The dataset is derived from the Reddit Pushshift database, which archives posts and comments from the internet forum site Reddit. The data was accessed via the Pushshift api (documentation found [here](https://github.com/pushshift/api)) using a python script to sequentially retrieve data from the database – necessary due to the data volume limitations enforced by the API. Due to this, data is still being retrieved from the database; however, a significant fraction has already been retrieved.

On Reddit, *users* make *posts* within specific communities called *subreddits*. Underneath these *posts*, users can reply in the form of *comments*; these *comments*, in turn, can also be replied to by additional *comments*. As a result, discussion under a reddit *post* expands in a tree layout, with each *comment* having an immediate parent they reply to. We are retrieving comments from the subreddits “sandersforpresident” and “joebiden” in the period from February 2,2020, to August 12, 2020. The data is preprocessed by first trimming unnecessary data from the obtained object, leaving only the fields: link\_id, which indicates the post the comment is made under; parent\_id, the immediate parent of the comment (either another comment or a post if it is a top level comment); author, the username of the user; body, the actual comment itself; and locked, a Boolean value representing whether discussion was restricted under a parent post. Comments from “locked” posts will be discarded, as these are prevented from expanding discussion and any graphical representation drawn from these will be skewed. Further preprocessing will use the parent\_id and link\_id fields to generate a graph representing the interactions between unique users.

With the information grabbed, we can model the interactions between different users on these forums and analyze the presence of echo chambers.

Due to technical issues, the initial data file containing 200k comments was lost; as such, the following statistics are derived from a backup data file of 60k comments. Ranging from May 15, 2020, to August 12, 2020, on the “sandersforpresident” subreddit, 59545 comments represent 21664 unique users, which each user making an average of 2.75 comments and the average length of each comment being 222 characters. 18783 of these comments – 31.5% of comments – are replies to other comments, with the remainder being direct replies to the parent post.