Modeling Polarization on Social Media

Matt Zhang | April 10, 2020

Abstract

The effect of political polarization in American life is a well-known topic, and has been discussed in great detail in recent years. Many researchers have linked this phenomenon to new trends in social media.

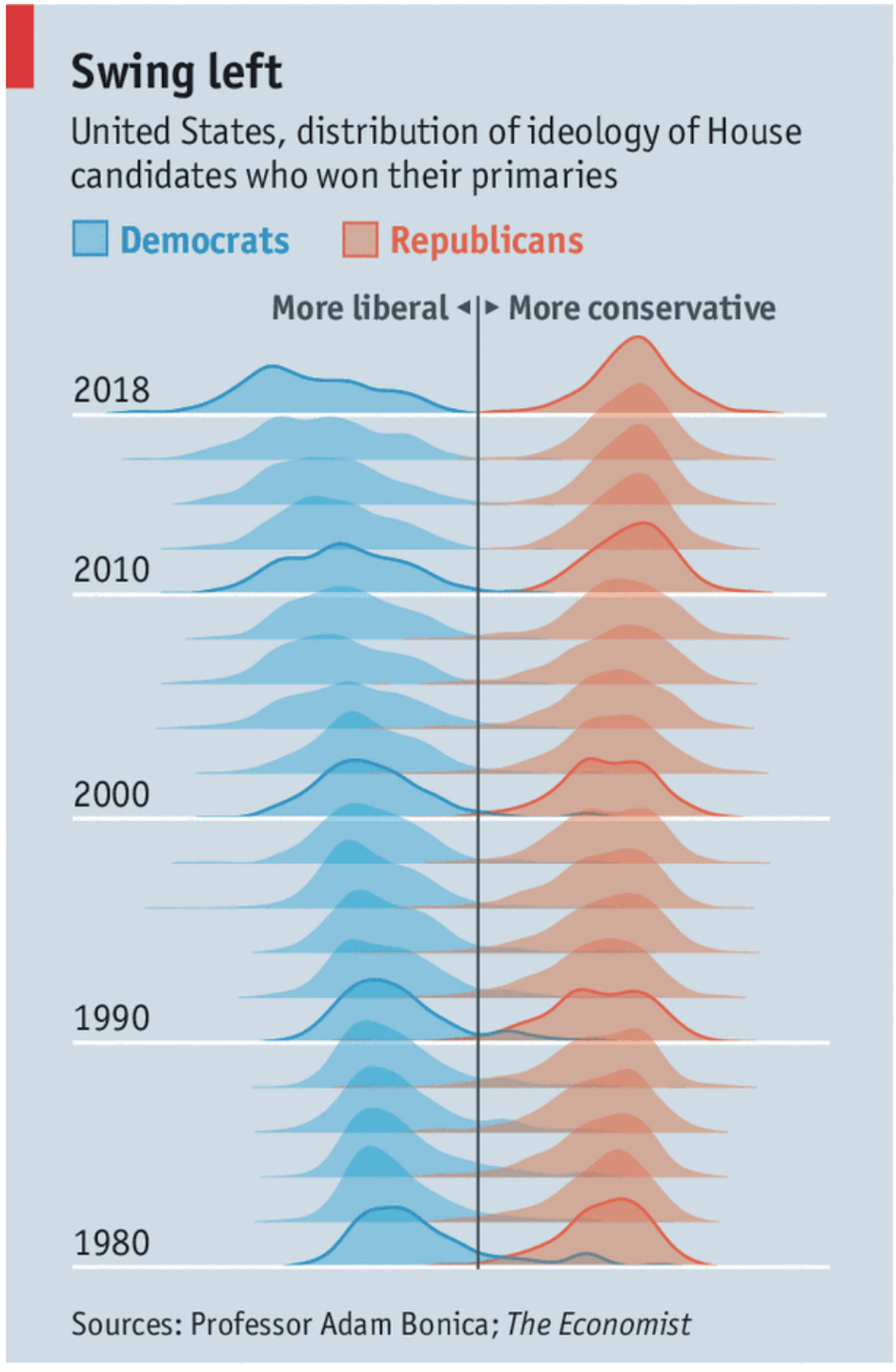
Media companies have taken this criticism seriously, and have spent much effort investigating methods of reducing the effects of online radicalization, including hate groups, echo chambers, search filters, astroturfing, troll farms, and fake news.

The best method of fighting radicalized online behavior remains unknown, and different companies have taken very different approaches. For example, recently this year Facebook announced that they would take a hands-off approach to all political discourse, even provably false political advertisement. On the other hand, Twitter took the opposite approach, banning all political advertisement entirely. Both approaches have led to criticism.

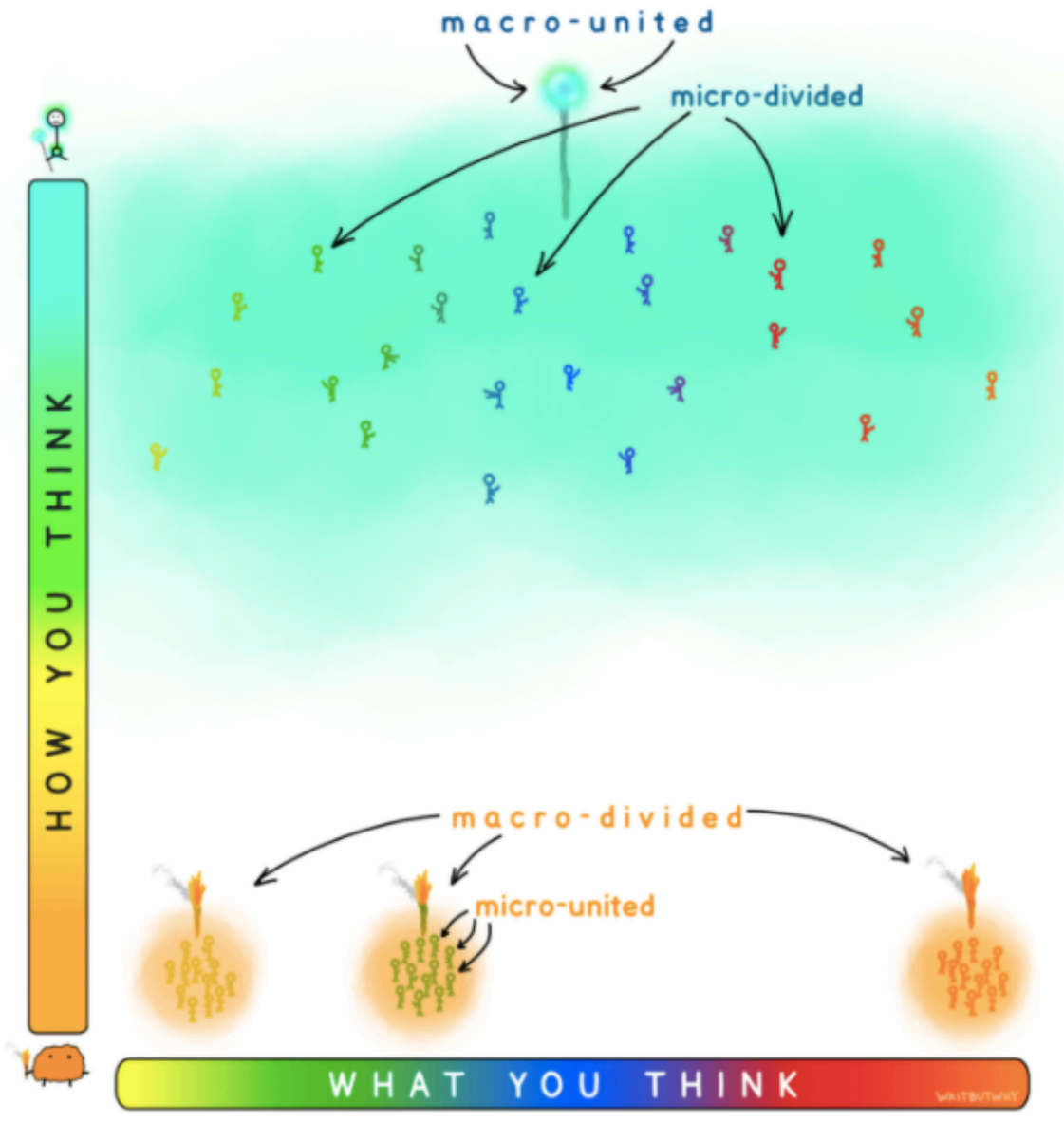
Here we take a different approach, examining the effects of different content and comment ordering systems on polarization in social media. We ask whether the comment-ordering system on a website can reduce radicalization without the need for active moderation.

Background

Increasing Political Polarization



The Economist [1]



[2]

Real-World Effects

New Fake Account Removals Highlight Twitter's Bot Problem Once Again

AUTHOR

Andrew Hutchinson

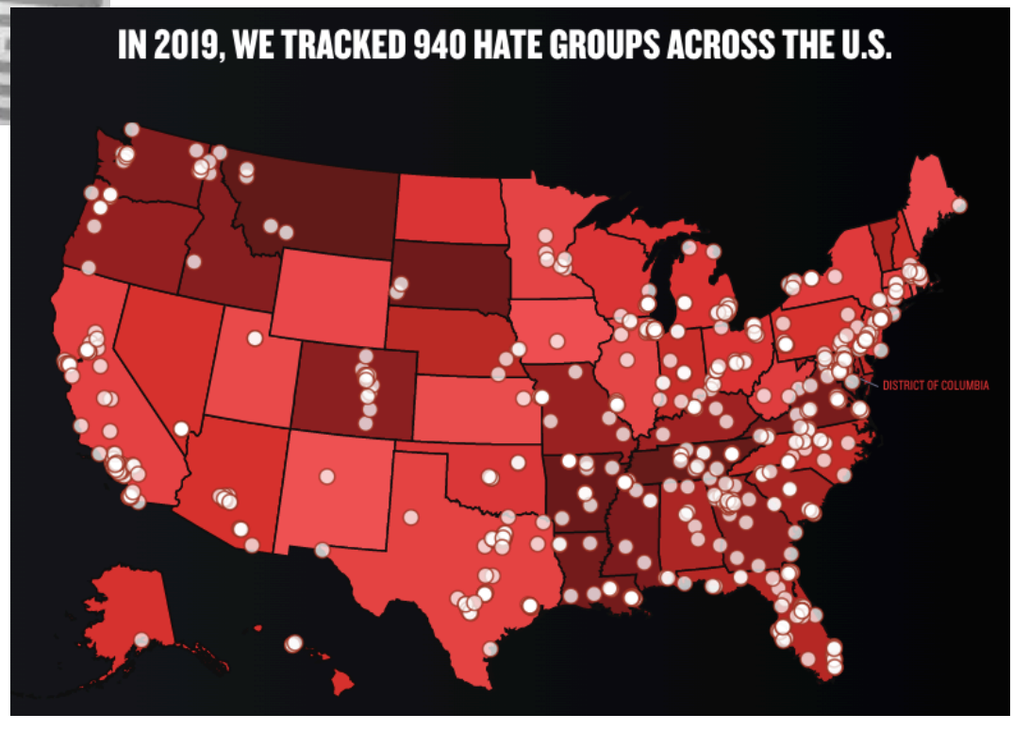
@adhutchinson

PUBLISHED

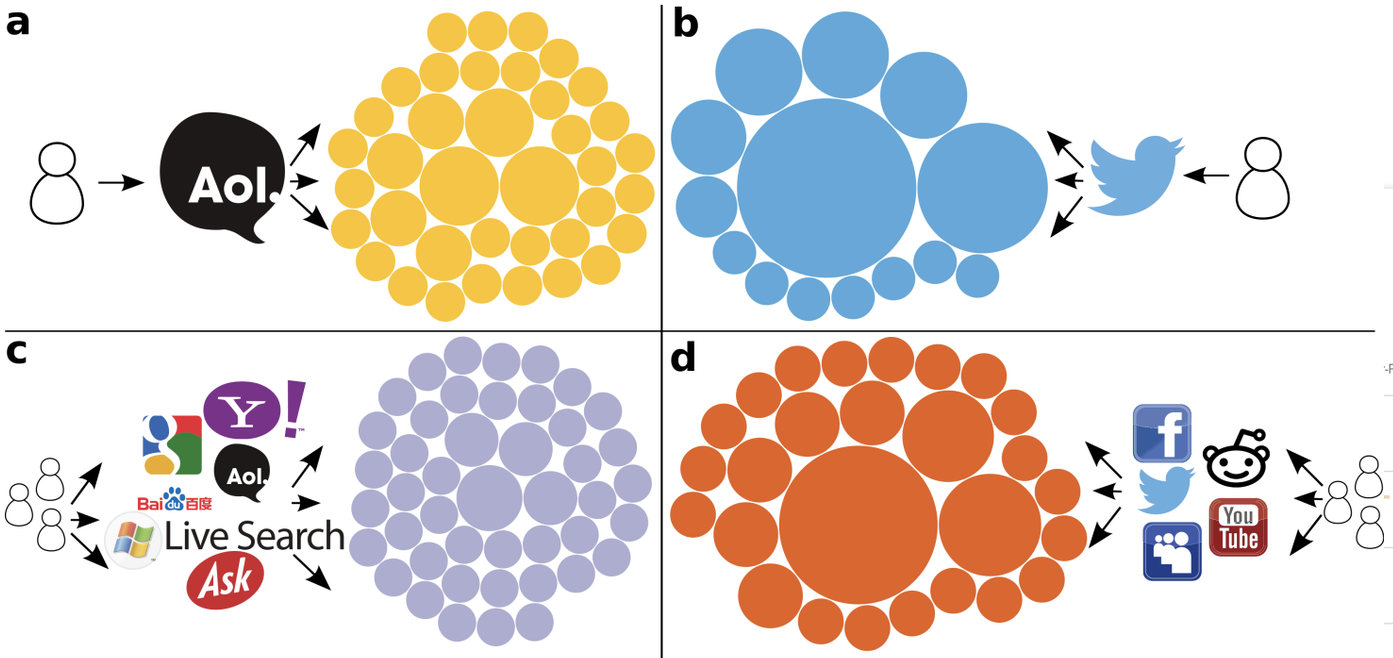
April 4, 2020

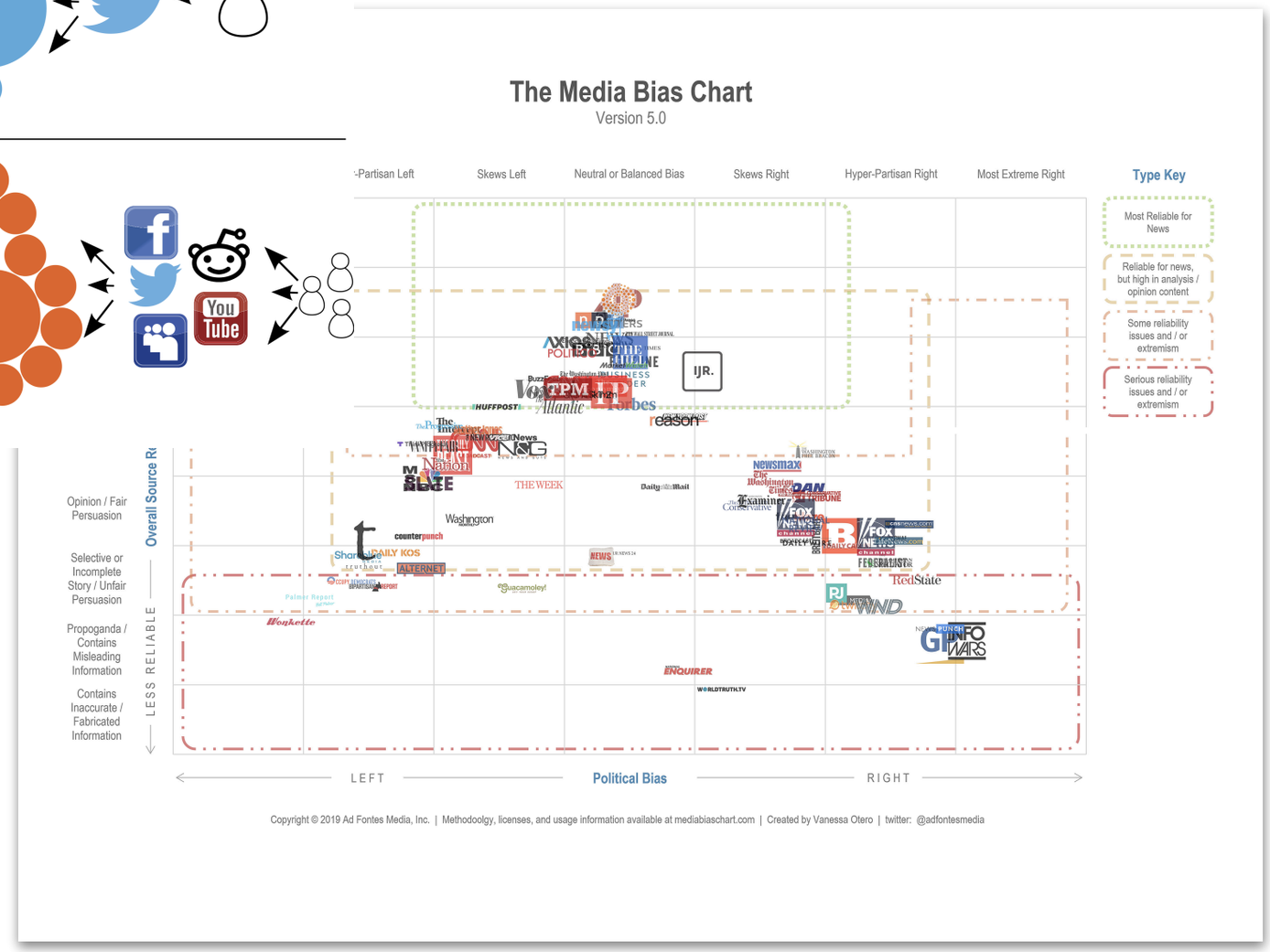
It's no secret to anyone involved in social media circles that Twitter has a bot problem.

For years, users have complained about the impact of bots and fake accounts on the platform, and while various research reports have pegged Twitter's fake profile levels at between 5% and 15%, their presence is likely more significant than that, with researchers repeatedly pointing to massive swarms of bot accounts being used for malicious purpose - in particular, to amplify



Potential Causes





Attempted Fixes

Facebook gives step-by-step instructions on how to spot fake news

By Hannah Sparks and Hannah Frichberg

March 26, 2020 | 7:25pm

TECH

Twitter unveils final details for political ad ban, but it's still looking murky

PUBLISHED FRI, NOV 15 2019•1:30 PM EST | UPDATED FRI, NOV 15 2019•4:16 PM EST

Lauren Feiner
@LAUREN_FEINER

Megan Graham
@MEGANCGRAHAM

SHARE

Opinion Modeling in Sociophysics

Prasetya and Murata *Comput Soc Netw* (2020) 7:2
https://doi.org/10.1186/s40649-019-0076-z

Computational Social Networks

RESEARCH Open Access

A model of opinion and propagation structure polarization in social media

Hafizh A. Prasetya* and Tsuyoshi Murata

*Correspondence:
hafizh@net.c.titech.ac.id
Murata Laboratory, Tokyo Institute of Technology, W8-59 2-12-1, Ookayama, Meguro, Tokyo 152-8552, Japan

Abstract

The issue of polarization in online social media has been gaining attention in recent years amid the changing political landscapes of many parts of the world. Several studies empirically observed the existence of echo chambers in online social media, stimulating a slew of works that tries to model the phenomenon via opinion modeling. Here, we propose a model of opinion dynamics centered around the notion that opinion changes are invoked by news exposure. Our model comes with parameters for opinions and connection strength which are updated through news propagation. We simulate the propagation of multiple news under the model in synthetic networks and observe the evolution of the model's parameters and the propagation structure induced. Unlike previous models, our model successfully exhibited not only polarization of opinion, but also segregated propagation structure. By analyzing the results of our simulations, we found that the formation probability of echo chambers is primarily connected to the news polarization. However, it is also affected by intolerance to dissimilar opinions and how quickly individuals update their opinions. Through simulations on Twitter networks, we found that the behavior of the model is reproducible across different network structure and sizes.

Keywords: Echo chambers, Polarization, Opinion modeling, News propagation

[3]

Table 1 Comparison of the proposed model to popular opinion model families and recent polarization models

Model name	Opinion spectrum	Stochastic process	Interaction mode	Model parameters	Modeled aspect
Families of opinion models					
Voter [13]	Binary	Node selection	Singular	Opinion	Majority influence
Sznajd [70]	Binary	Pair selection	Pair	Opinion	Social validation
Averaging [16]	Continuous	None	Global	Opinion	Opinion averaging, positive influence
Bounded confidence [48]	Continuous	Pair selection	Pair	Opinion, deviation threshold, confidence value	Positive influence, selective exposure, confidence
Recent polarization models					
Argument exchange [50]	Continuous	Pair selection, argument selection	Pair	Opinion, arguments	Argument exchange, homophily
Social feedback [6]	Binary	Pair selection	Global	Opinion, opinion perception, learning rate, deviation probability	Social validation
Algorithmic bias [63]	Continuous	Pair selection	Pair	Opinion, deviation threshold, bias strength	Positive influence, selective exposure, algorithmic bias
Bounded confidence w/ propaganda [71]	Continuous	Pair selection	Pair	Opinion, deviation threshold, confidence value, propaganda threshold	Positive influence, selective exposure, confidence, propaganda influence
Bounded confidence w/ emotion [67]	Continuous	Pair selection	Pair	Opinion, deviation threshold, confidence value	Positive influence, selective exposure, confidence, emotion
Algorithmic bias [74]	Continuous	Pair selection	Pair	Opinion	Positive influence, selective exposure, algorithmic bias
Confidence-tolerance [59]	Continuous	Pair selection	Pair	Opinion, deviation threshold, confidence, rewiring probability	Positive influence, selective exposure, confidence dynamic, connection rewiring
News percolation [72]	Continuous	Seed selection	Propagating broadcast	Opinion, news fitness, sharing threshold	News influence, homogeneity
Proposed model	Continuous	Seed selection, propagation success	Propagating broadcast	Individual opinion, connection strength, news opinion, update rate, tolerance	News influence, selective exposure, connection strength

[3]

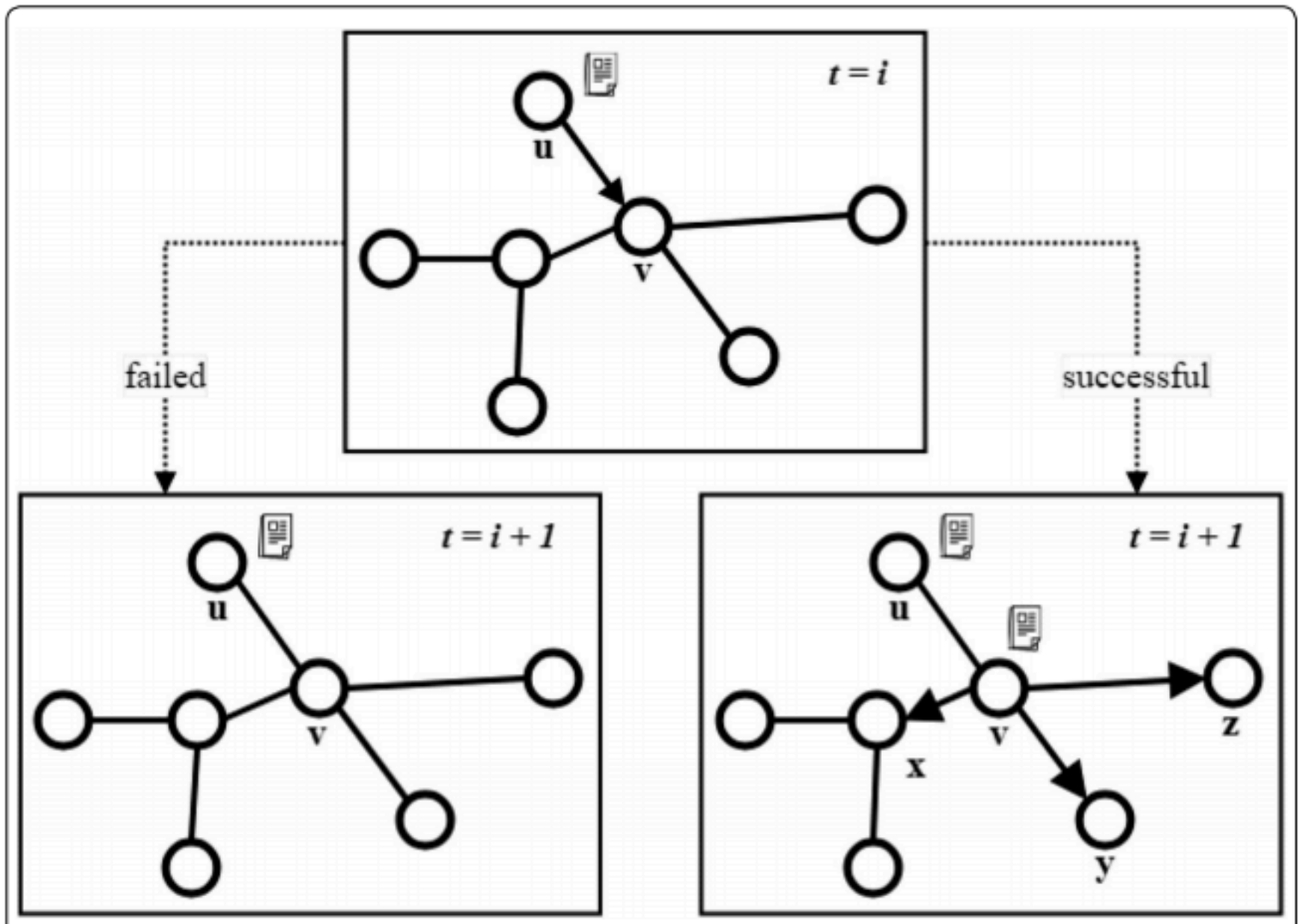


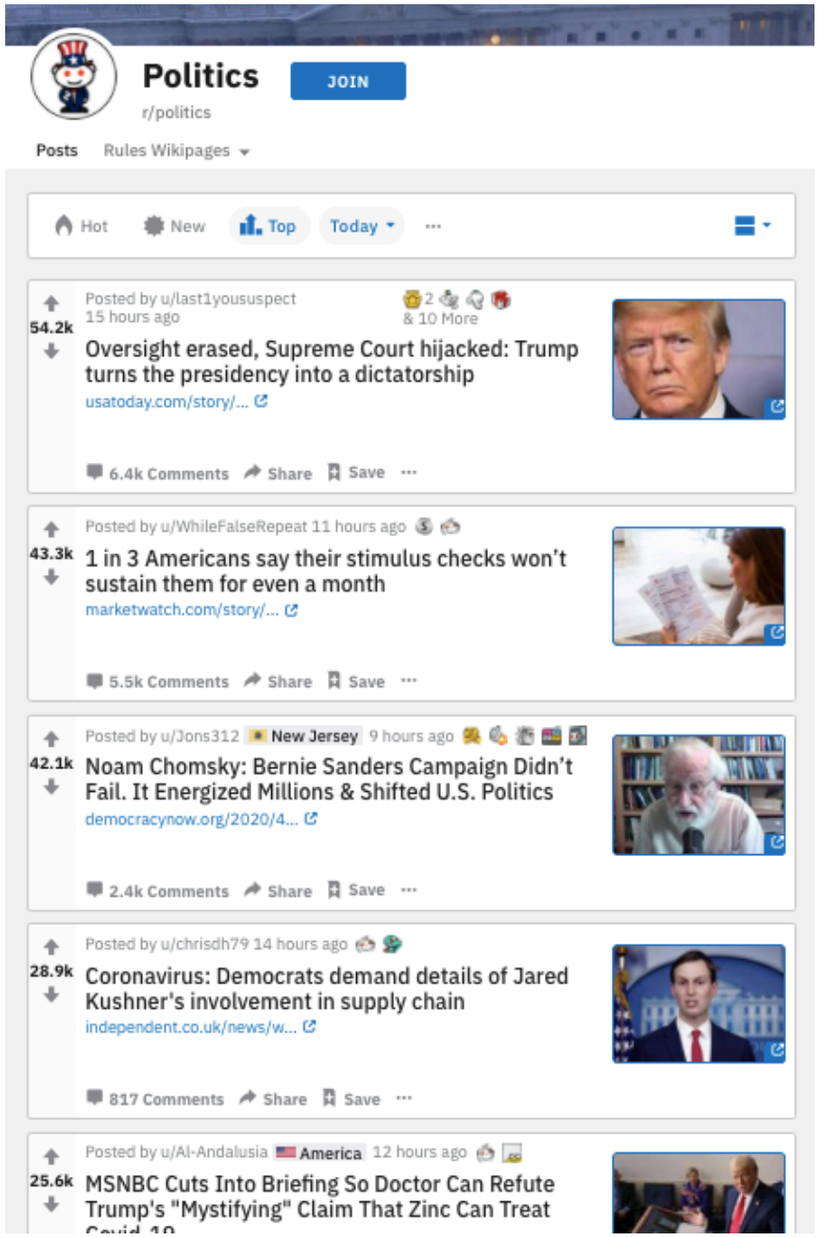
Fig. 1 An example of one propagation under the IC model: node *u* tries to propagate to node *v* in $t = i$ (top). If the propagation is successful (bottom, right), node *v* then propagates to all of its neighbor in $t = i + 1$. If it fails (bottom, left) the cascade stops because there are no more propagation

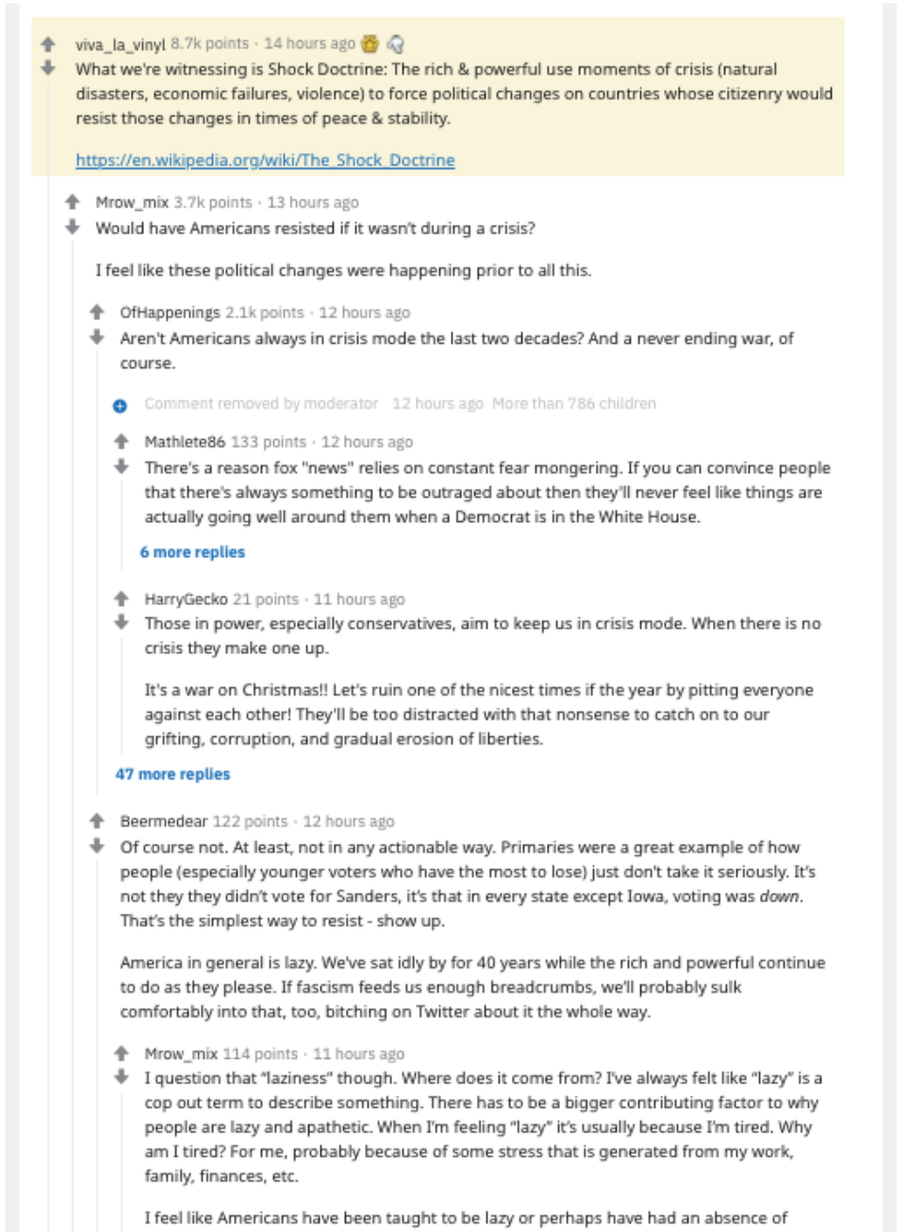
[3]

Project Overview

Description of Project

Many models use graphs such as the IC network seen earlier. This is good for modeling webs of connected individuals, as may be seen on e.g. Facebook or Twitter. However, I'm more interested in modeling Reddit, since that's what I use. So in these studies I create a new type of model, to reflect the Reddit structure.





Project Setup

In this project we create an environment where individuals are allowed to submit posts and to comment on them.

Here we will use a continuous opinion spectrum model. We have a population of individuals, where each individual has an opinion on the issue that can change over time.

We examine a simplified problem - a single issue with two sides.

Opinions are updated asynchronously, with one individual at a time altering their opinion based on interaction with their environment. They will read comments, change their opinion, and leave a comment.

We submit posts and allow the individuals to talk to each other, and we examine how the opinions of the population change over time.

Though most people do not vote or comment on social media, here we will examine only the subset of people who do vote and comment, since they determine the political climate of a website.

Individual-Environment Interaction Process

In these experiments we allow individuals to read and leave comments one at a time. Comments are voted on, with the most-upvoted comments showing up at the top.

First the individual will read a certain number of comments and have their opinion affected accordingly. If we are ranking comments by popularity, they will upvote or downvote comments according to how much they agree with the comment and the persuasiveness of the commenter. Then they will leave their own comment with their own opinion.

After everybody leaves their comments, we start over with a new thread.

Code Repository

Available on GitHub at:

git@github.com:BucketOfFish/SocialPolarization.git

Project Details

Details - Individuals

Individuals are randomly initialized with certain values:

Opinion - how the individual feels about the topic being discussed.
Opinion is in the range [-1, 1], with 0 being neutral.

Bias - how strongly the individual will hold their own opinion and find opposing opinions unconvincing.
Bias is in the range [0, 1]. High-bias individuals will find even very persuasive arguments to be unconvincing.

Persuasiveness - unpersuasive individuals (e.g. very rude ones) cause others to disagree with their opinion.
Persuasiveness is in the range [0, 1], with 0 being unpersuasive.

Openness - how quickly an individual will change their opinion.
Openness is in the range [0, 1], with 0 being unwilling to change and 1 being quick to change.

Patience - how many comments an individual is likely to read in a thread before leaving their own opinion.
Patience has a lower bound of 0.

Details - Reading and Leaving Comments

We use a nested comment system, where individuals can either leave a top-level comment, or respond to another comment. However, each individual ends up leaving exactly one comment.

At any given comment nesting level, individuals read a number of comments at that level determined by their patience.

For each comment, the individual can decide to read replies to the comment, leave a reply, or do nothing. If the individual does not see any comments that they want to reply to, or if they are done reading, they leave their own comment at the level where they left off.

The individual will be more likely to read replies to a comment if there are a lot of them. The formula used is:
P = 0.2 + 0.8 * (1 - np.exp(-comment.n_total_replies/20))

The individual is more likely to respond to a comment if it expresses an opinion either similar or very different to their own. The formula is:
opinion_diff = abs(comment.opinion - self.opinion)
if (opinion_diff > 1) P = opinion_diff/2
elif (opinion_diff < 0.2) P = (0.2 - opinion_diff) * 5
else P = 0

Details - Changing Opinions

After reading a comment, an individual's opinions are changed as follows:

opinion_diff = comment.opinion - self.opinion
persuasiveness = comment.persuasiveness - self.bias * abs(opinion_diff)
self.opinion = min(max(opinion_diff*self.openness*persuasiveness + self.opinion, -1), 1)

If an individual is biased enough, the perceived persuasiveness of a comment can be negative, so that the individual is pushed further towards their initial viewpoint. This demonstrates the concept of confirmation bias.

Experiment #1

Modeling Polarization

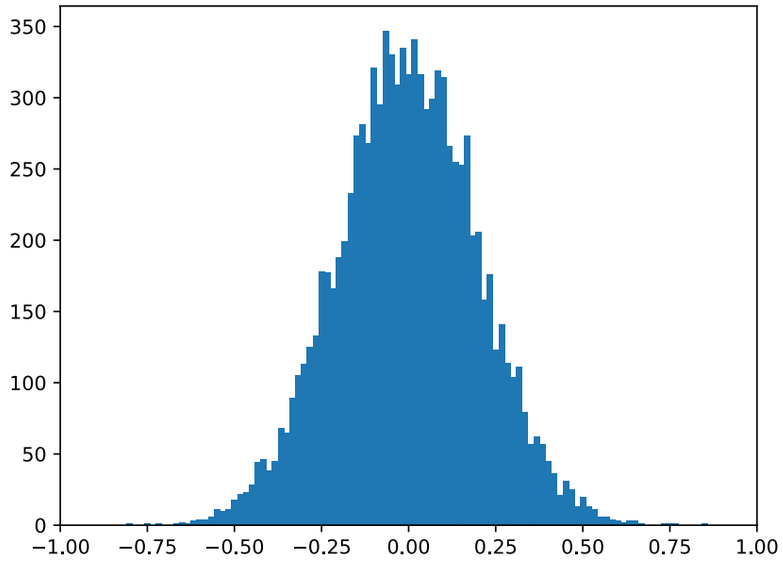
General Population

Numbers set to what I think people online are actually like:

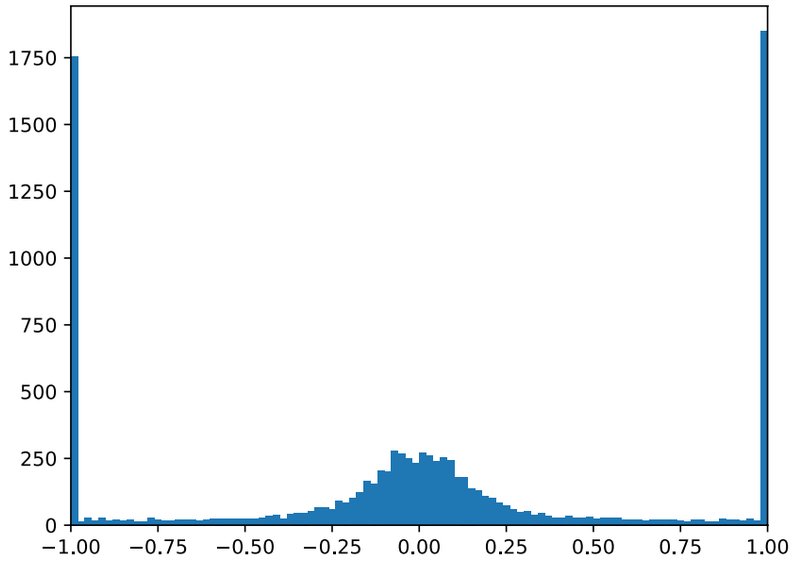
Opinion is initialized as a Gaussian of mean 0 and standard deviation 0.2.
Openness is Gaussian with mean 0.4 and standard deviation 0.05.
Persuasiveness is Gaussian with mean 0.2 and standard deviation 0.1.
Openness is Gaussian with mean 0.3 and standard deviation 0.05.
Patience is Poisson with a lambda of 5.

Using a population of 10,000 across 25 comment threads.

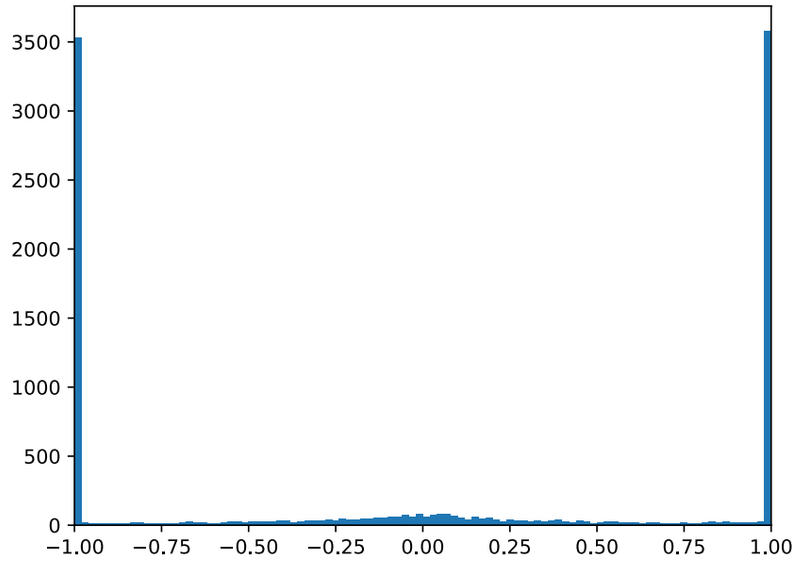
Results on next slide. We clearly see polarization developing almost immediately. Note that this is in a situation where no new information enters the system. Rather, all changes in opinion are only based on people talking to each other based on their initial opinions.



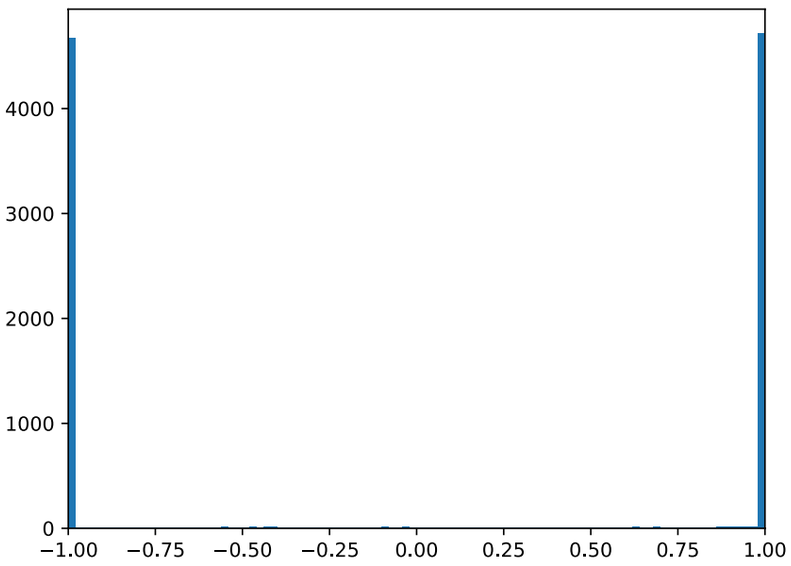
initial



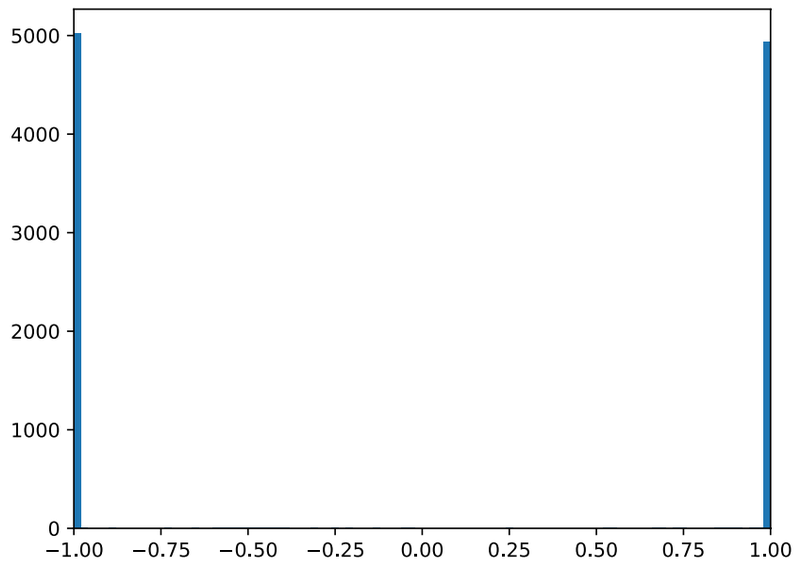
after 5 threads



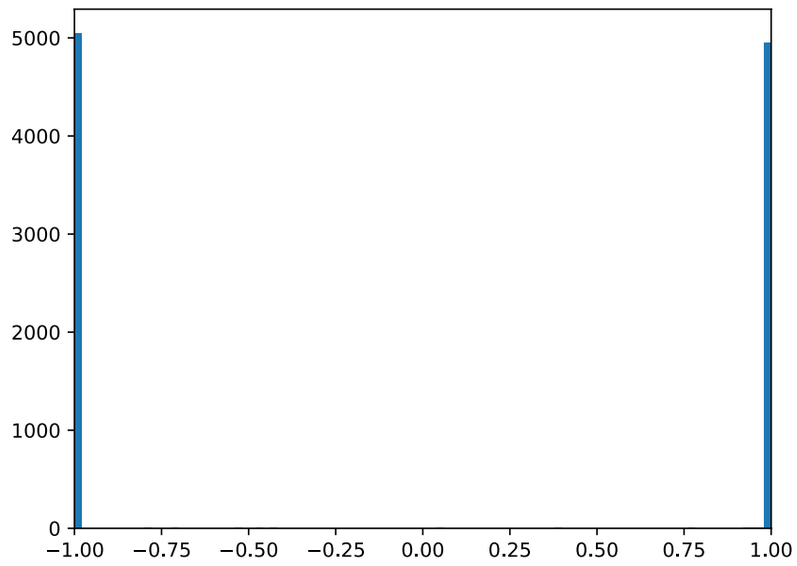
after 10 threads



after 15 threads



after 20 threads



after 25 threads

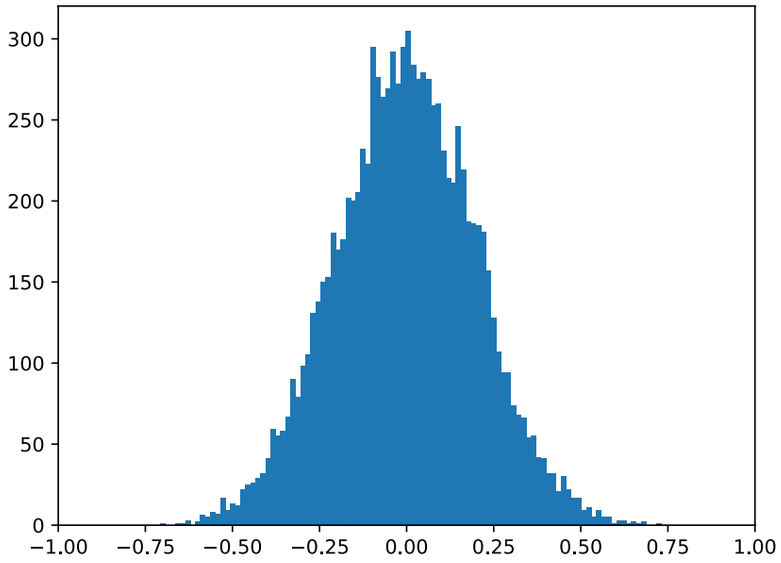
Better Comment Quality

We increase persuasiveness by a little.
Comments which are entirely unpersuasive (e.g. ones saying "I agree" or offering an insult) are less common.
Most comments have at least some argumentation, even if it's weak.

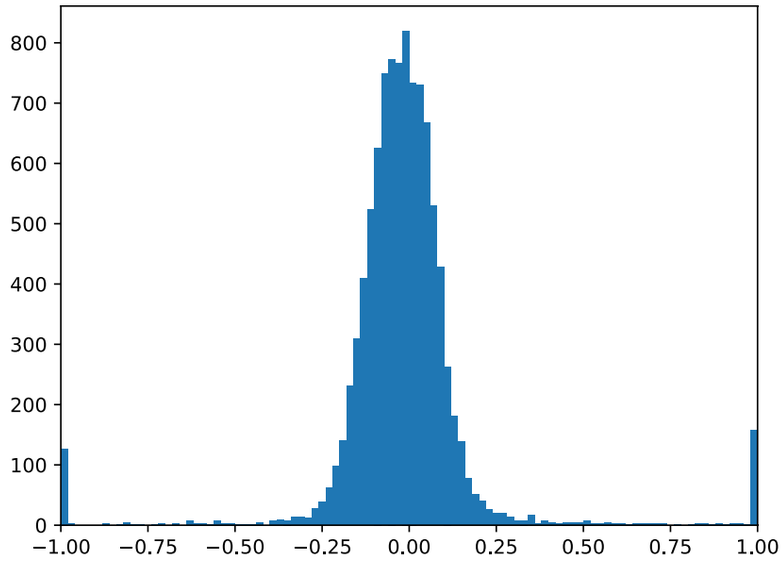
Persuasiveness is Gaussian with mean 0.4 and standard deviation 0.1.

Results on next slide. We see less extremists, but they still exist. Most people tend to converge towards an opinion, but it skews away from the midpoint of the initial population.

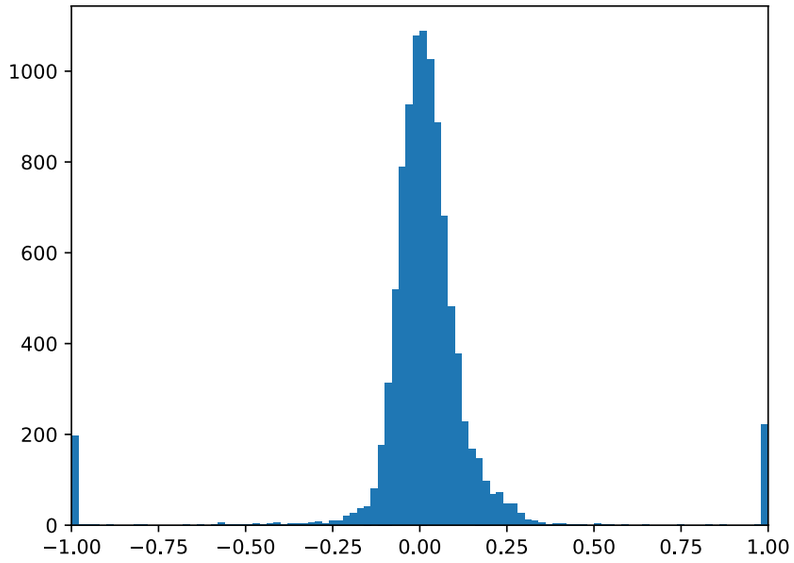
Note that people are still as defensive of their opinions as before (high bias).



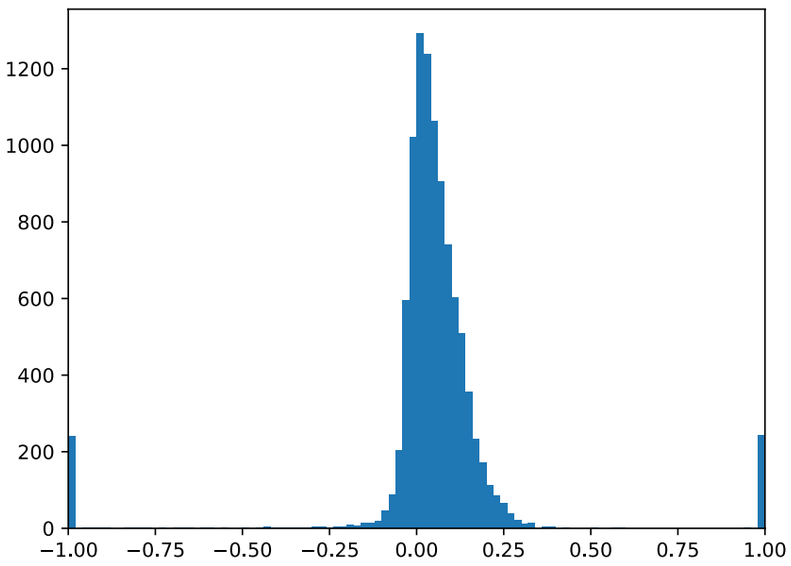
initial



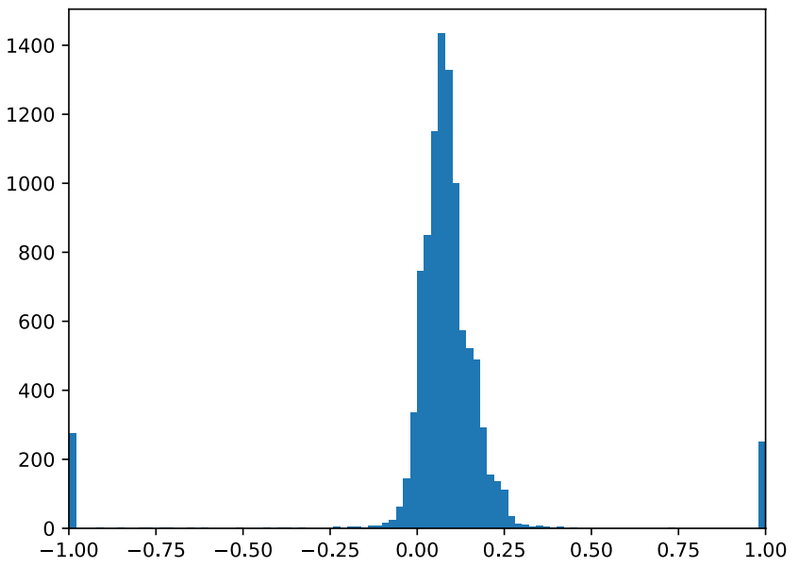
after 5 threads



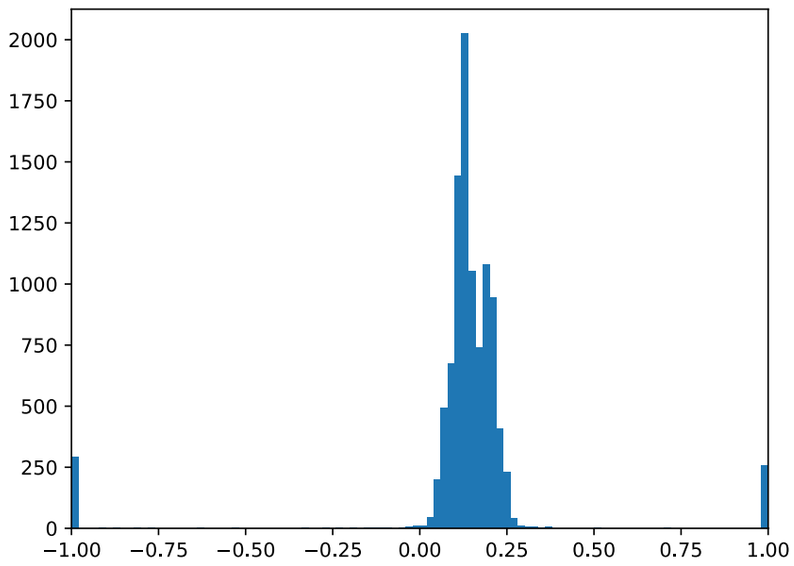
after 10 threads



after 15 threads



after 20 threads



after 25 threads

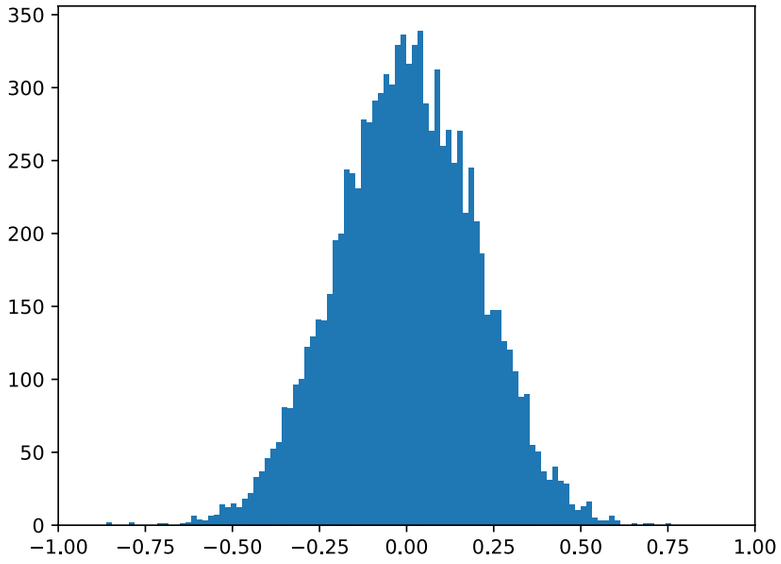
Academic Comment Quality

Persuasiveness is increased by a lot. Comments are now all cited academic works with high rigor and neutral language.

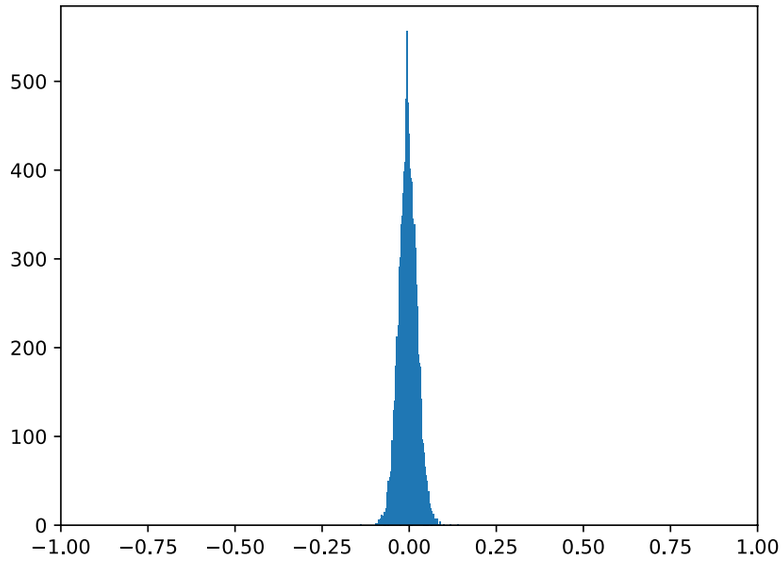
Persuasiveness is Gaussian with mean 0.8 and standard deviation 0.1.

Results on next slide. Opinion very rapidly converges to a centralized consensus. You can't really see it on the plots, but there are still a very small number of extremists! These are people with abnormally high bias who happened to start with an opinion slightly different from the mainstream.

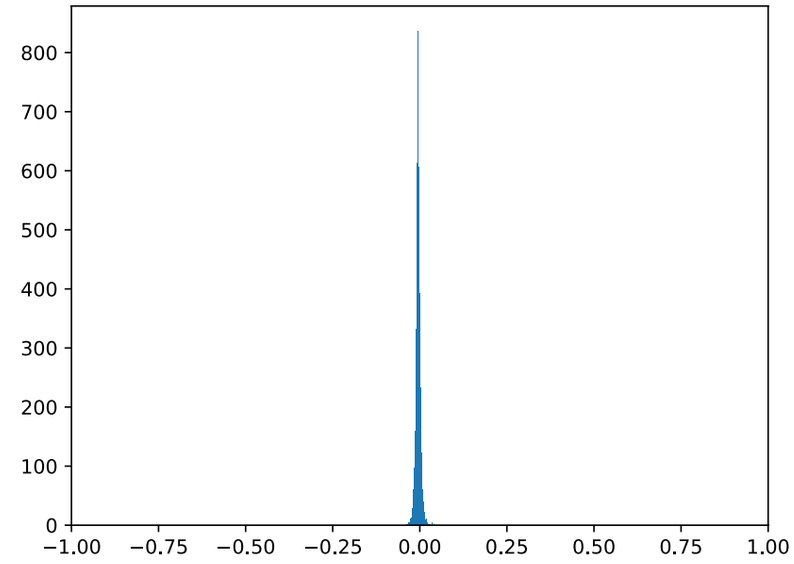
Note that people are still as intrinsically biased as before.



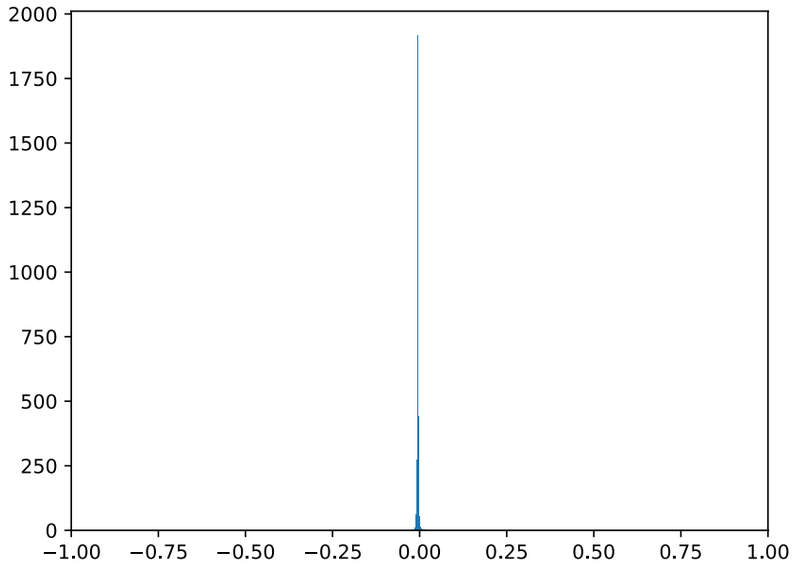
initial



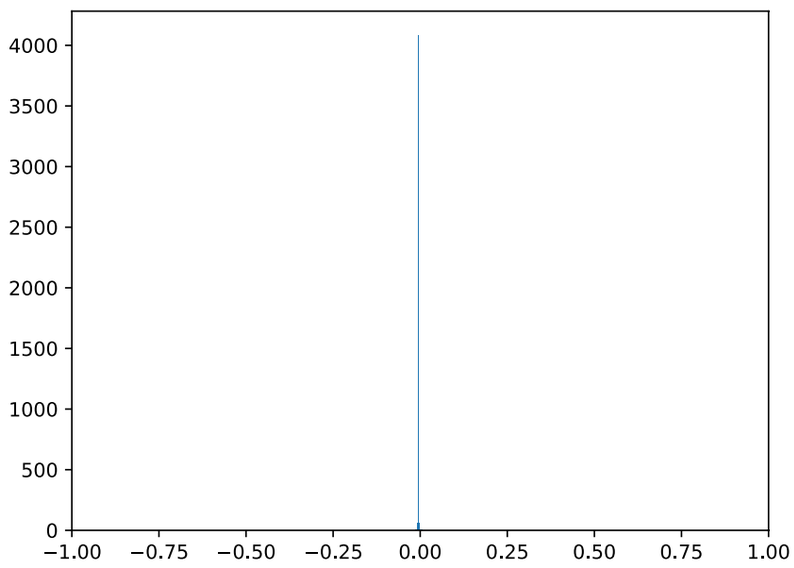
after 5 threads



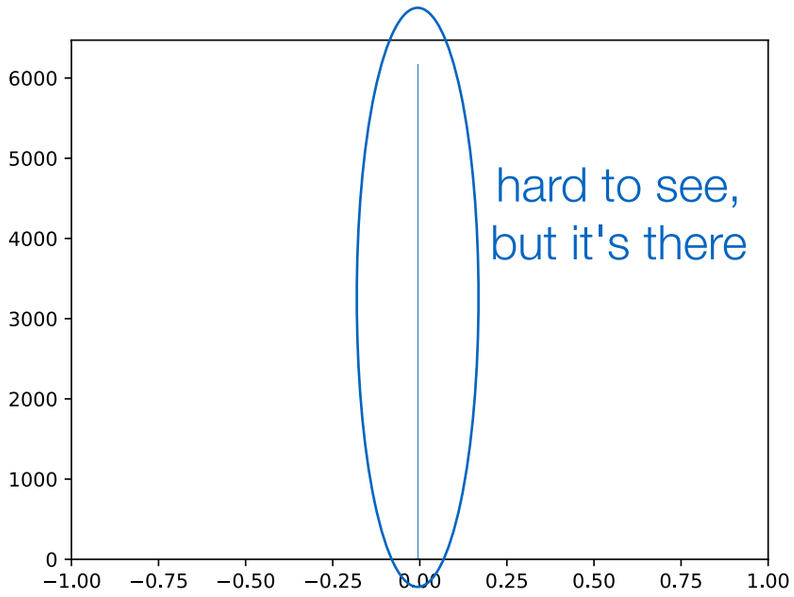
after 10 threads



after 15 threads



after 20 threads



after 25 threads

Other Modeling Options

You can look at the following situations on your own by editing the code.
These are just some examples.

Apathetic population - low bias, low persuasiveness
Close-minded population - low openness
People who don't pay attention to social media - low patience
Mostly non-experts - low persuasiveness for most of the population, with high persuasiveness for a few

Experiment #2

Bimodal Opinion Distribution

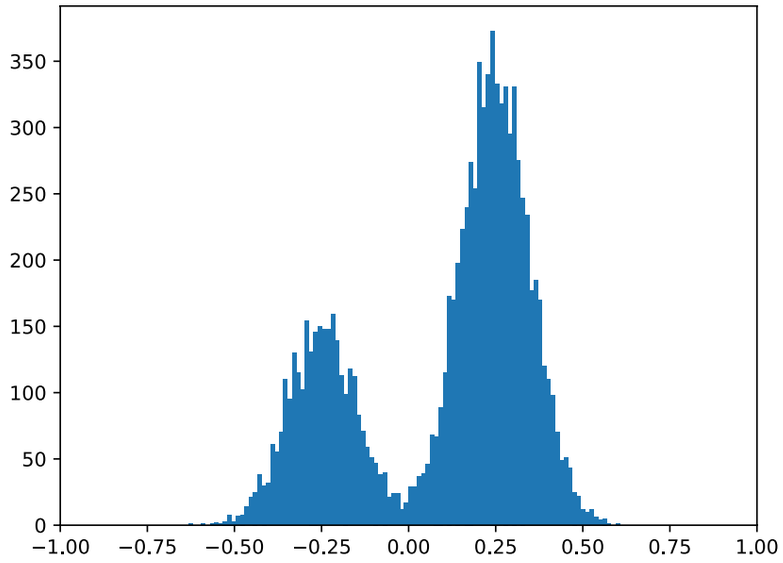
Two Initial Opinion Groups

Let's start with a population where people are already divided into two camps.

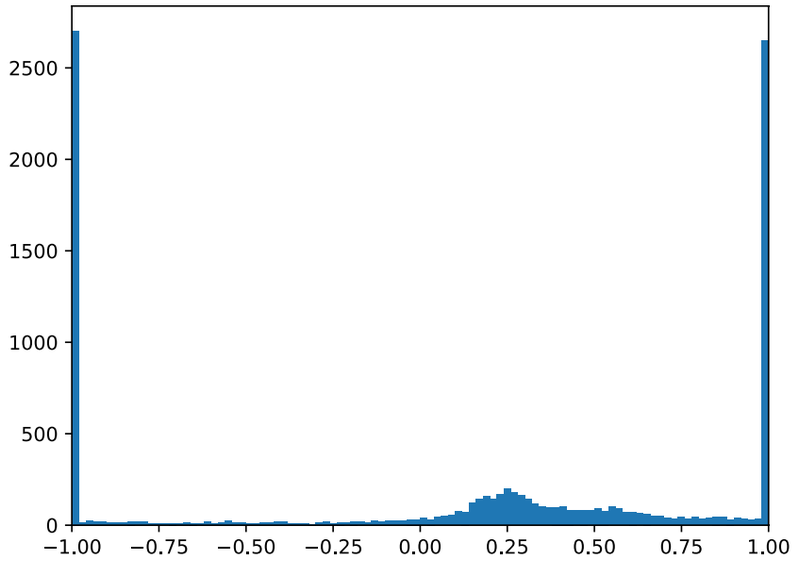
We will use the same parameters as before, for the three populations we examined.

The only difference will be that initial opinions are distributed as the combination of two Gaussians.
30% of the population opinion will be distributed as a Gaussian of mean -0.25 and standard deviation 0.1.
70% of the population opinion will be distributed as a Gaussian of mean 0.25 and standard deviation 0.1.

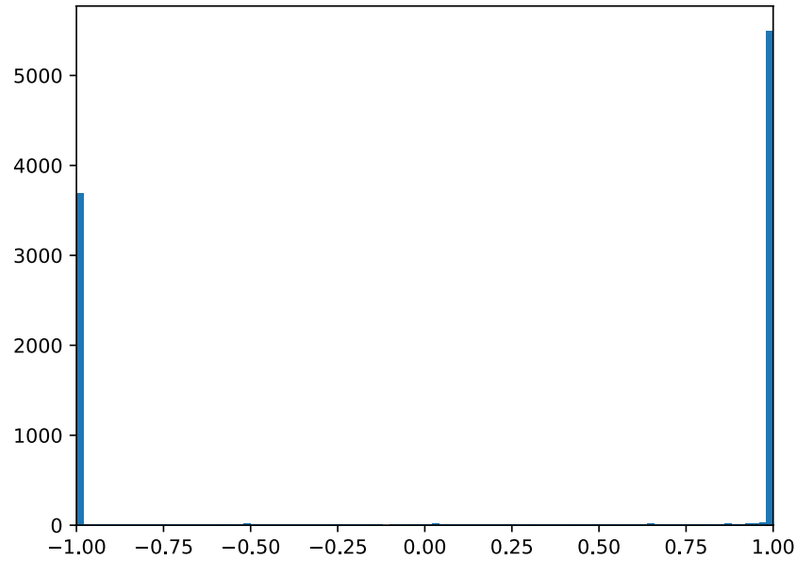
On the next slide we show results for the general population.
The slide after that shows results for the population with slightly better comments.
After that are academic-level comments.



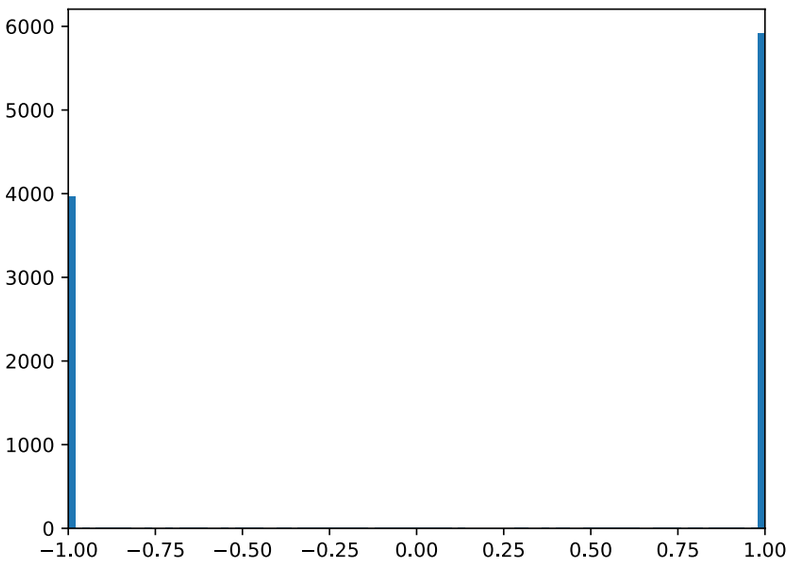
initial



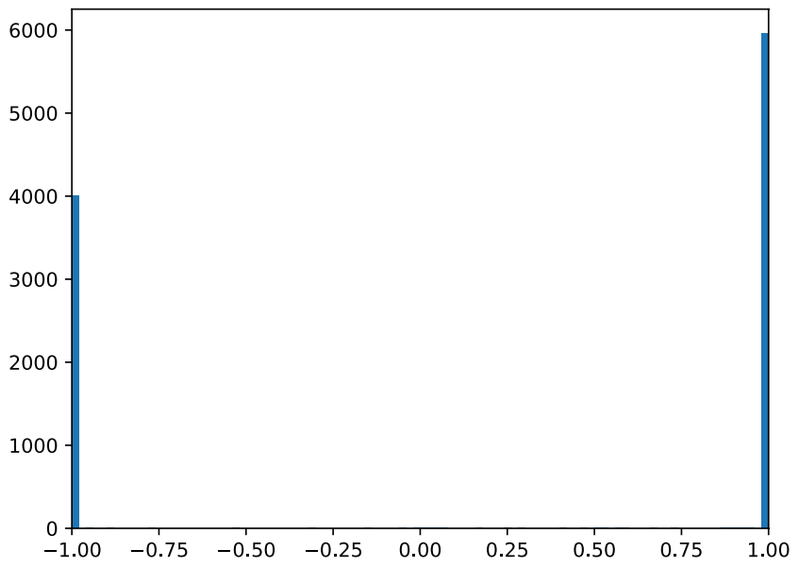
after 5 threads



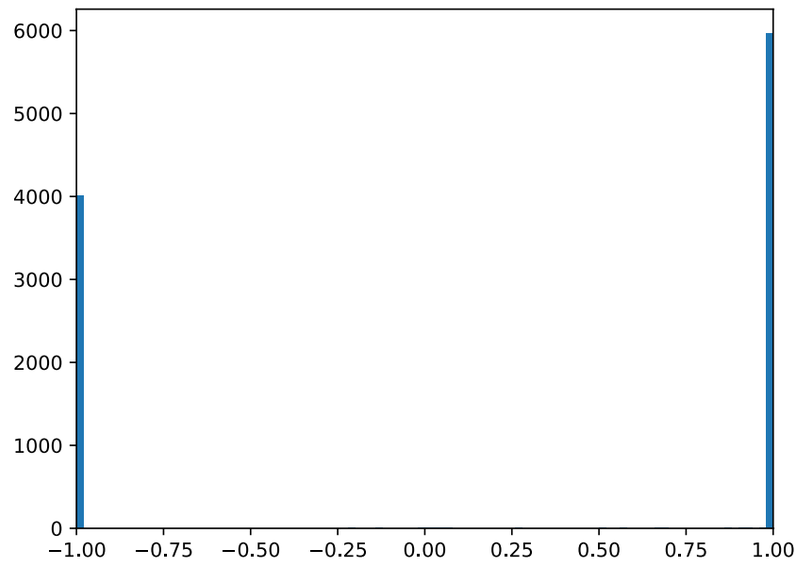
after 10 threads



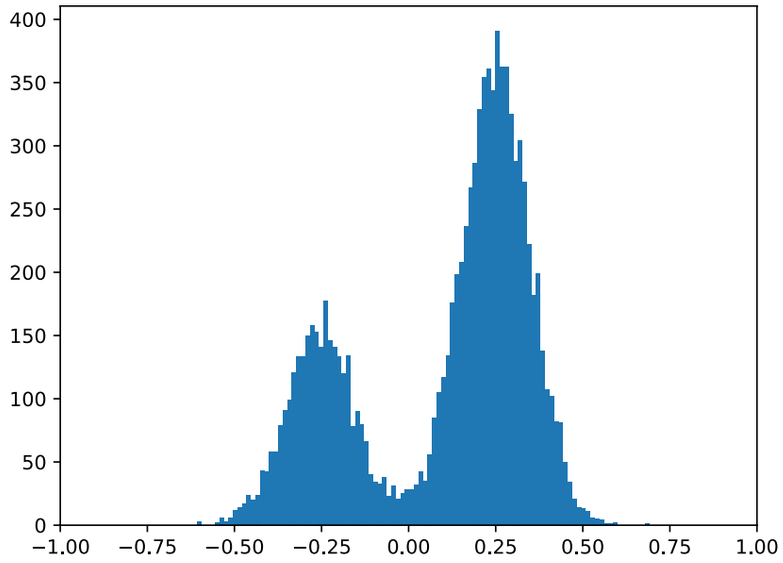
after 15 threads



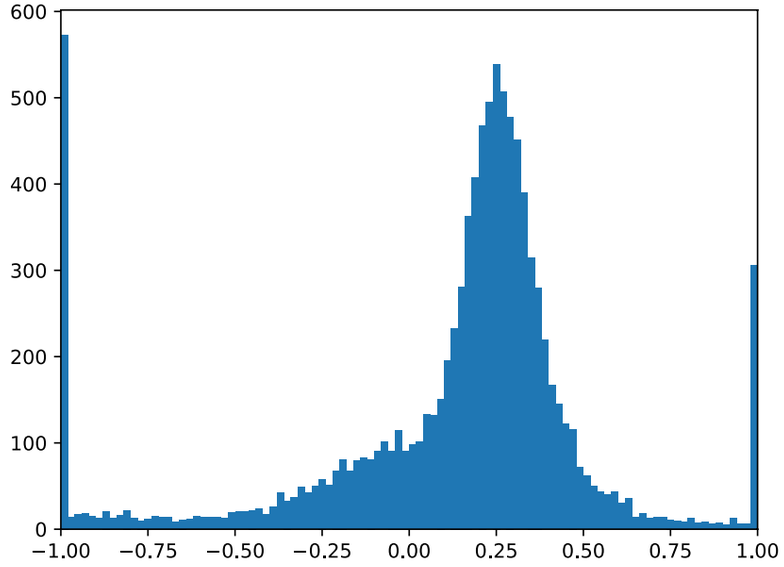
after 20 threads



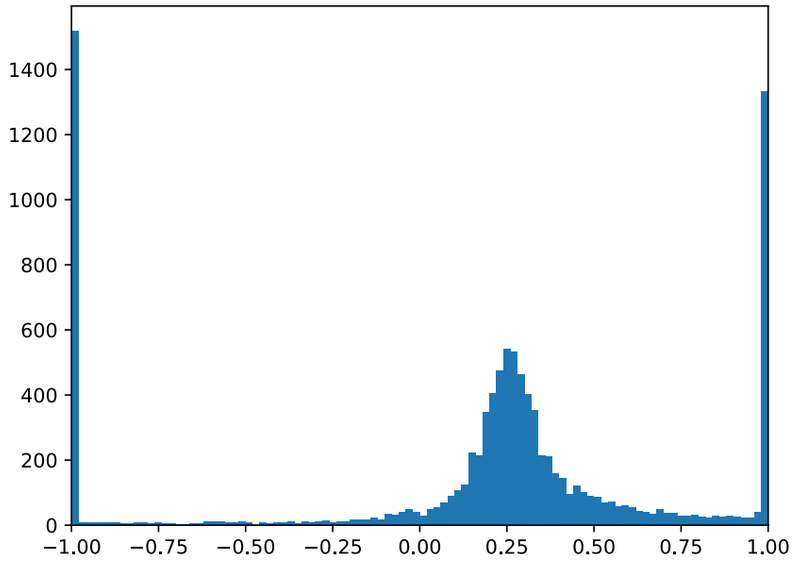
after 25 threads



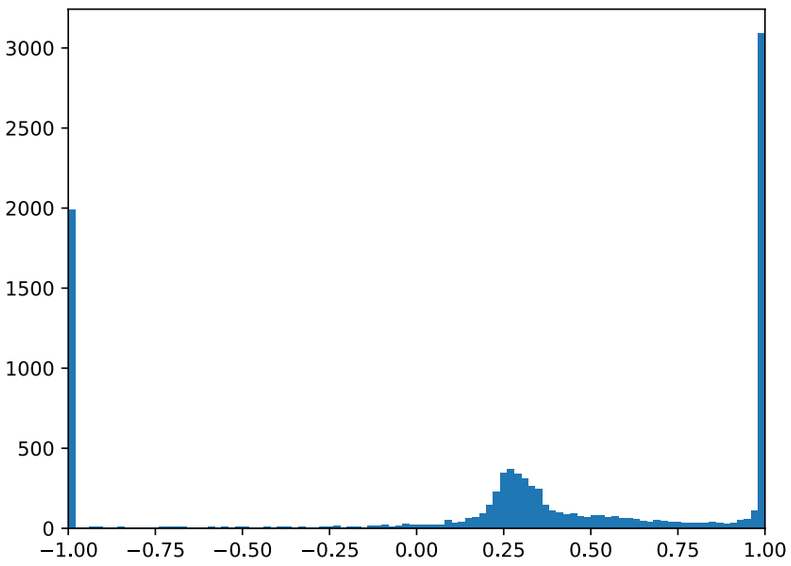
initial



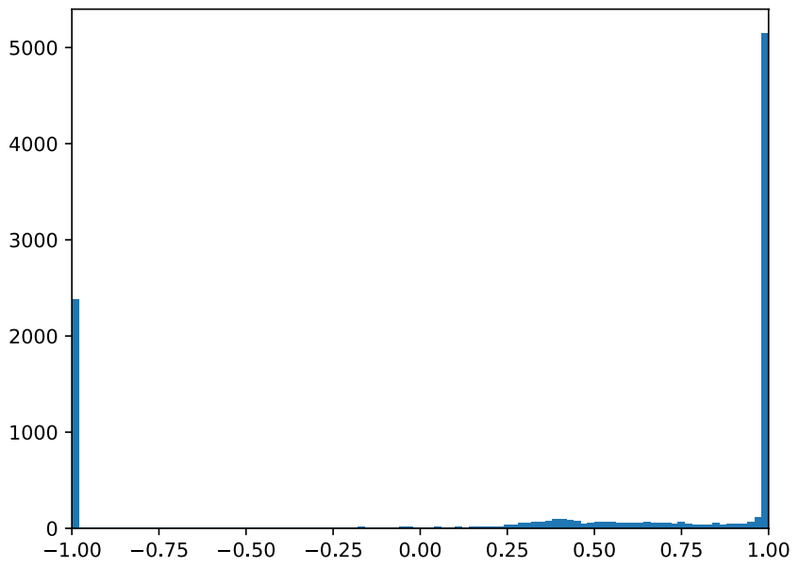
after 5 threads



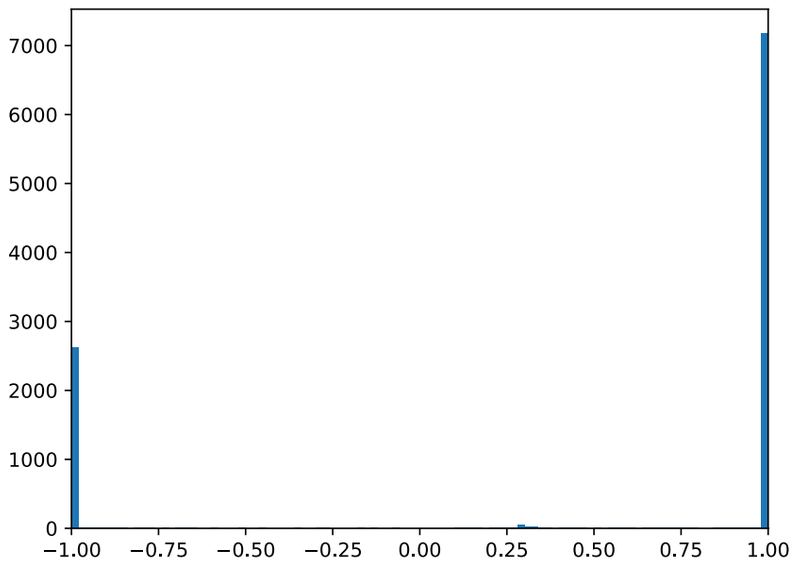
after 10 threads



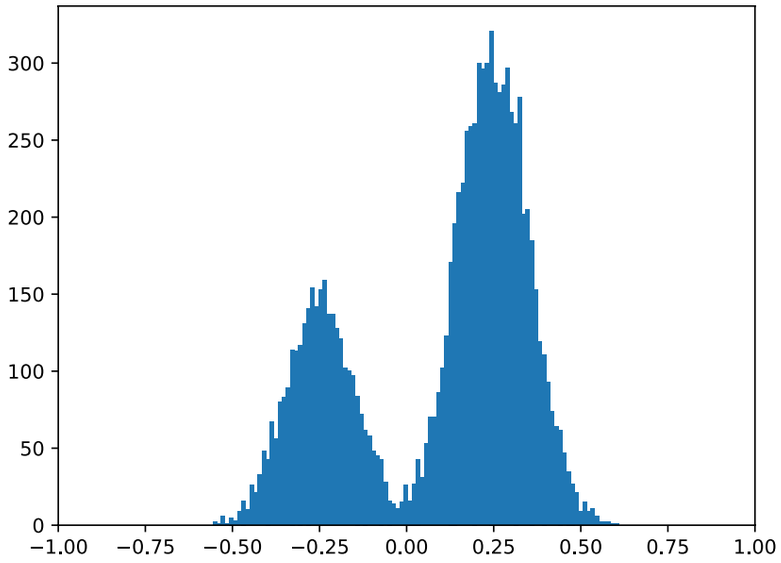
after 15 threads



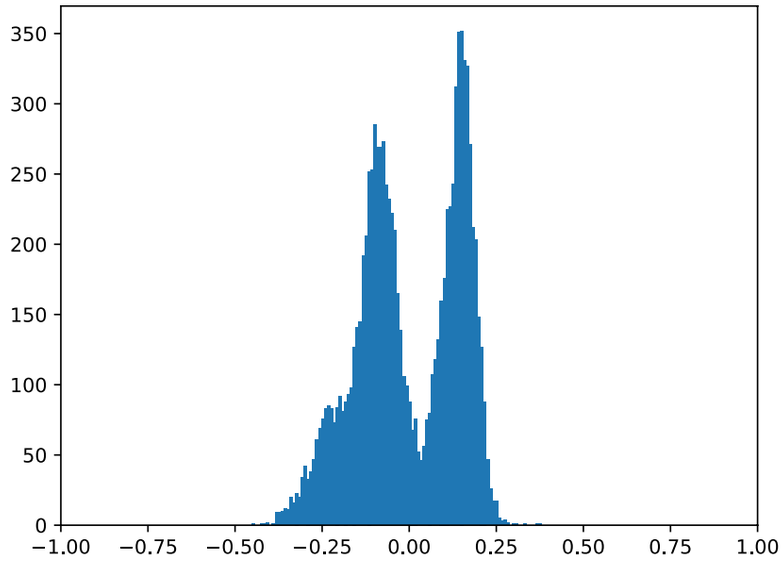
after 20 threads



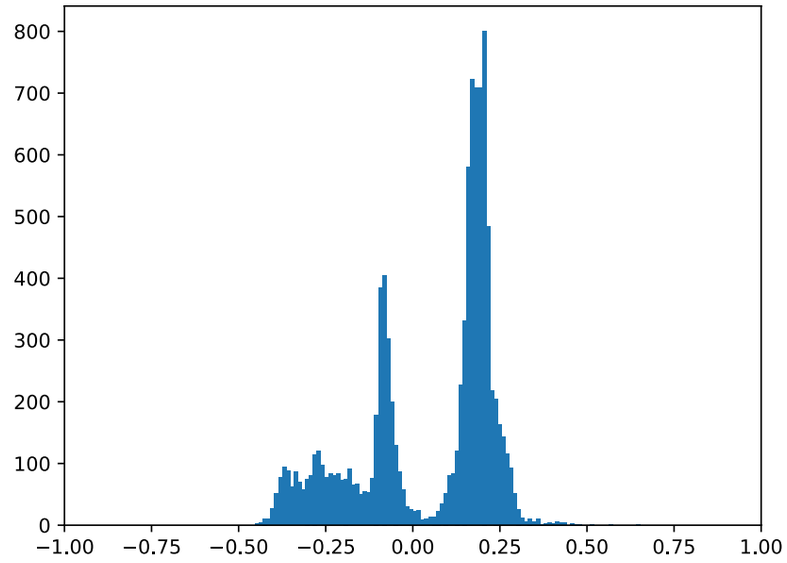
after 25 threads



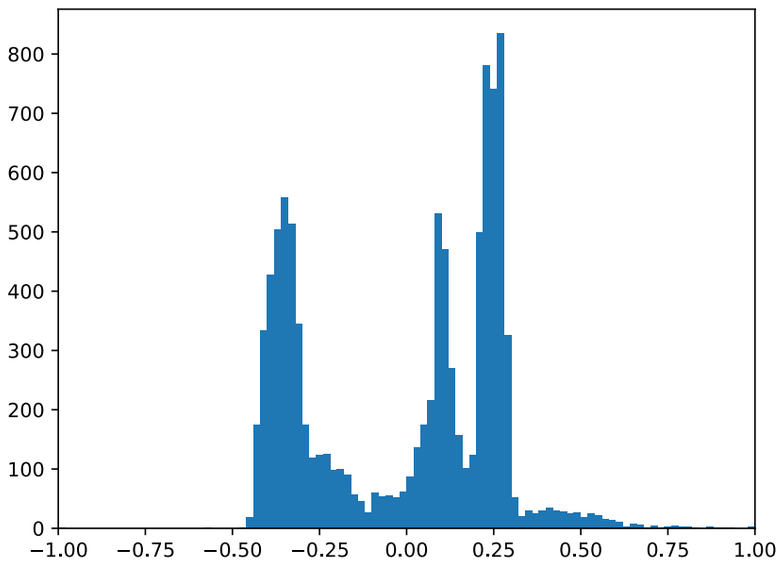
initial



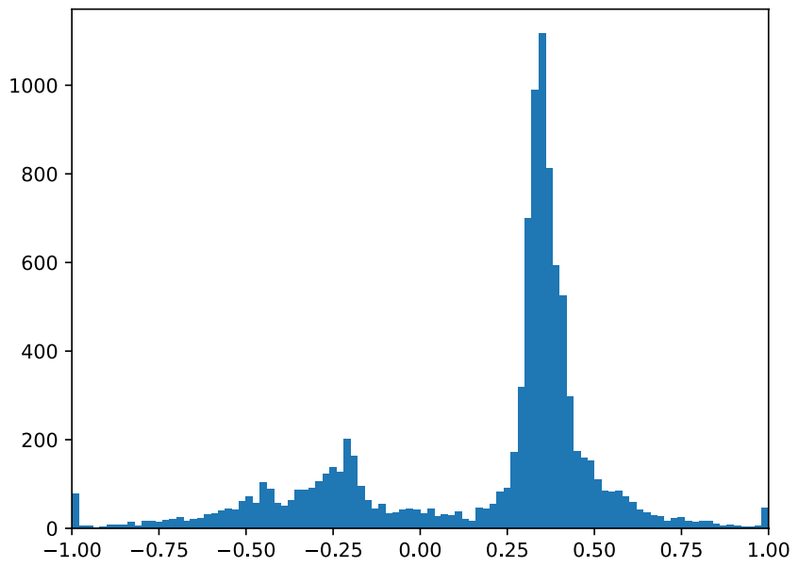
after 5 threads



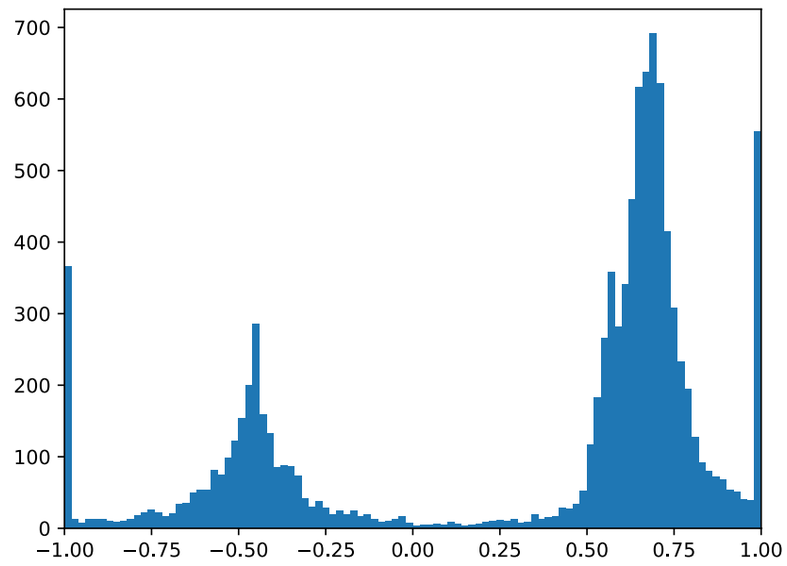
after 10 threads



after 15 threads



after 20 threads

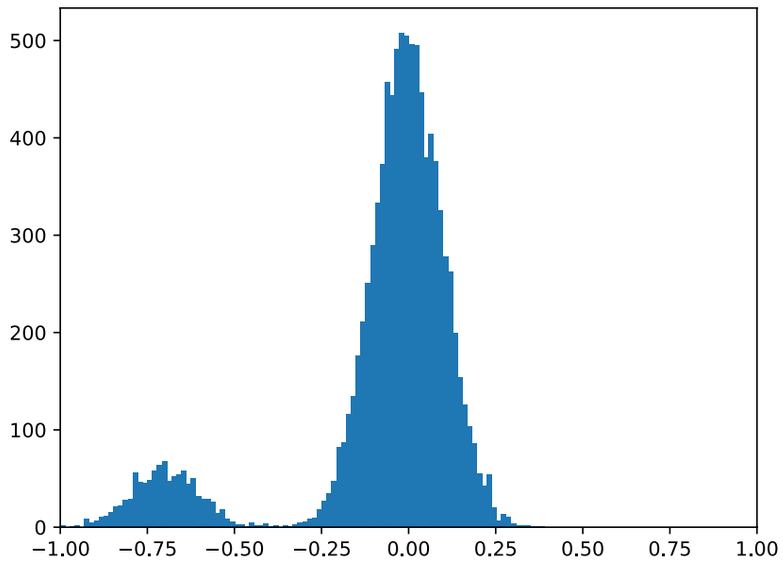


after 25 threads

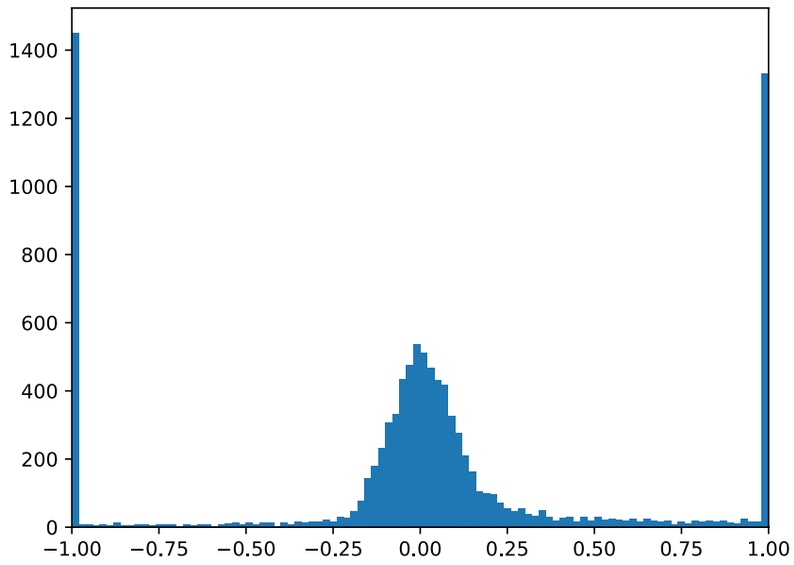
Initial Small Extremist Group

10% of the population opinion will be distributed as a Gaussian of mean -0.7 and standard deviation 0.1.
90% of the population opinion will be distributed as a Gaussian of mean 0 and standard deviation 0.1.

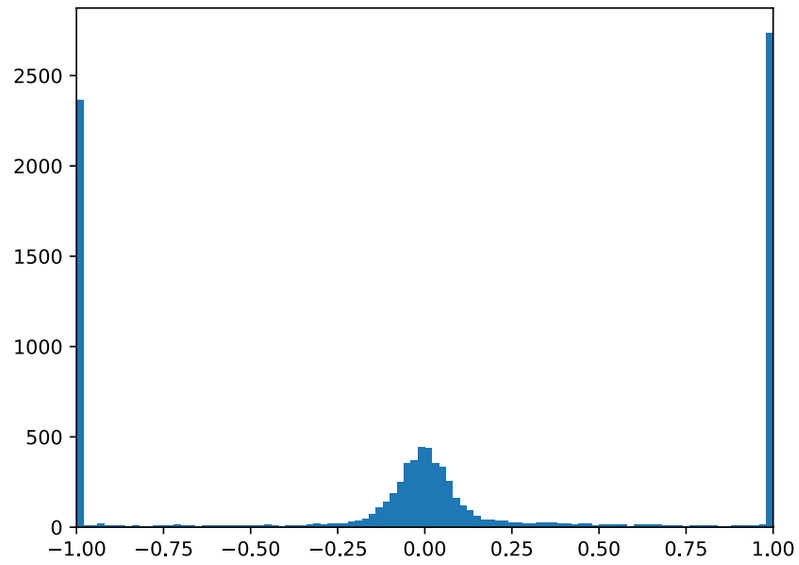
On the next slide we show results for the general population.
The slide after that shows results for the population with slightly better comments.
After that are academic-level comments.



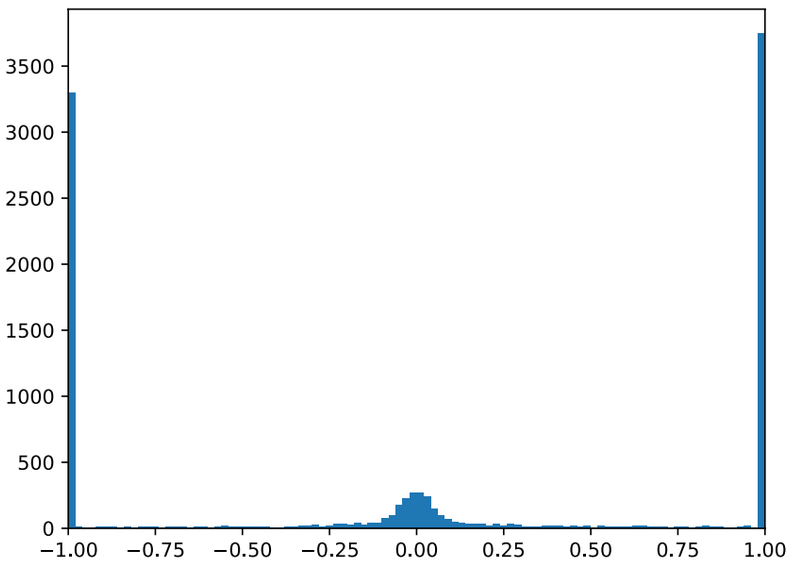
initial



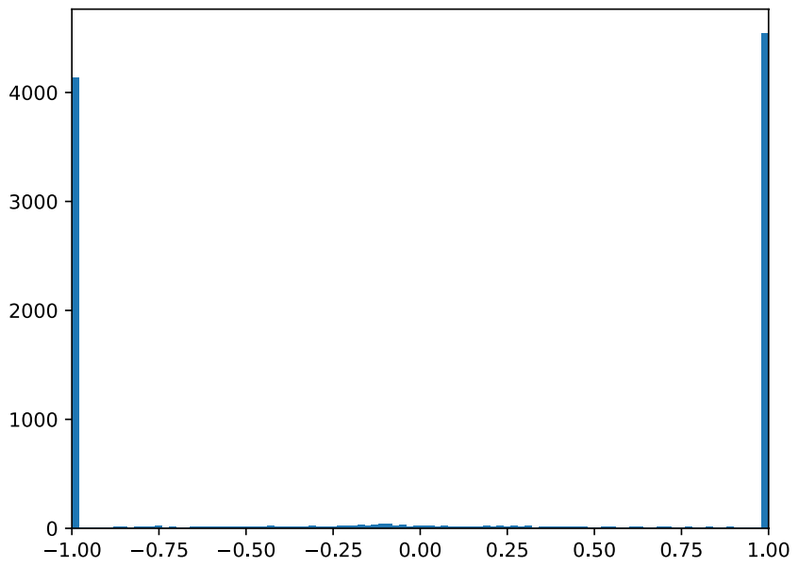
after 5 threads



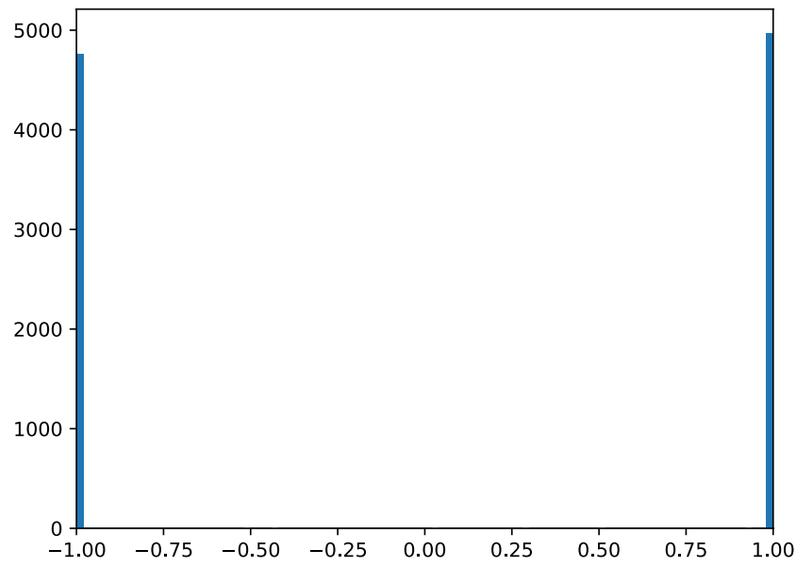
after 10 threads



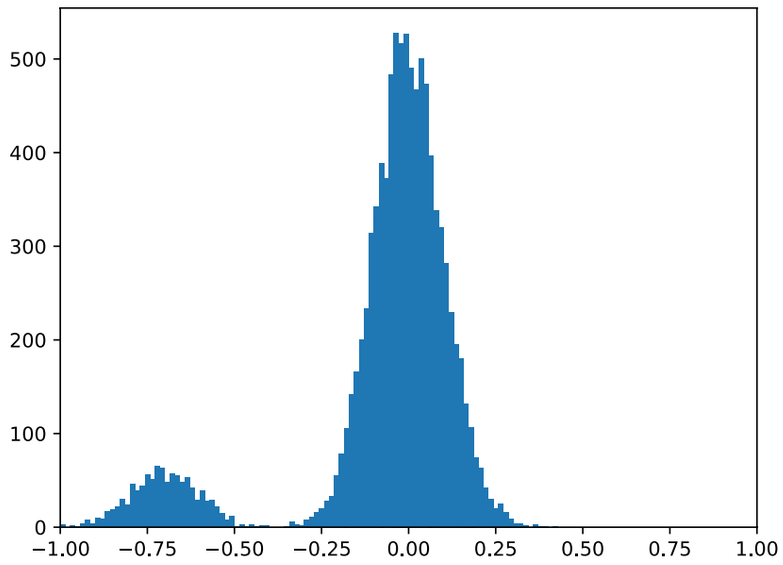
after 15 threads



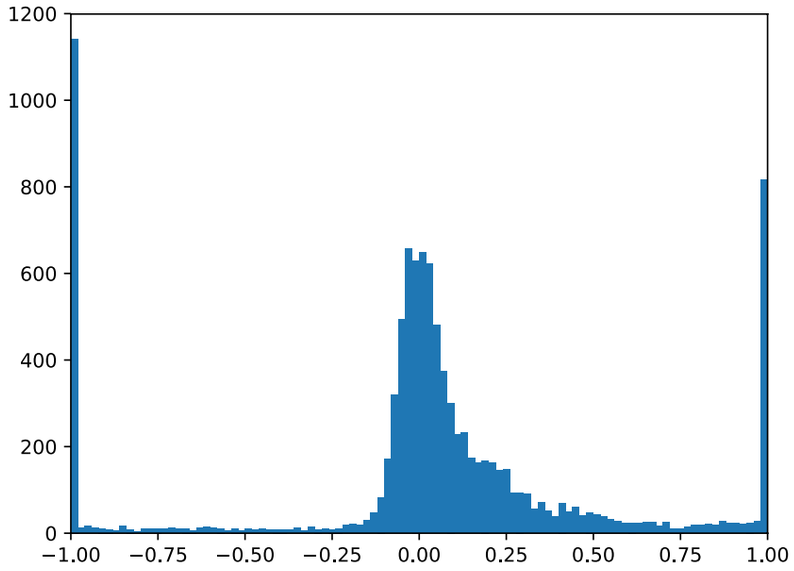
after 20 threads



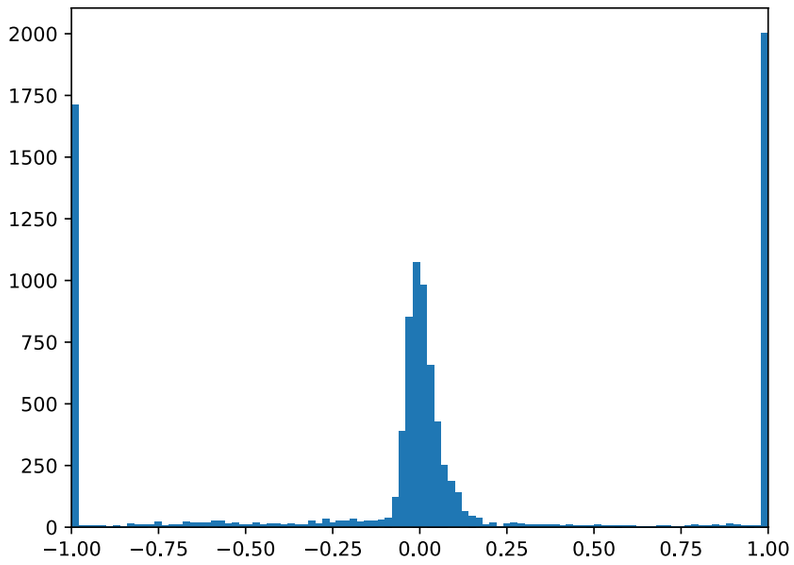
after 25 threads



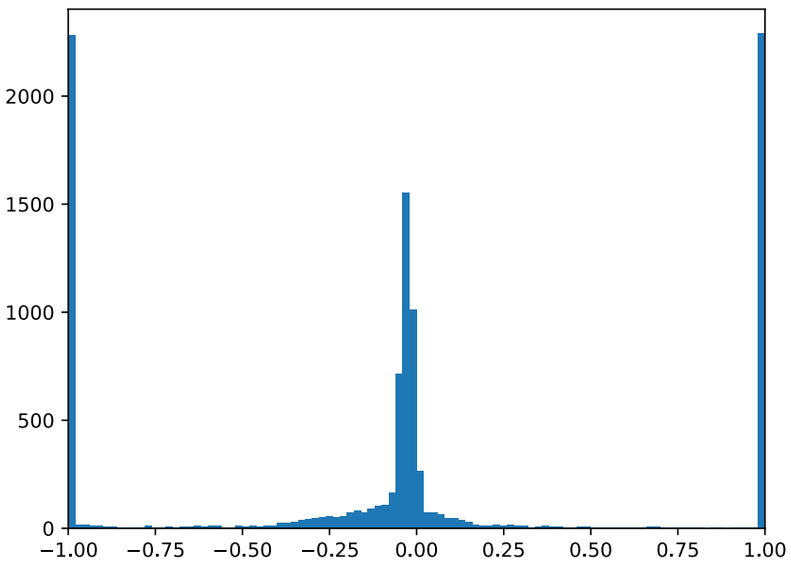
initial



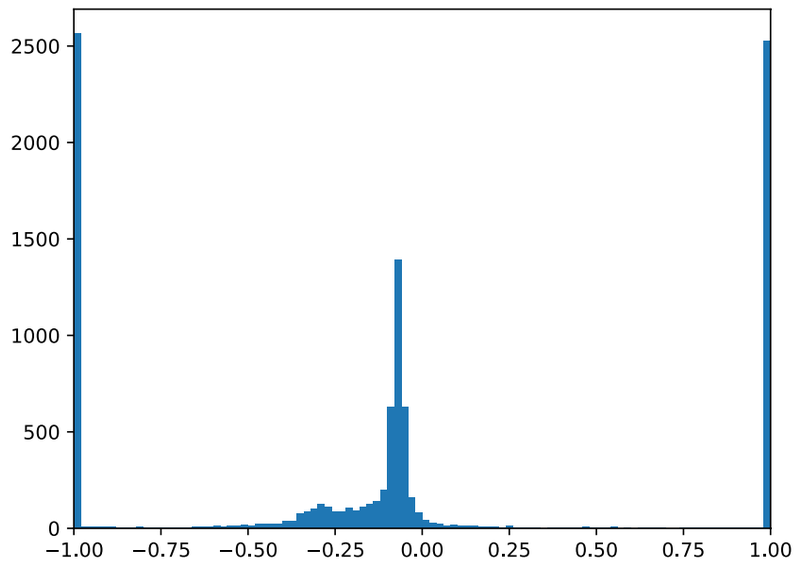
after 5 threads



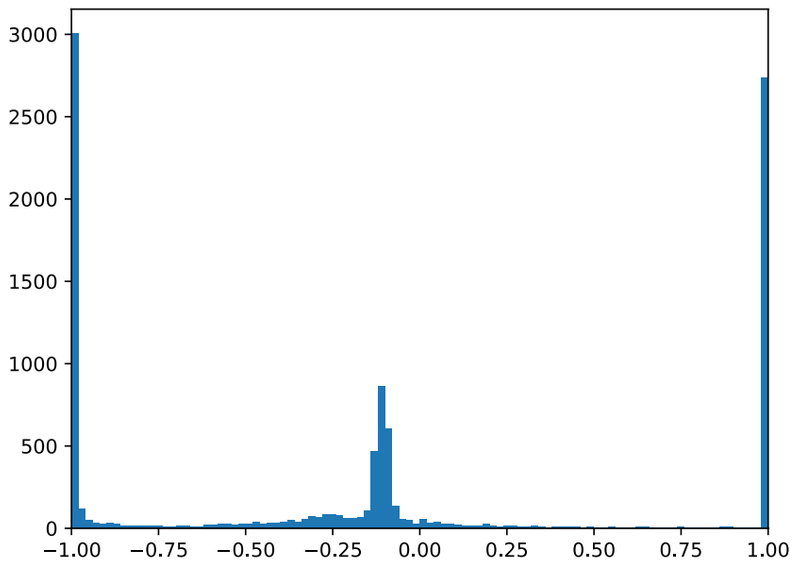
after 10 threads



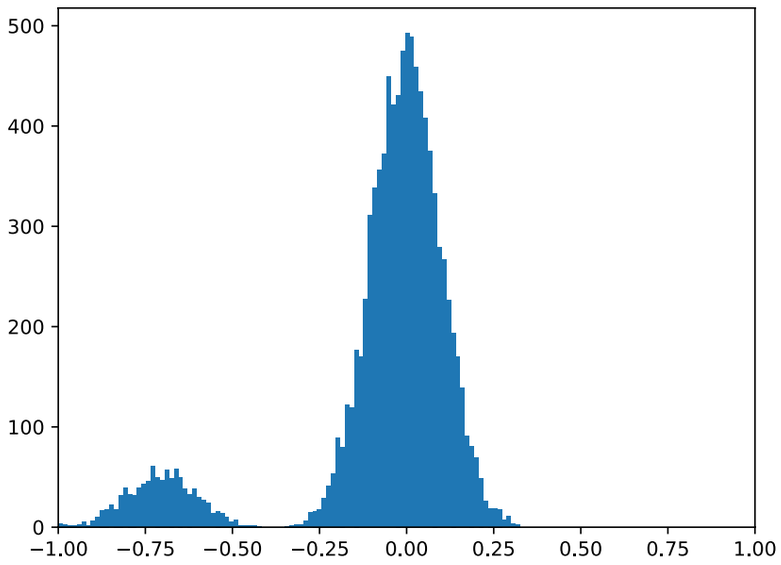
after 15 threads



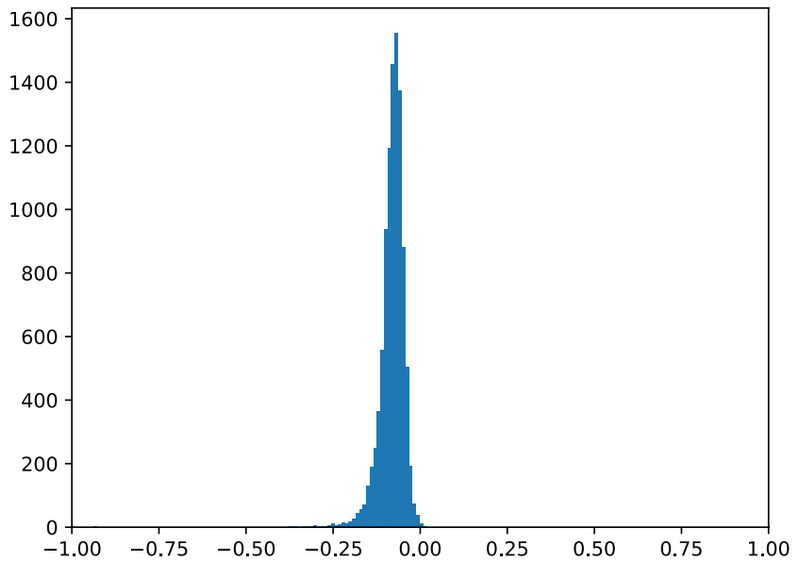
after 20 threads



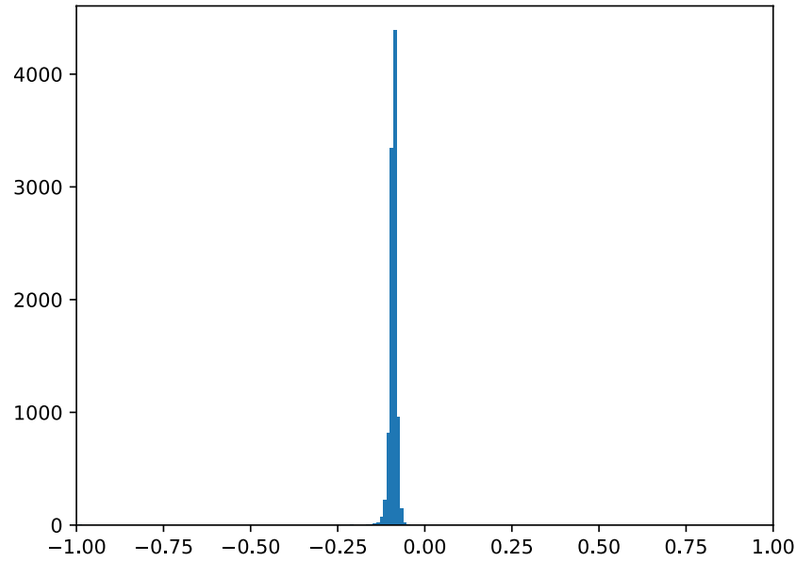
after 25 threads



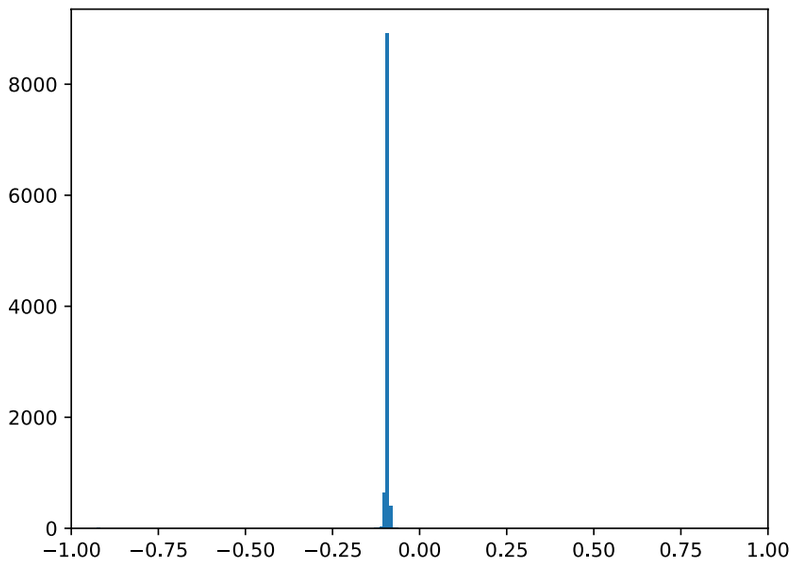
initial



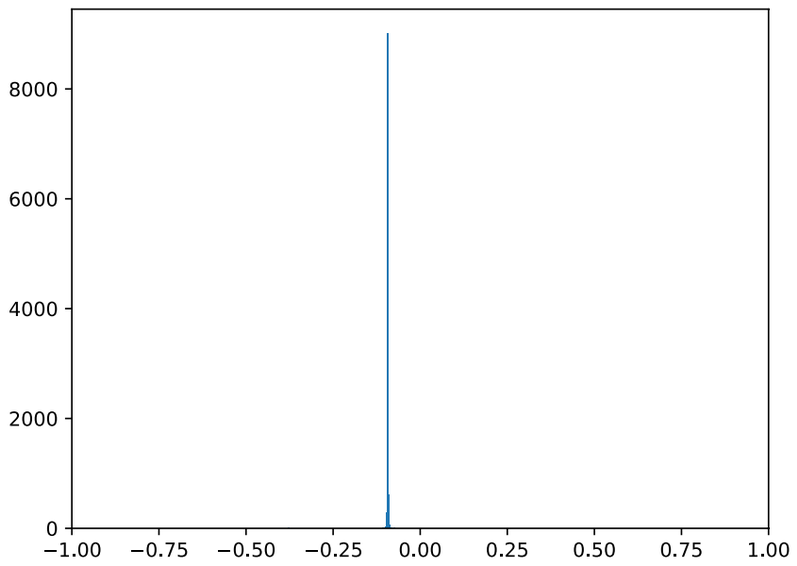
after 5 threads



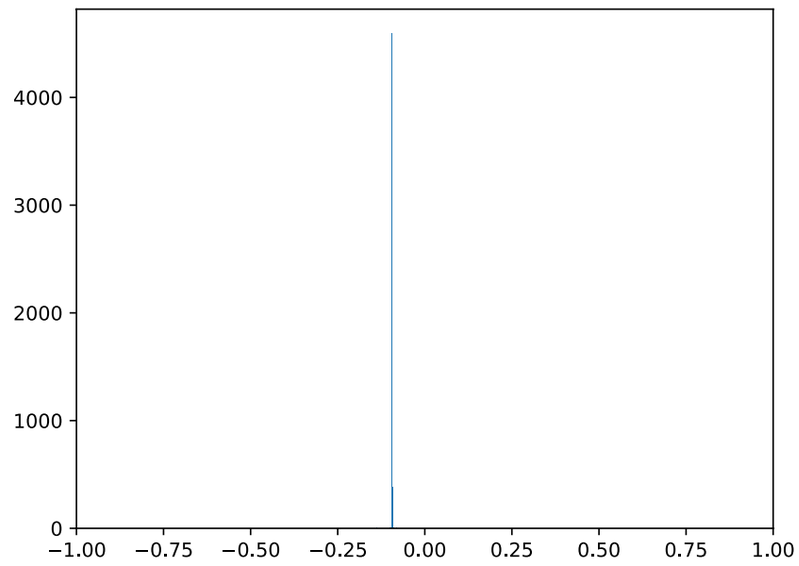
after 10 threads



after 15 threads



after 20 threads



after 25 threads

Other Modeling Options

Here are some more modeling suggestions:

You can try adding in the concept of objective truth by boosting the persuasiveness of arguments based on their closeness to the "true" opinion value.

See what happens when the initial population opinions are more uniformly distributed.

Experiment #3

Moderation

Adding Moderation

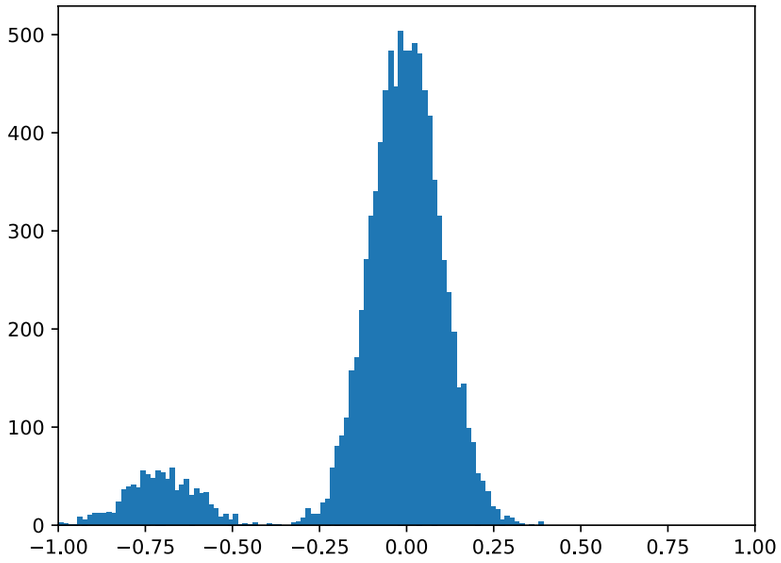
We have seen that when overall comment qualities are low, cognitive biases are able to kick in and polarize people into groups.

Even small groups of extremists are able to get traction in these kinds of environments.

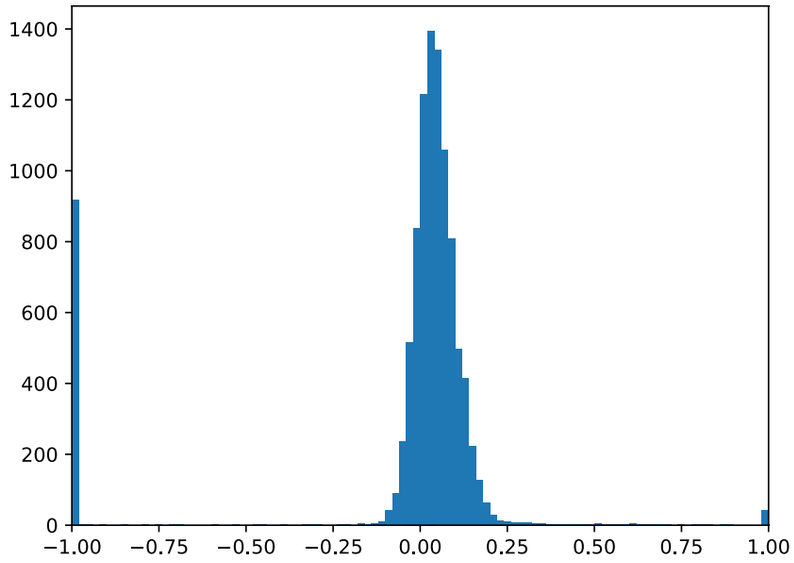
What happens if we remove comments with low persuasiveness?
These would include memes, insults, and other comments which have no informative value.

Let's remove all comments with persuasiveness below a certain threshold.
We'll use the population with a small group of extremists from before.

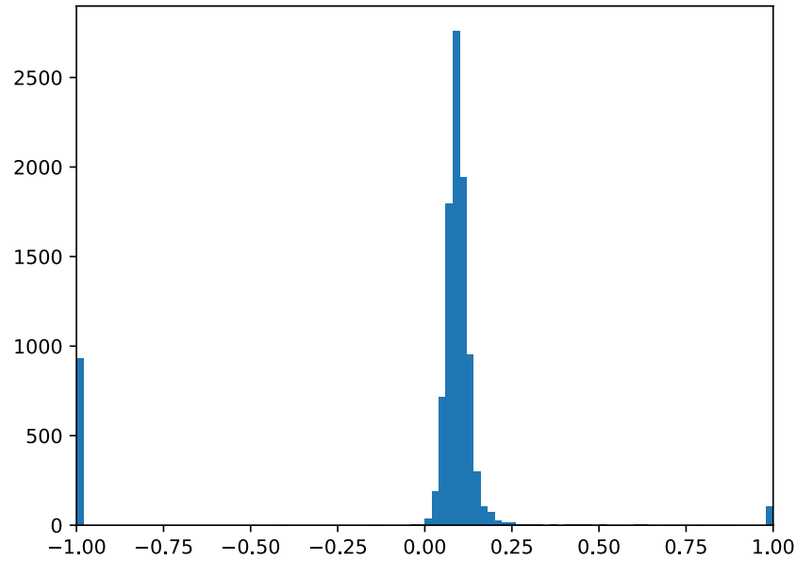
We'll look at the group of semi-reasonable commenters, and apply a comment persuasiveness floor of 0.8.



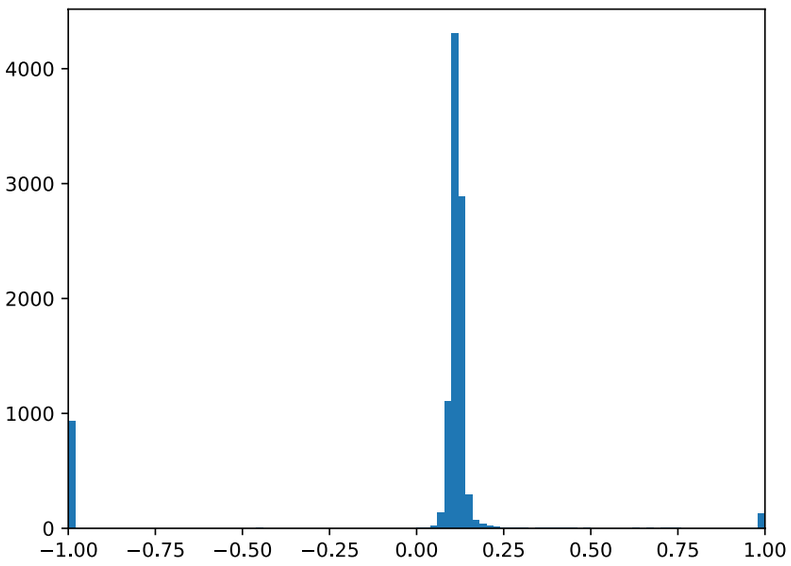
initial



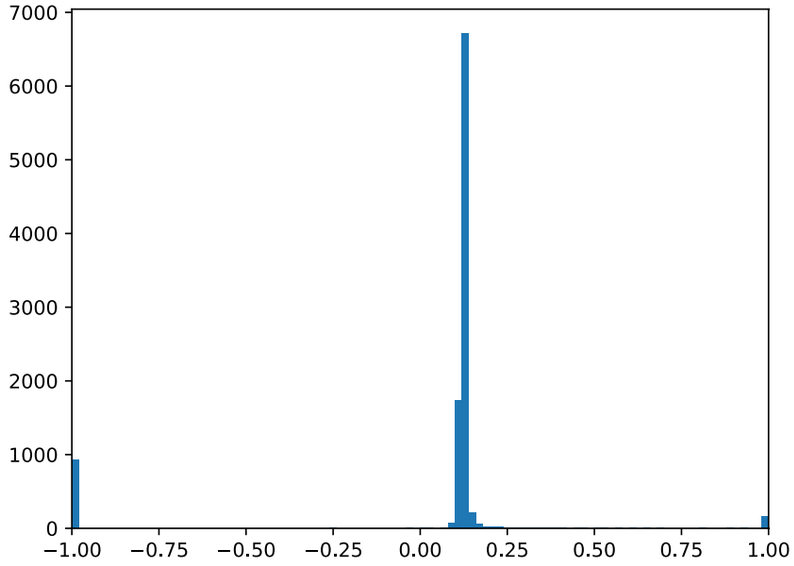
after 5 threads



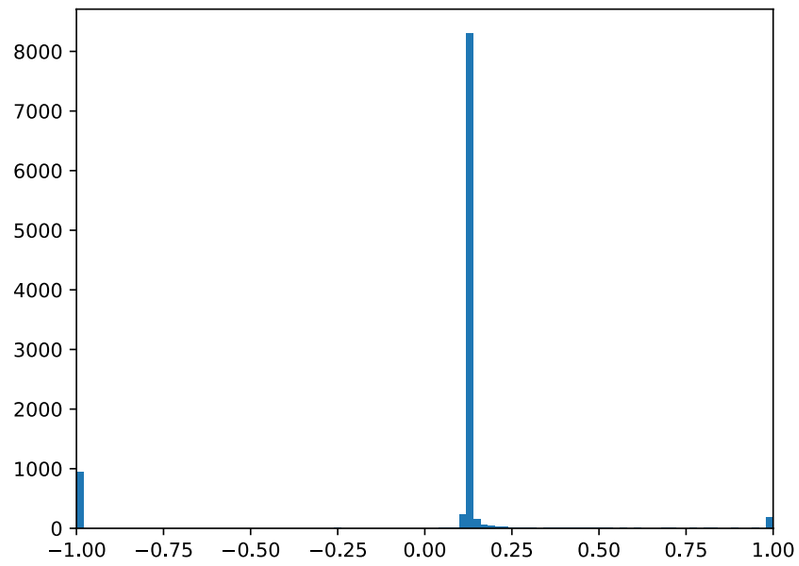
after 10 threads



after 15 threads



after 20 threads



after 25 threads

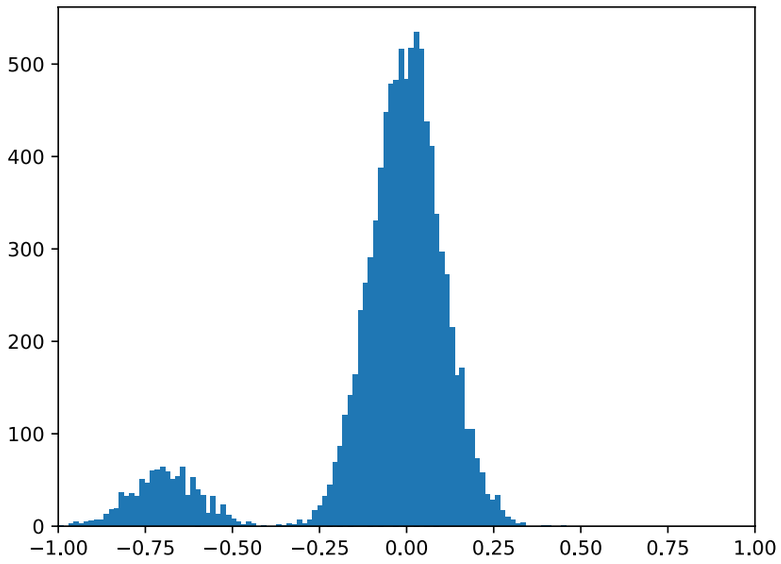
Biased Moderators

An objective moderator who removes low-quality comments can prevent polarization and extremism.

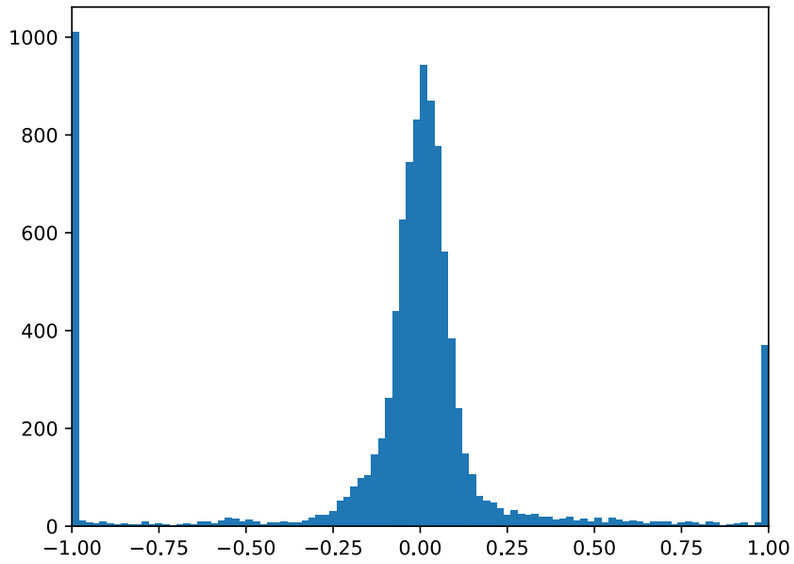
However, in real life, the individuals acting as moderators have biases as well. This means that low-quality comments which affirm their biases may be allowed through, and medium-quality comments with contrary opinions might be censored.

Using the last scenario, let's make one of the people with outlier opinions the mod for this sub. Only comments that they deem to be high-quality will be allowed through.

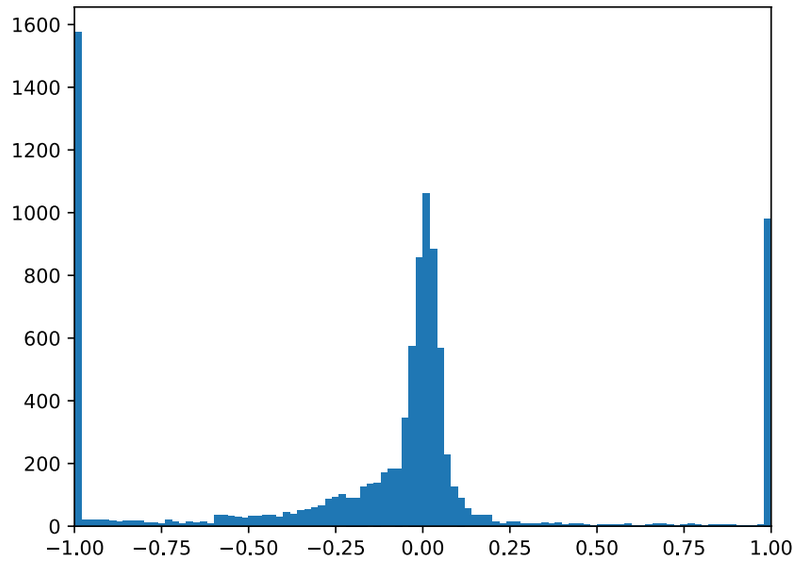
On the next slide, we see that having a biased moderator still results in polarization. In fact, the polarization is worse than in the case where there is no moderator. The difference is that most of the dissenting population's comments are now hidden.



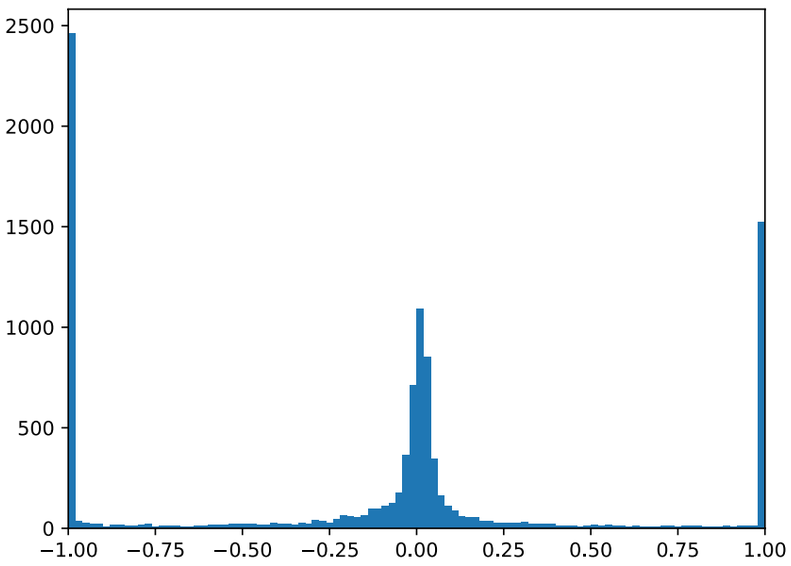
initial



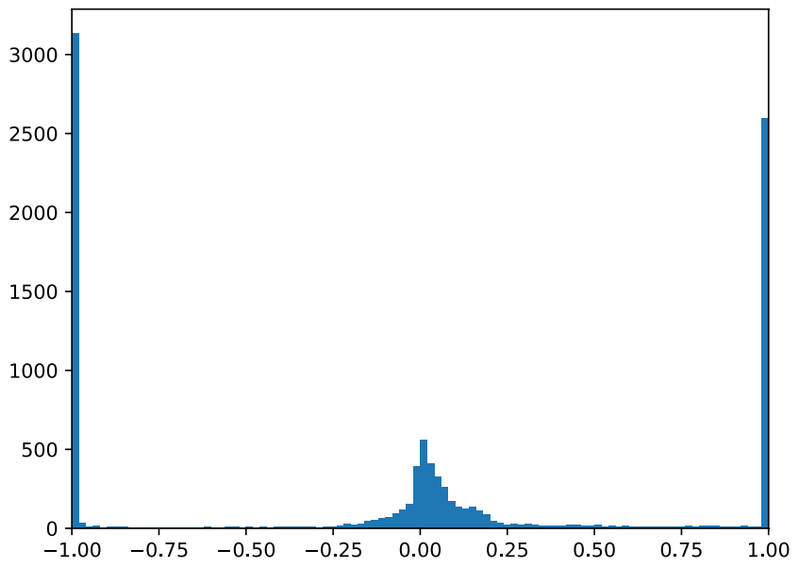
after 5 threads



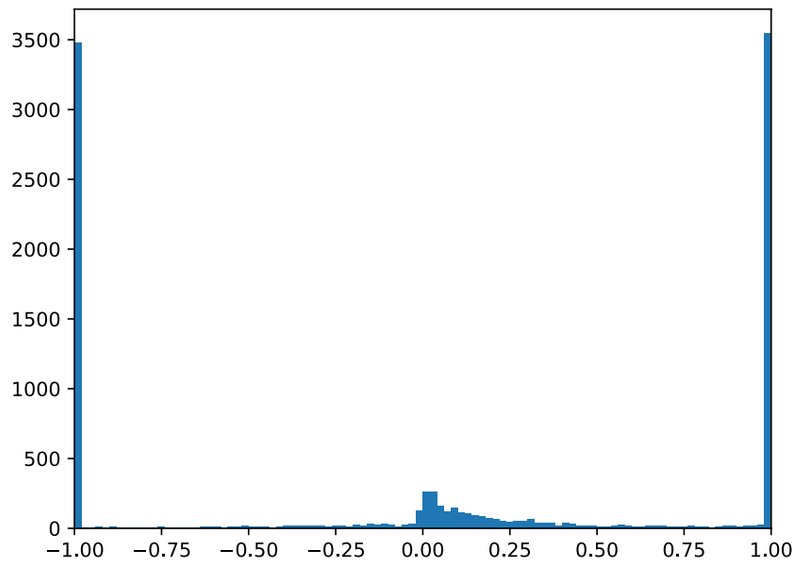
after 10 threads



after 15 threads



after 20 threads



after 25 threads

Other Modeling Options

Here are some more modeling suggestions:

Try adding in multiple moderators. Note that doing this will probably involve refactoring the code.

See what happens when the moderation is intermittent instead of being omnipresent.

Experiment #4

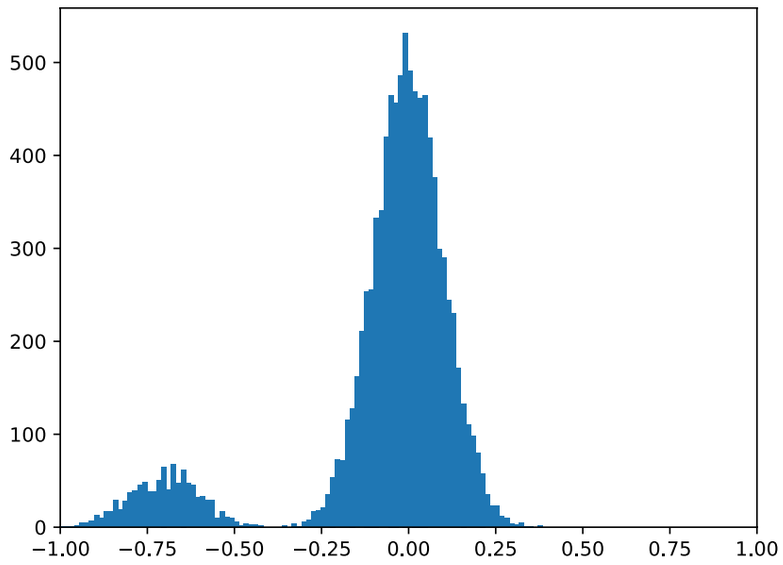
Comment Ordering

Random Comment Ordering

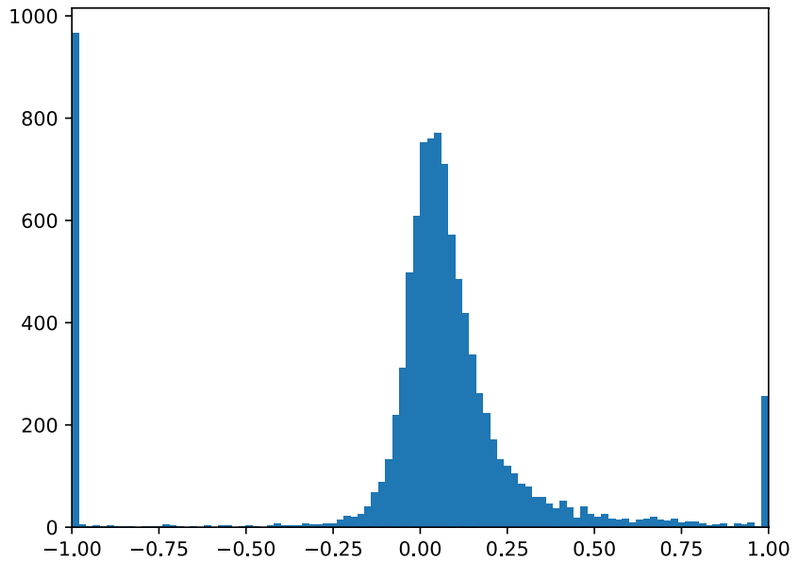
We have seen that in a forum-type commenting system, if the overall comment qualities are low enough then the population can become more polarized just by talking to each other. Moderation by a human actually makes the polarization worse.

What if we tried something different? It's well known that in comment threads the first couple of commenters can drastically change the bias of the thread. Whichever comments come earliest usually end up getting voted to the top, where they are seen by more people.

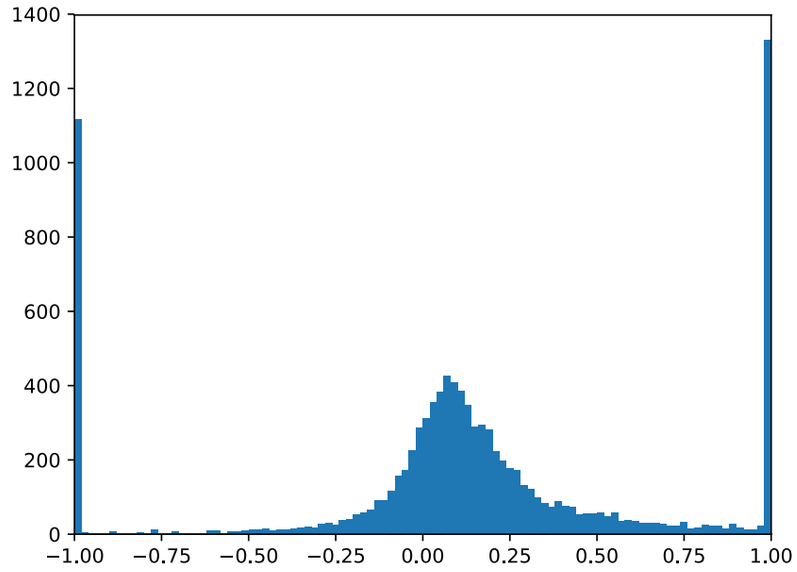
Can changing the order in which comments are displayed have an effect on polarization? First, let's check random ordering. In this scheme, we list in random order the comments at each level.



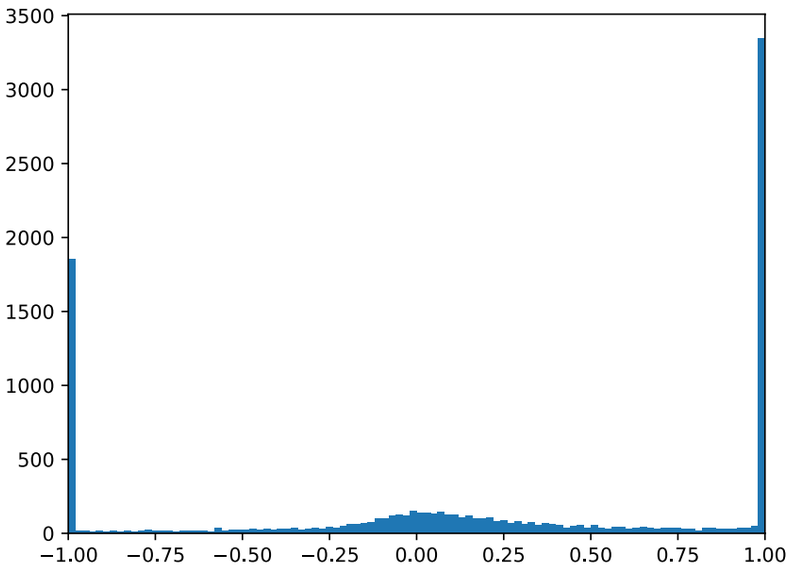
initial



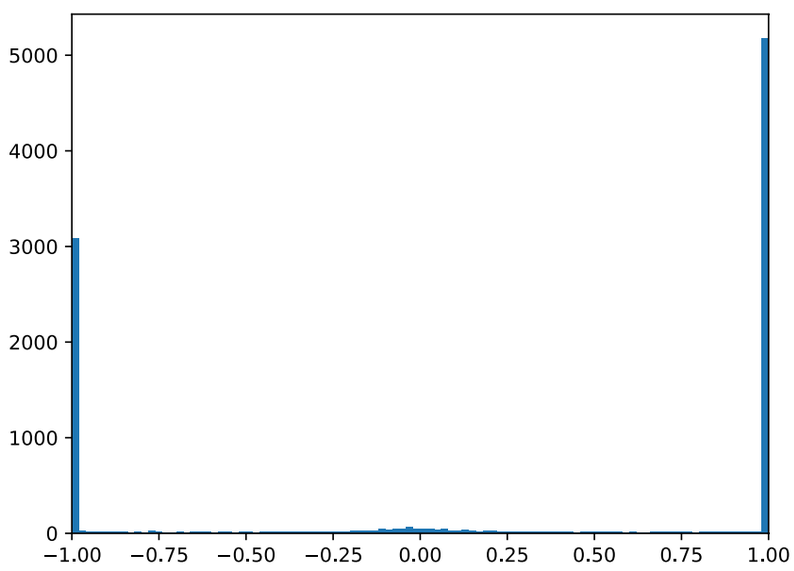
after 5 threads



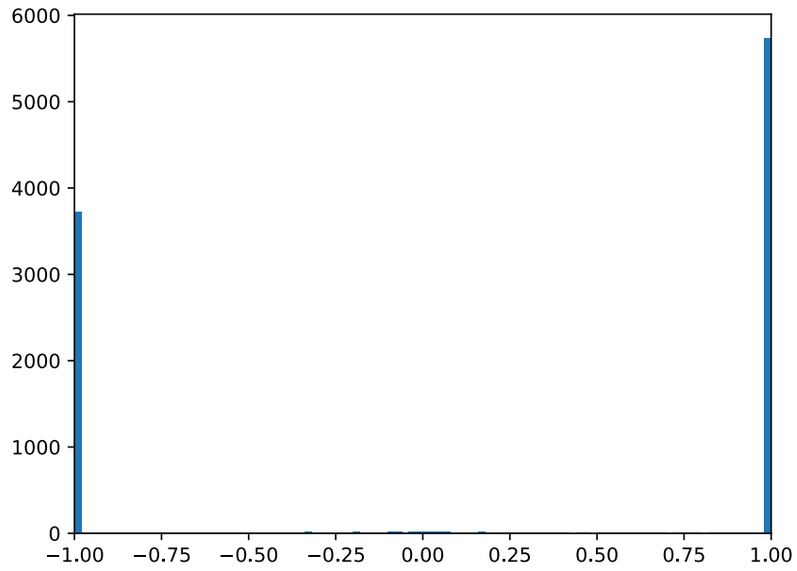
after 10 threads



after 15 threads



after 20 threads



after 25 threads

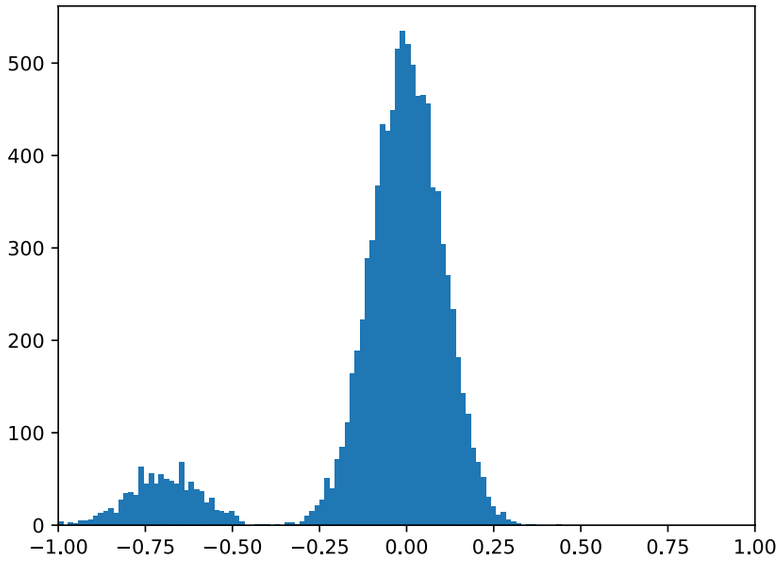
Similar Comment Ordering

It makes sense that random ordering doesn't do much. In a situation where we have a cognitively-biased population sharing mostly low-quality comments, it doesn't really matter whether an individual sees a comment that agrees or disagrees with them. The net effect is still overall towards polarization.

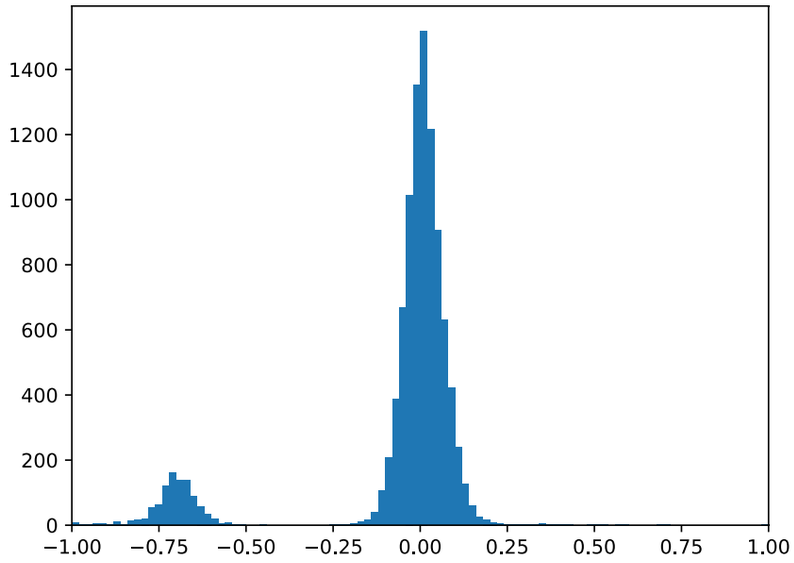
After all, it is common for members of one echo chamber to visit another an opposing echo chamber and come back even stronger in their original beliefs. When differences in belief are large and persuasiveness of comments low, seeing opposing views is rarely helpful.

However, what if we preferentially showed views which were similar, but not too similar? For example, it is easy to find clustering in social media data, among individuals who upvote each other's posts and comments. If we took someone who was 95% associated with group A and showed them comments that were only 90% approved by group A, could this help fight polarization?

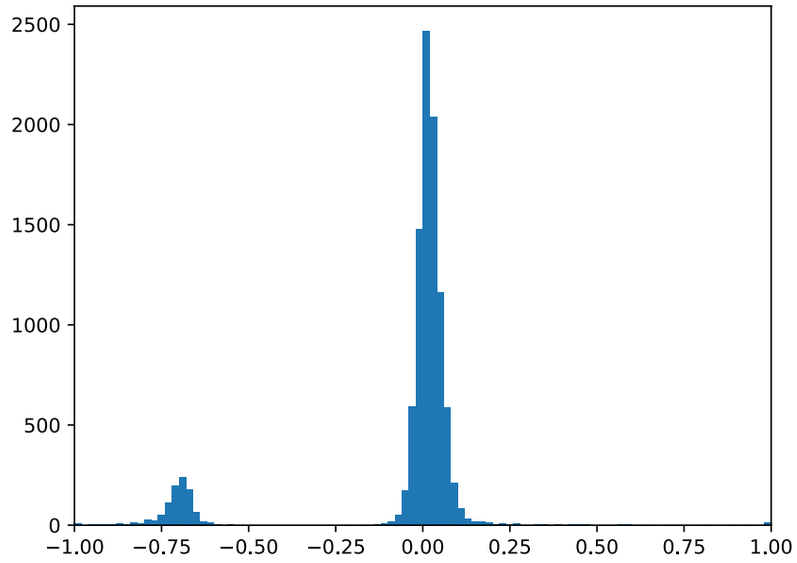
In this next scenario, we will preferentially show comments which are about 0.1 away from each individual's own opinion score, sampling using a Boltzmann distribution.



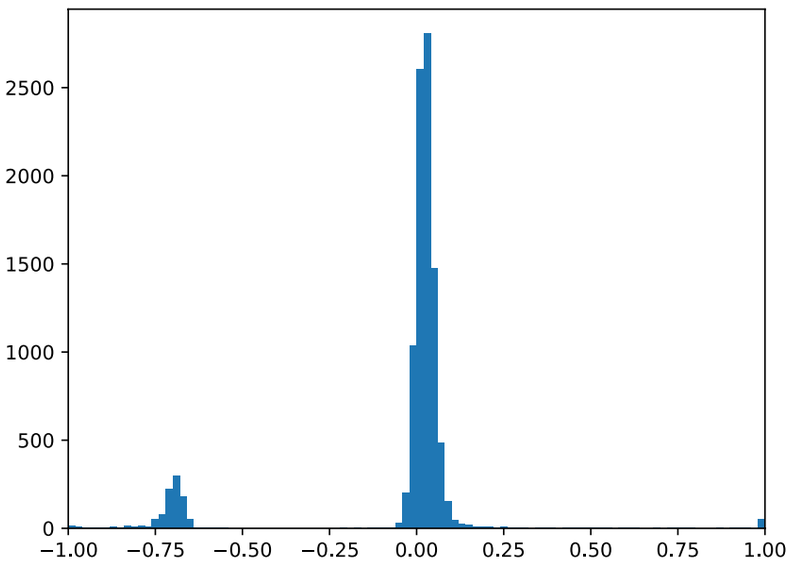
initial



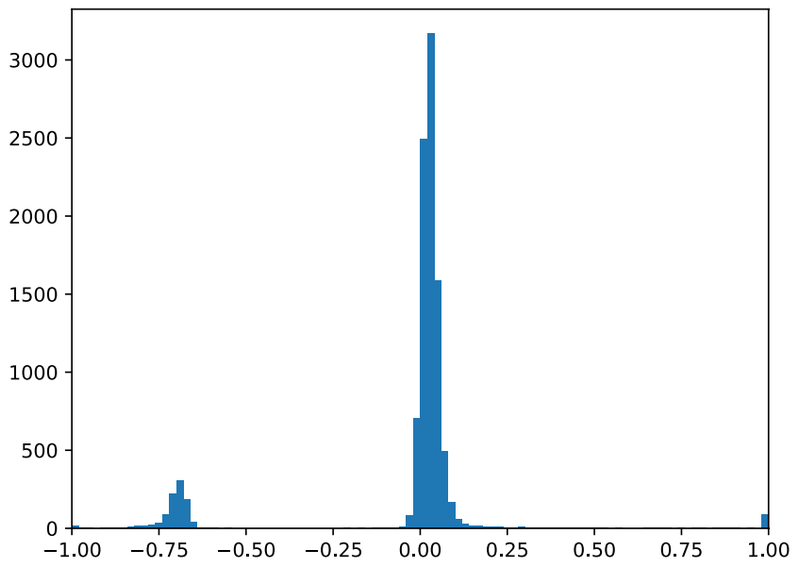
after 5 threads



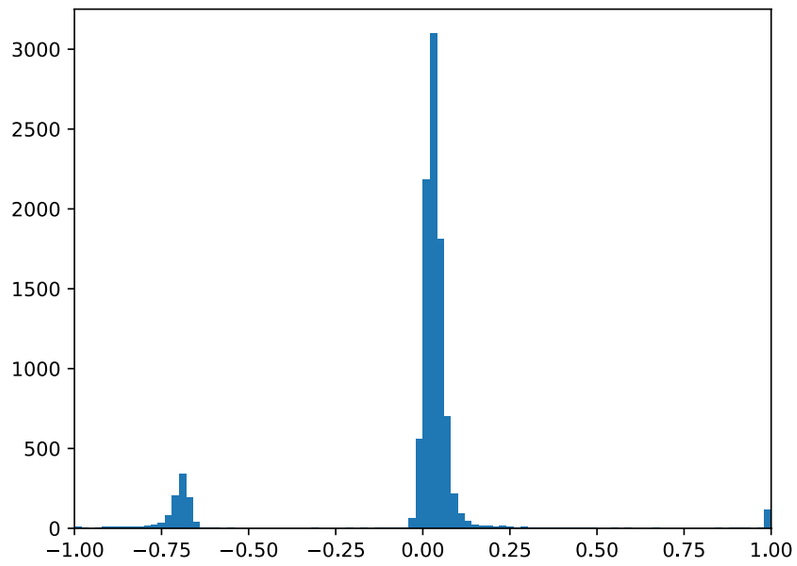
after 10 threads



after 15 threads



after 20 threads



after 25 threads

Wrapping Up

Conclusions

In this project we examined what happens when people talk to each other in a system where there is only one topic of discussion, and in which no new information enters the system.

We see that if comment quality is low (rude, uninformative, or otherwise unlikely to persuade), then polarization quickly occurs. People form into extreme opinion camps.

Furthermore, if there are moderators, given the inevitability that the mods have personal biases, we see an even greater drive towards polarization, albeit one in which a subset of the population gets preferentially silenced.

Finally, we showed that a system in which similar-but-not-identical comments are shown to a user, polarization does not occur. This may indicate that preferentially showing a range of opinions within an individual's comfort range can offset the effects of echo chambers.

In practice, this can be implemented with a small tweak to existing search algorithms. We can use the same clustering methods, but instead of returning items which most match a user's preferences, we can return results that are ideally just a little bit different.

Other Modeling Options

Here are some more modeling suggestions:

What happens if within the range of comfortable opinions, we rank those higher which are closer to the overall median population opinion?

What happens once we add in the concept of objective truth? Does this comment ordering system help to move the consensus to that point?

Bibliography

[1] AllSides. "Political Polarization in America, in Two Fascinating Charts". https://www.allsides.com/blog/political-polarization-america-two-fascinating-charts.

[2] Wait But Why. "The Story of Us". https://waitbutwhy.com/2019/08/story-of-us.html.

[3] Hafzh A. Prasetya and Tsuyoshi Murata. "A model of opinion and propagation structure polarization in social media". January 09, 2020. https://link.springer.com/content/pdf/10.1186/s40649-019-0076-z.pdf.