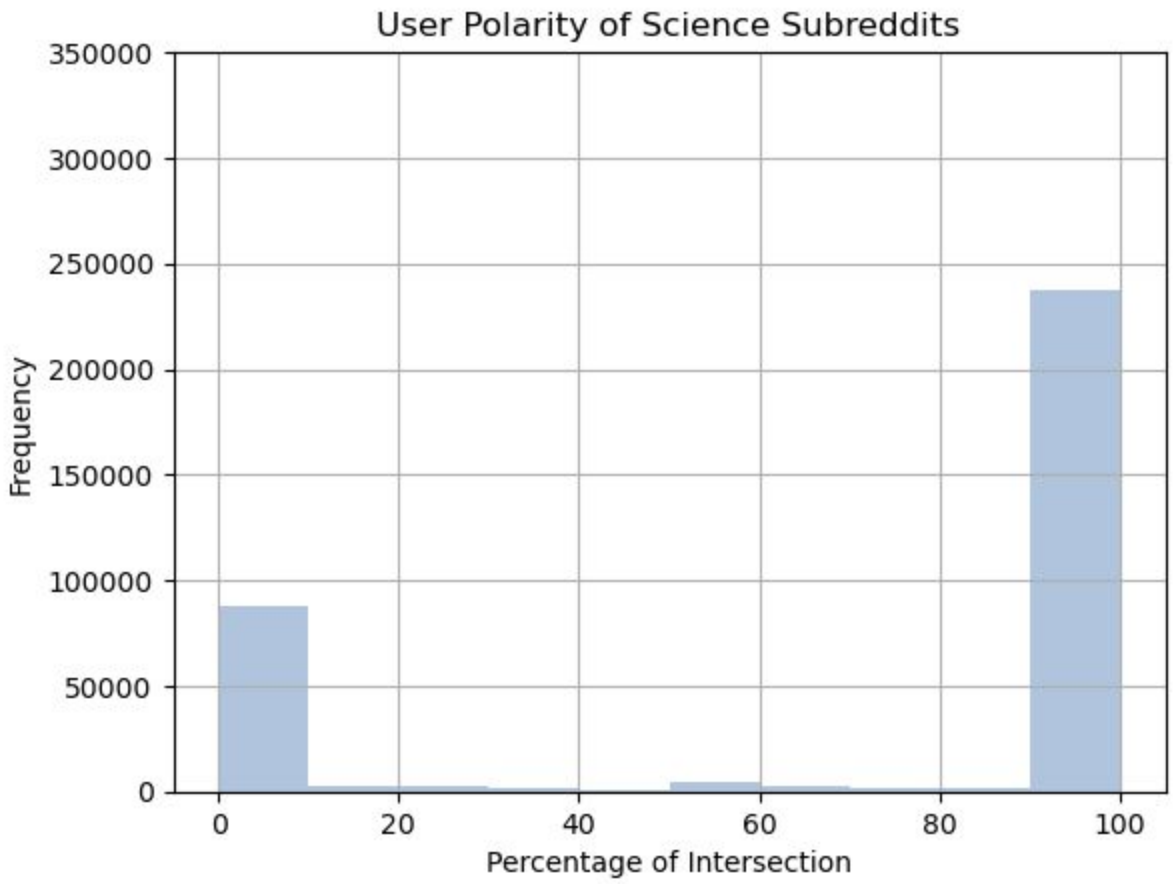


Fig 3. Histogram of the scientific polarity of users

The three graphs above represent the political, myth, and science subreddits respectively. These represent the user polarities of these subreddits, on a range from 0% to 100%, where 0% would mean that a user has not posted any comments within the subreddit, while 100% would mean that all of a user's comments are within one subreddit. Somewhere in the middle of this spectrum, like a 40% percentage of intersection in the myth subreddit, would represent that 40% of a user's past posts were in myth subreddits. The other 60% of their posts can be found distributed between the political and scientific subreddits. Thus, all polarities within each subreddit category will add up to 100%, or all of a user's past posts. The political (*Fig 1*) and myth (*Fig 2)* subreddits are quite similar in terms of their trend- they have a majority 0% intersection among users, which means that many Reddit users have not ever posted in the subreddits before. This wasn't too surprising for the myth subreddits, as it is somewhat a smaller community among Reddit users, since we looked into topics like Sasquatch, Bigfoot, and aliens. However, we had predicted the opposite for the political subreddits. Because we had pulled data from the months of March - June 2020, which happened to be right before the November presidential election, we thought there would be more political discussion happening on this platform. Perhaps we had chosen political subreddits that weren't as mainstream for political

discussion, or maybe political discussion is kept to other social media platforms like Twitter or Facebook.

Meanwhile, the science (*Fig 3*) subreddit had the most versatile user base. There are more users that are very loyal to the science subreddits than myth or political users. There were also more of a variety of users, as you can see in the slight bump in the 60% region. This finding was also somewhat surprising, as we had previously hypothesized there to be more of a normal distribution rather than a bimodal distribution we see here. We had initially thought that because scientific findings shouldn't be too polarizing, there would be a variety of different users that frequent other subreddits other than science subreddits.

Once we have obtained the user polarities, our next step is to visualize the aggregated browsing patterns of the users we have collected to detect an echo chamber. For each possible combination of two subreddits, we will record both the count and average polarities of the subset of users participating in both. By calculating these metrics and visualizing them in a heatmap, we can obtain a general sense of how likely followers of a certain subreddit are apt to follow, comment in, and spread potentially misinformative ideas in another. We hypothesize that users are more likely to follow subreddits aligning with their currently existing beliefs, and, given the social and political tensions of the 2020 presidential election, that political misinformation will spread faster than myth.

To generate the heatmap of shared users, we took the number of shared users between any two subreddits, and divided by the total number, or union, of users between said subreddits. This helped to account for the large user bases within the science subreddits, as opposed to the myth subreddits.

$$Heatmap\ Value\ =\ \frac{Subreddit\ A\ Users \cap Subreddit\ B\ Users}{Subreddit\ A\ Users \cup Subreddit\ B\ Users}$$

Our next two heatmaps rely on the average myth and political misinformation polarities of the common users of any two subreddits. To remove other types of misinformation affecting the visualization, the polarities will be reweighted to just a misinformation category and science. Our values are generated as follows:

$$Myth - Science\ Polarity\ =\ \frac{Average\ Myth\ Polarity}{Average\ Myth\ Polarity + Average\ Science\ Polarity}$$

$$Political - Science\ Polarity\ =\ \frac{Average\ Political\ Misinformation\ Polarity}{Average\ Political\ Misinformation\ Polarity + Average\ Science\ Polarity}$$

Results and Analysis

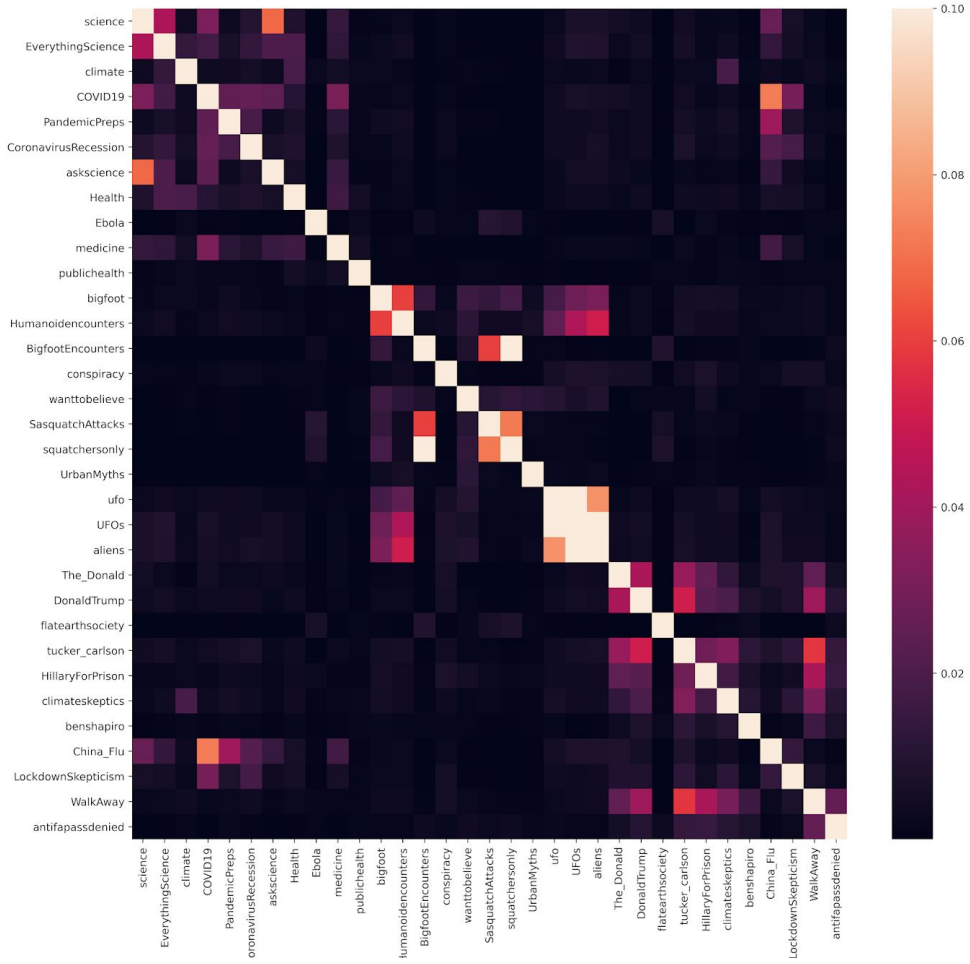


Fig 4. Heatmap of shared users between any two subreddits

The colormap of Figure 4 has been rescaled for visibility purposes, as most subreddit combinations had less than 10% of the same users. Regardless, we are able to make out three red squares along the diagonal, each square confined our established subreddit types. Participants of science subreddits will often visit their similar subreddits, while mostly staying out of our other categories. Our outlier is a result of our scientific COVID-19 related subreddits crossing into the conspiracy COVID-19 related subreddits.

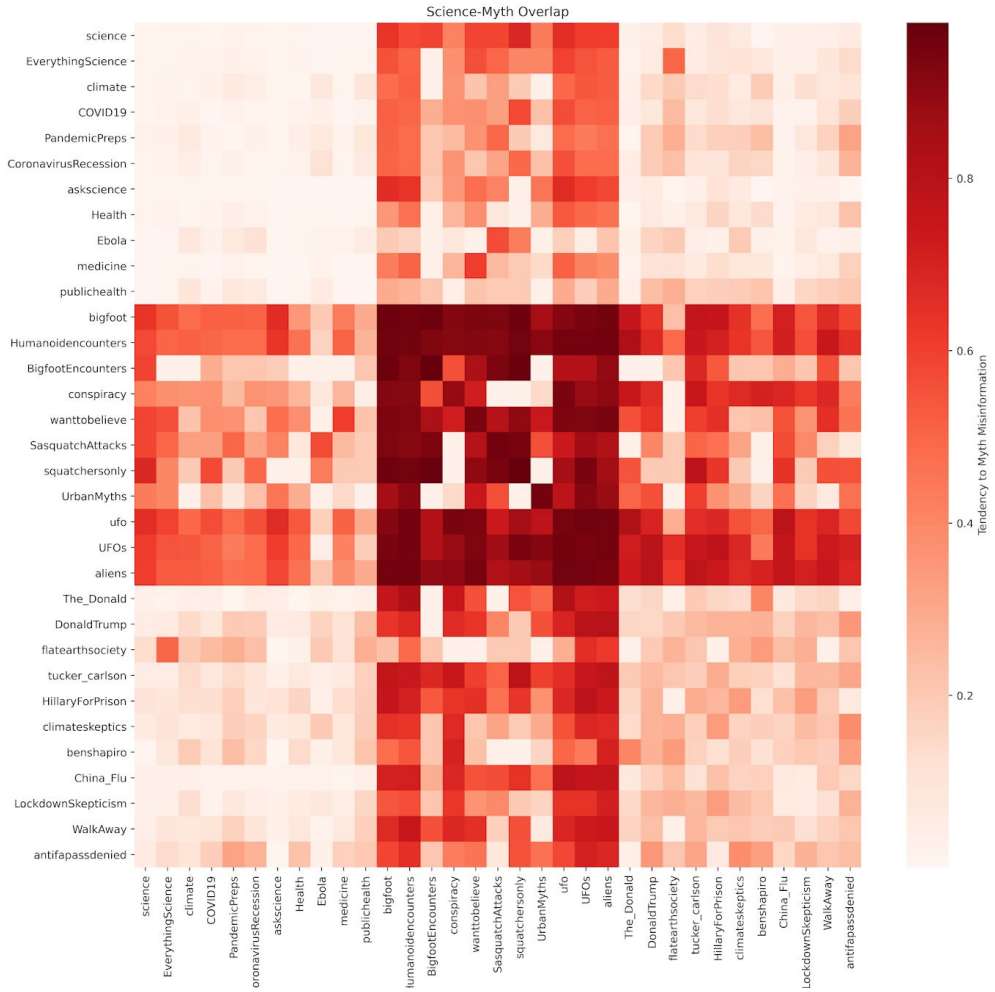


Fig 5. Heatmap of Average Myth Polarity vs. Average Science Polarity

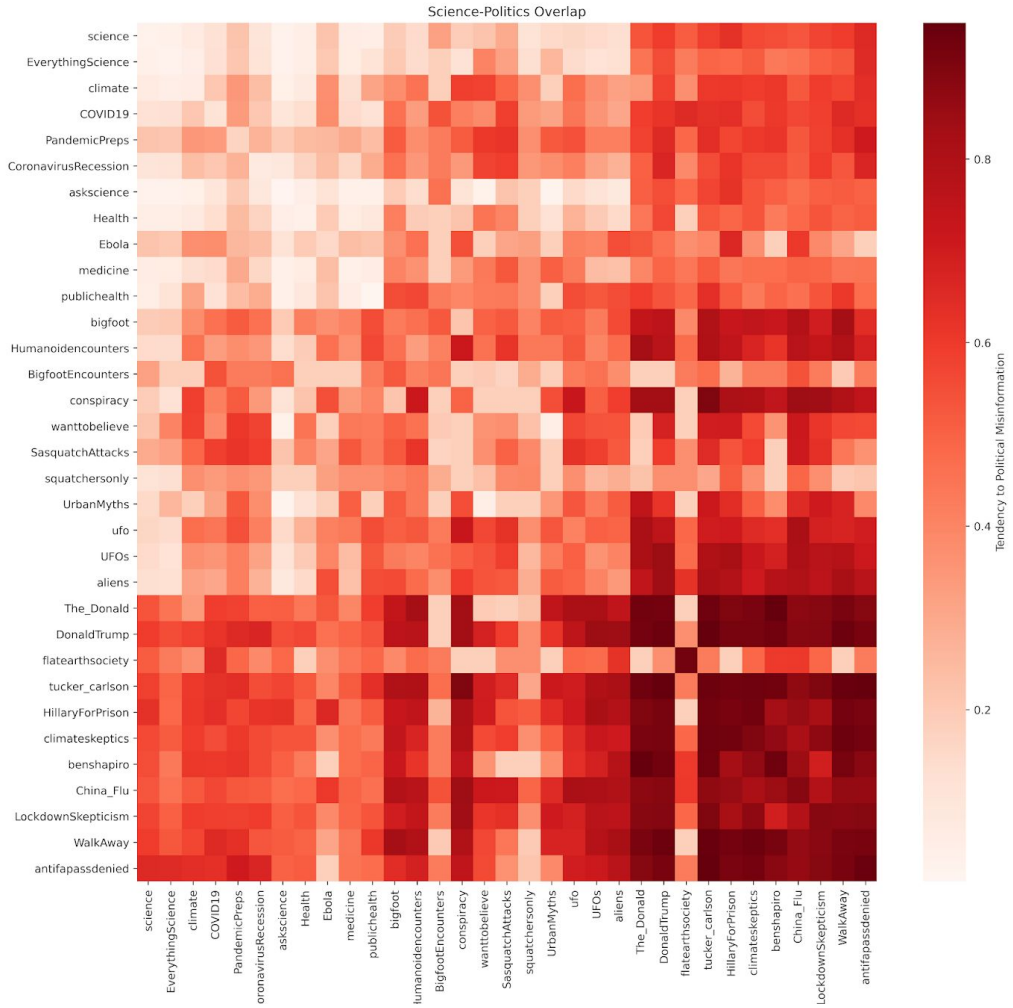


Fig 6. Heatmap of Average Political Misinformation Polarity vs. Science Polarity

Myth polarities in Figure 5 appear to peak when observing the common users of two myth subreddits. The polarities taper off when comparing the user bases of a myth subreddit and a non-myth subreddit, and colors are significantly lighter. Polarities reach their lowest when a myth subreddit is not at all involved. The shared users of two political misinformation subreddits, interestingly enough, seem to harbor the most myth subreddit users, although not by much. In that sense, myth subreddits appear to be incredibly niche -- user bases with no association to a myth subreddit will have a low polarity.

Figure 6, on the other hand, shows how much more easily political misinformation spreads into other subreddits. Generally, Figure 6 follows a similar trend to Figure 5. Our highest average polarity occurs in the shared user bases of two political subreddits, and the concentration

of these users tapers off as politics are less and less involved. However, the average polarity of any two non-political subreddits is significantly higher than that of the myth subreddits. This subset of users, despite the supposed lack of association with political misinformation, has had a remarkably high exposure to politically misinformative content.

It is clear that users of any one subreddit grouping tend to stay within that grouping, with the exception of subreddits closely related in subject matter (i.e. COVID-19 misinformation subreddits permeating into other COVID-19 subreddits). When these users do venture into other subreddits, however, their ideas meet with varying reception. Myth subreddit users seem to stay within their own group, and most other users will almost never participate in a myth subreddit. Political misinformation, on the other hand, is noticeably diffusing into other subreddits. Perhaps the longevity of urban legends can be attributed to the small but dedicated group surviving these stories. Meanwhile, as "populations beset by social change and economic inequality are uniquely susceptible to end-of-times conspiracy theories" [2], our results confirm that the stresses of the pandemic and election have enabled a greater spread of political misinformation.

[2] Alt, Matt, et al. "The Flashing Warning of QAnon." The New Yorker, 26 Sept. 2020, www.newyorker.com/culture/cultural-comment/the-flashing-warning-of-qanon.