Birdwatch and the Polarization of the Crowds

Ryan Champaigne

28 April 2022

Introduction

Any social media platform with a large user base has to balance opposing pressures when deciding how to enforce their community guidelines (Gorwa *et al.* (2020)). Importantly, platforms must decide what posts they will not allow, and how to go about identifying them. Due to the sheer volume of content posted to social media sites daily, this problem can be monumental. Many companies have turned to advanced algorithms and artificial intelligence to flag problematic content, which can be very efficient in processing large amounts of data, but often leads to crude enforcement due to computers' relative inability to detect subtleties of speech ("How social media firms moderate their content," s.d.). Despite the complexity, government and consumer pressure has been mounting for companies to reign in what is considered dangerous speech on their platforms ("Do digital echo chambers exist?" 2019). As a possible solution, Twitter and Facebook as well as other platforms are testing out crowd-sourced content moderation as a means of accurately categorizing large volumes of posts.

Collaborative Content Moderation

Wikipedia is often cited by proponents as a positive model of the power of crowds to produce accurate and reliable information(Yasseri & Menczer (2022),Liu & Ram (2011)). People have always been skeptical that Wikipedia's model, which allows anybody to edit and contribute information to their site, could produce an online encyclopedia that is as reliable as traditional encyclopedias. Certainly, there are examples of pages that are poorly written or contain false information, however, research has shown that Wikipedia articles are generally very factually accurate, and resilient to contributions of false or unverifiable claims(Yasseri & Menczer (2022)).

There has been much debate about how Wikipedia is able to maintain its relatively high accuracy while also allowing anyone to contribute anything. The wisdom of crowds is a phenomenon that often the combined judgment of a large group of people can give better results than any one person could have given. Lui argues that the process of reaching consensus between a wide variety of editors utilizes this wisdom of the crowds to achieve a result that is better than any one contributor could have made (Liu & Ram (2011)). This outcome does not arise automatically, though, and is instead a product of the process of collaboration and consensus reaching in which Wikipedia editors engage. The fact that anyone can contribute also implies that anything posted to wikipedia can be edited by others. A consensus must be reached between editors if they disagree, and this process contributes to reliable and relatively bias free information(Guilbeault *et al.* (2018)).

Whether or not crowd-sourcing can be used by social media platforms to effectively parse through large amounts of content and prevent the spread of false or misleading information is still an open question. Major companies like Twitter and Facebook have already rolled out crowd-source style content moderation in some form, and we are beginning to see the preliminary results on its efficacy(Pröllochs (2021)).

Birdwatch

Twitter has launched Birdwatch, a pilot test for a crowd-sourced content moderation on their platform. Users can sign up to be a part of the test program, which works on two levels. First, users can write notes on tweets that they believe are misleading, dangerous, or need more context. Creating the note involves answering multiple choice questions classifying why they believe the content is problematic, as well as writing text that could be displayed alongside the tweet. It is also possible to write notes that classify a tweet as not misleading or problematic, though these notes will not be displayed on tweets. Second, after notes have been written, users can rate the notes posted by other users in terms of whether or not they believe the note is helpful. Notes that receive sufficient positive reviews will eventually be displayed publicly on tweets.

Twitter also cites that in addition to a high rating, notes must receive a high rating from a diverse base of users, although it is unclear exactly how this will be implemented. Birdwatch is still in its pilot phase, which means that no notes are ever displayed on tweets, but users can still generate and rate notes.

Another key tenet of Birdwatch is transparency. Currently, all data containing user generated notes and ratings is updated daily and available for download on Birdwatch's website. Algorithms decide which note will eventually be displayed on the tweet, and twitter encourages active community participation both in developing Birdwatch and analyzing and critiquing its results.

Hypothesis 1: Since many tweets are political in nature, I predict there will be distinct groups of Birdwatch users that rate tweets in alignment with their political ideology. That is, two users from one ideological group will tend to have the same opinions on whether a note is helpful or not helpful, while users in different ideological groups will not have similar helpfulness ratings of notes.

If there is significant polarization between users, I believe that it will be difficult to generate consensus between two groups, rather than simply promote the note that gets the most support from its group. While the goal of Birdwatch is to foster collaboration that can combat misleading content, it could become another ideological battle ground.

Hypothesis 2: Polarization of users will differ according to the content of the tweet. Issues that generate strong debate–and subsequently lots of tweets–tend to be polarizing. I hypothesize that for some issues which are very polarizing we will see a corresponding polarization in Bridwatch user ratings, while for content that is more apolitical, we will see less polarization. While less testable, I also hypothesize that tweets that are flagged will be disproportionately related to topics which are highly polarized.

Hypothesis 3: There will not be significant polarization in the tweets or ratings that users write and rate notes for. While the idea of social media echo chambers has been proposed as a driving factor for increased polarization in public discourse, recent research has shown that their effect may be overstated ("Do digital echo chambers exist?" 2019). What is more, the Birdwatch platform makes specific recommendations to ensure that notes are rated by a diverse set of users. If there was polarization in the content that users were rating, we would see distinct clusters when mapping which notes users rate regardless of if they agree or disagree.

Data

The Birdwatch data comes in two data sets, one containing information on notes, and one containing information on ratings. The data set I obtained contains all Birdwatch data from its launch up until April 1, 2022. In total, there are 8,134 users, 219,395 ratings, 20,221 notes, with 14,187 tweets reviewed. I combined the datasets according to the note_id which produced a dataset containing all information on the ratings of notes alongside the note data itself.

To establish whether a user was generally in favor of a note or not, I used the helpfulness rating. When users rate a note they are asked if the note is "Helpful", "Somewhat Helpful", or "Not Helpful". In the older data, users were not given the "Somewhat Helpful" option, and the helpfulness was recorded with two binary variables, "Helpful" = 0 or 1, "Not Helpful" = 0 or 1. To combine the old data and the new data, I simply recorded the new data in the old binary format, and recorded "Somewhat Helpful" as a 0 for both.

Birdwatch does not release any information on the tweet itself other than the tweetId, a unique id number assigned to all twitter content. I used the rtweet package in r to retrieve the tweet data from twitter's API. I then joined the tweet's text to the birdwatch notes dataset to have both the note and the tweet it refers to.

Methods and Results:

In order to map polarization in the Birdwatch data, I used gephi and Igraph to plot weighted undirected graphs related to various connections between users. For all graphs, I used the Fruchterman Reingold layout, which models edges as springs in order to orient vertices(Hansen *et al.* (2020)). For network analysis, there are two important subsets of a graph–vertices and edges. Because I am looking to model relationships between users, it is natural to assign the vertices to be individual users. An important question in this research, though, is how to connect users based on their activity on the Birdwatch site.

For my first analysis, I created a bipartite graph, where both users and notes are vertices, and there is an edge that connects each user to the notes that they rated as helpful. Importantly, this graph has two classes of vertices, users and notes, and every edge connects a vertex in one class to a vertex in the other class. Because of this property, it is possible to condense the graph into a new graph containing only user vertices, where an edge between two vertices represents that two users both thought the same note was helpful. Additionally, the weight of the edge is the number of such notes on which they agreed (i.e., if two users both rated the same 3 notes as helpful, the edge weight between them would be 3).

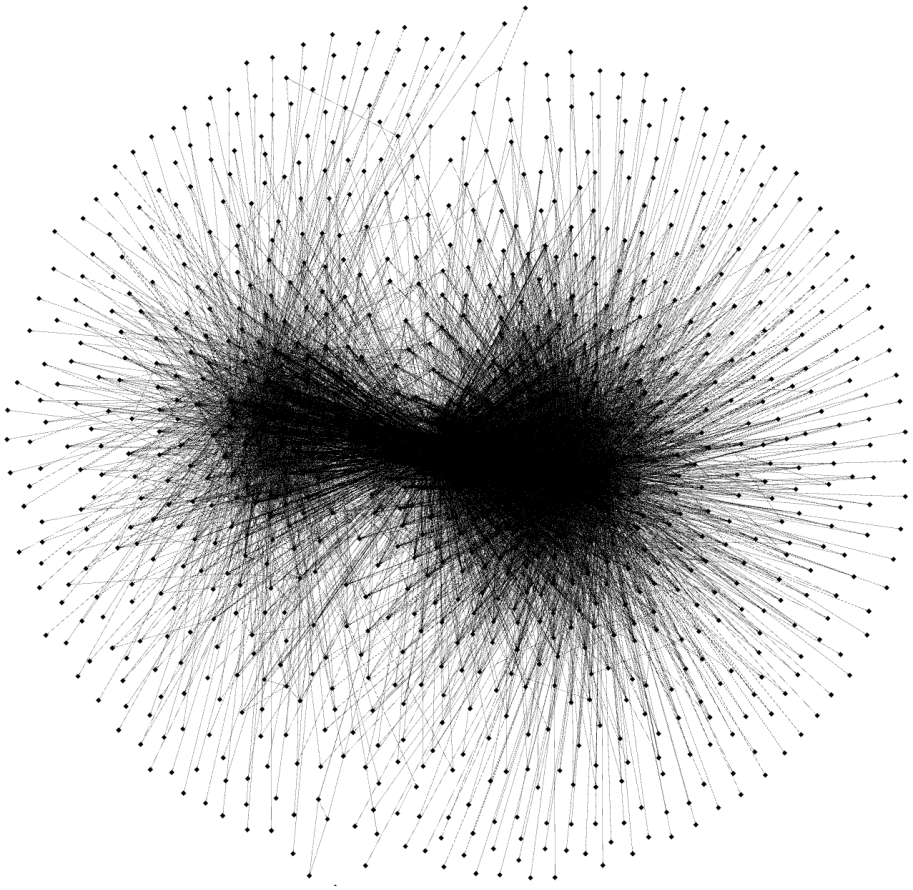


Fig 1. Bipartite Condensed graph, vertices connected based on agreement about the same notes

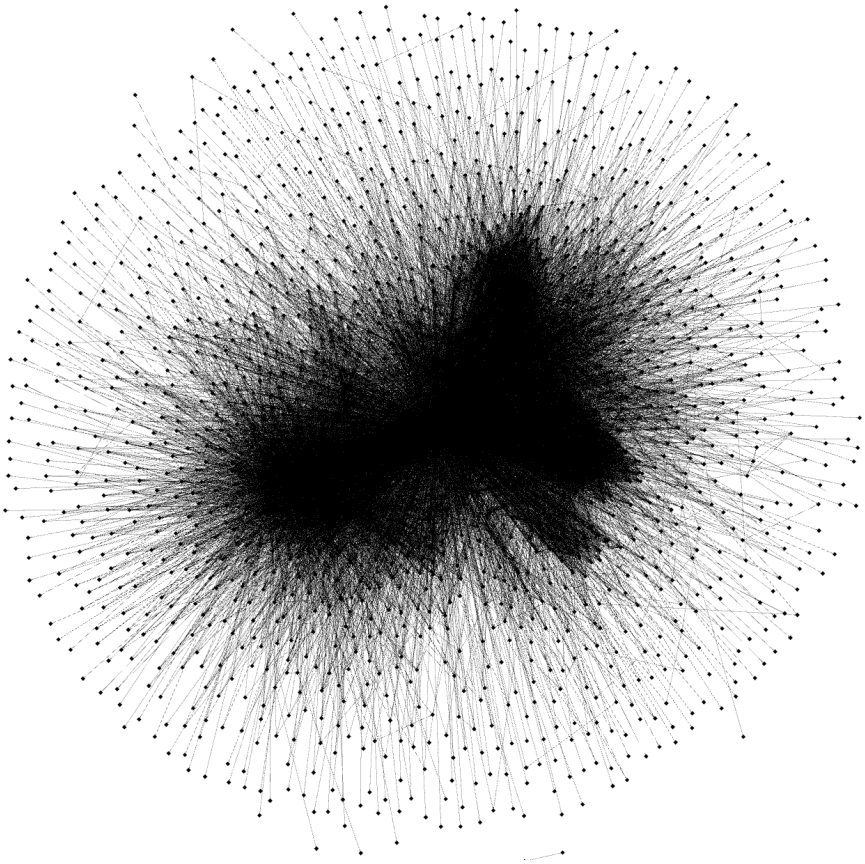


Fig 2. Bipartite condensed graph, vertices connected based on disagreement about the same notes

From the figures, we get a clear sense that there is clustering in the data. Users tended to form groups based on agreement with each other. In the next graph, users are similarly connected based on disagreement between the same notes. To make the graphs easier to understand, only edges with weight higher than 10 and higher than 5 respectively are plotted.

From exploring the graphs however, it was evident in both that there were a select few users that were connected to a large portion of all other vertices. Indeed, the more ratings a user makes, the more likely they are to be connected to other users, even if their ideologies do not align perfectly (i.e., if one user agrees with another user on 10 tweets but disagrees on 10 other tweets, they will still be connected in the agreement graph, despite their overall dissimilarity in note ratings).

To create a better measure of overall agreement between users, it would be beneficial to base edges on average note agreement rather than total instances of agreement. First, I created a new measure of helpfulness equal to 1 if a user rated a note as "Helpful", -1 if a user rated a note as "Unhelpful", and 0 if the user rated the note as "Somewhat Helpful" or did not rate the note. Next I created a large matrix with entries (i , j) equal to the new helpful measure of the j-th note by the i-th user. From this matrix, I created a vector for each user with entries v-j equal to that user's helpfulness index of the j-th note. With these vectors, for any two users, we can calculate the dot product of their vectors, which gives a new measure of connectedness, what I will call the dot-product-weight. One can see that if two users both find a note helpful, this contributes 1x1 = 1 to their dot-product-weight. If one user finds a note helpful while the other finds the note unhelpful, this contributes -1x1 = -1 to the dot-product-weight. Similarly, if two users both find a note unhelpful, this contributes -1x-1 = 1 to their dot-product weight.

This new measure has the property that large positive values reflect that two users tend to agree, large negative values represent that two users tend to disagree, and values close to 0 reflect that users agree and disagree at a similar rate.

Finally, I created an adjacency matrix with edge weights equal to the dot-product-weight between users. Since edges cannot have negative weight, I filtered out negative values.

Due to the size of the matrix needed to create this measure I reduced the size of the data set in two separate ways. For the first graph, I plotted only users who have rated 150 or more notes. This graph represents the connectedness of the most active birdwatch users.

For the second graph, I took a random sample of one third of all users to get an idea of the connectedness of the average birdwatch user.

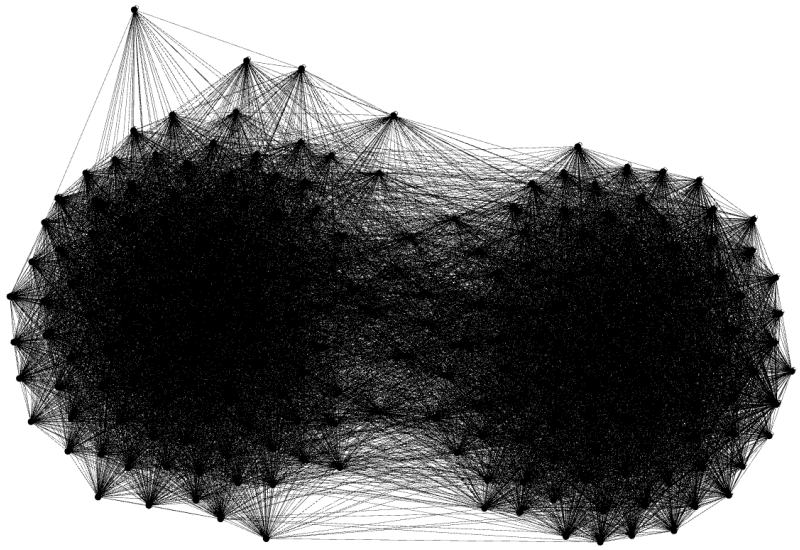


Fig 1: Most active birdwatch users, graphed by dot-product-weight

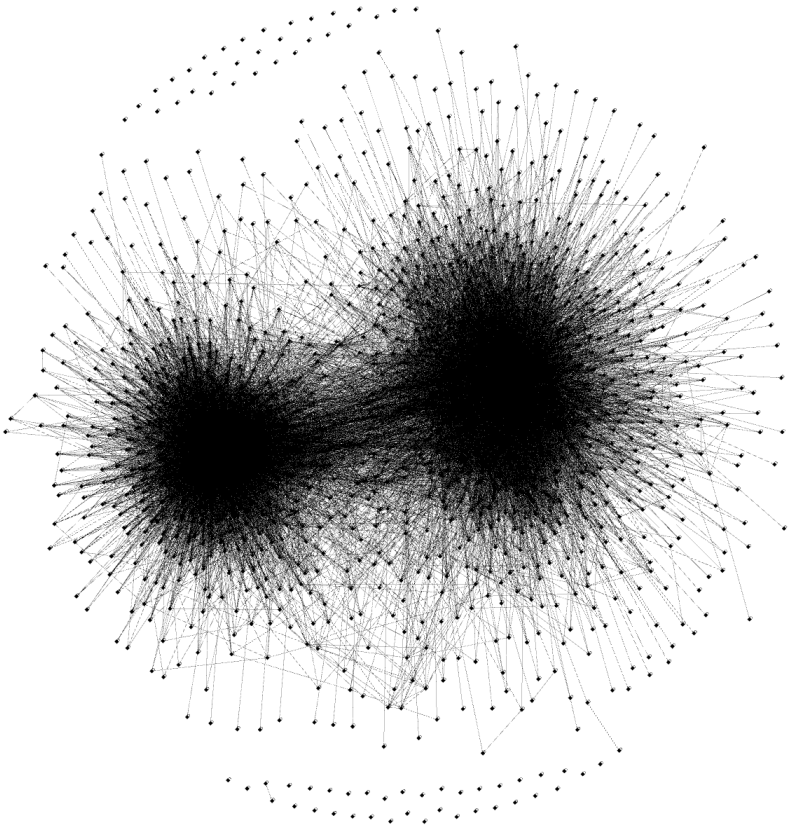
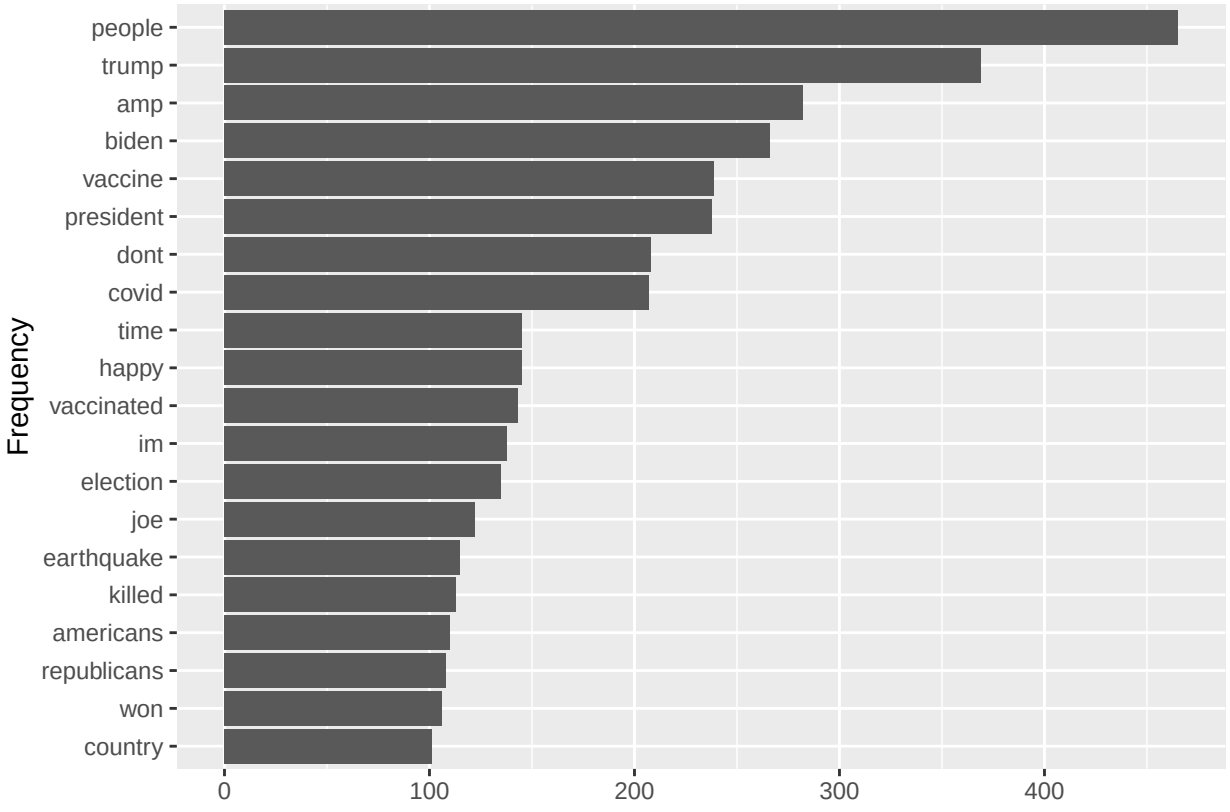
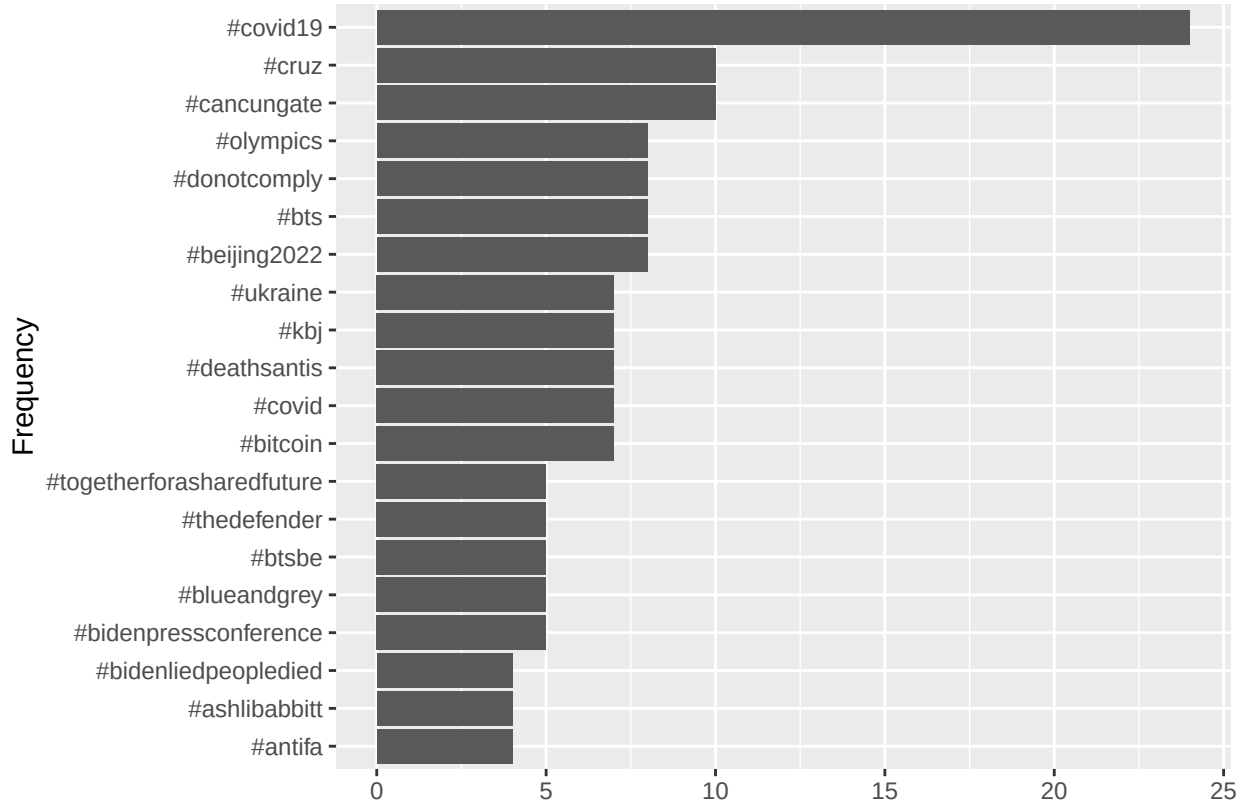


Fig 2: Random sample of birdwatch users graphed by dot-product-weight

As can be seen, both graphs demonstrate large polarization between users. The most active users form an incredibly clear bi-modal network while the average user network forms a similarly bi-modal network but which is more sparse.

Topic Analysis: From the tweets gathered from the twitter API, I counted the occurrence of each word and each hashtag in each tweet. Then after removing words that contain little to no information about the content of the tweet itself (referred to as stop words), I created a ranked list of the most common words.


people
trump
amp
biden
vaccine
president
dont
covid
time
happy
vaccinated
im
election
joe
earthquake
killed
americans
republicans
won
country
Frequency
0
100
200
300
400



With a few exceptions, the most common words and hashtags are political in nature. One of these exceptions is certainly "earthquake" however, after looking into these tweets I found that the large majority of notes relating to "earthquake" were created by one user, who has repeatedly flagged the same twitter account for posting unsubstantiated earthquake predictions. For the other common words that can be considered topics, they clearly are mostly political and consequently polarizing in nature.

Networks by topic:

In order to find if birdwatch users were more polarized on some issues than others, I created graphs that represent only agreement connections on tweets that contain certain words.

For the first graph, I filtered out by tweets that contain "biden", "trump, or "president" and for the second graph I filtered out by tweets that contain "bitcoin", "crypto", "olympics", or "beijing". While it is not certain from the words chosen that tweets containing these words would or would not be polarizing, I chose the words that I felt represented the respective categories best (indeed, I would argue that all of the topics represented in the most common words can be considered polarizing).

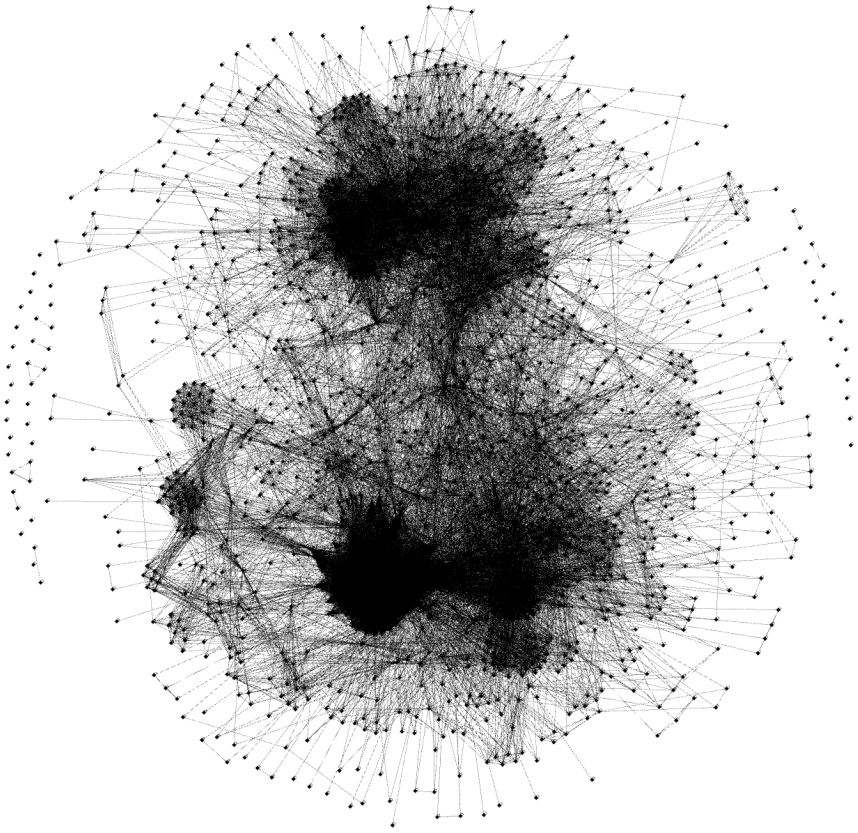


Fig 5: Connectivity based on biden/trump/president tweets

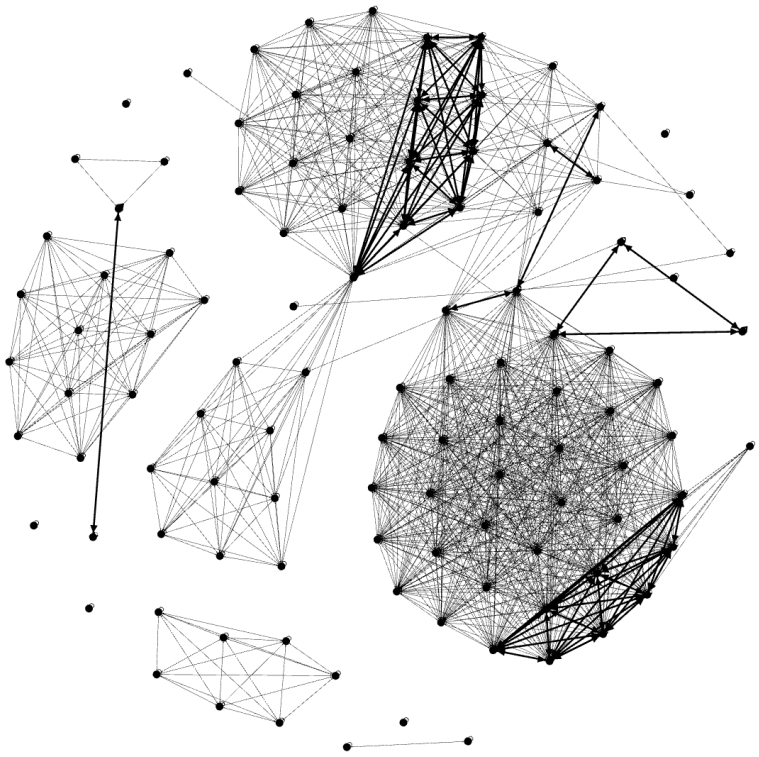


Fig 6: Connectivity based on bitcoin/crypto/olympics/beijing tweets

While some polarization is clear in the first graph, polarization in the second graph is certainly less prominent, although both graphs clearly reflect an underlying structure of the tweets themselves, due to the low number of tweets in each category.

Finally, to test the third hypothesis, I graphed users based on which notes they rated, regardless of whether or not they rated the notes helpful or not. If there exists an echo chamber like effect which affects which tweets and notes users see we would expect to see a similar bi-modal structure in this graph.

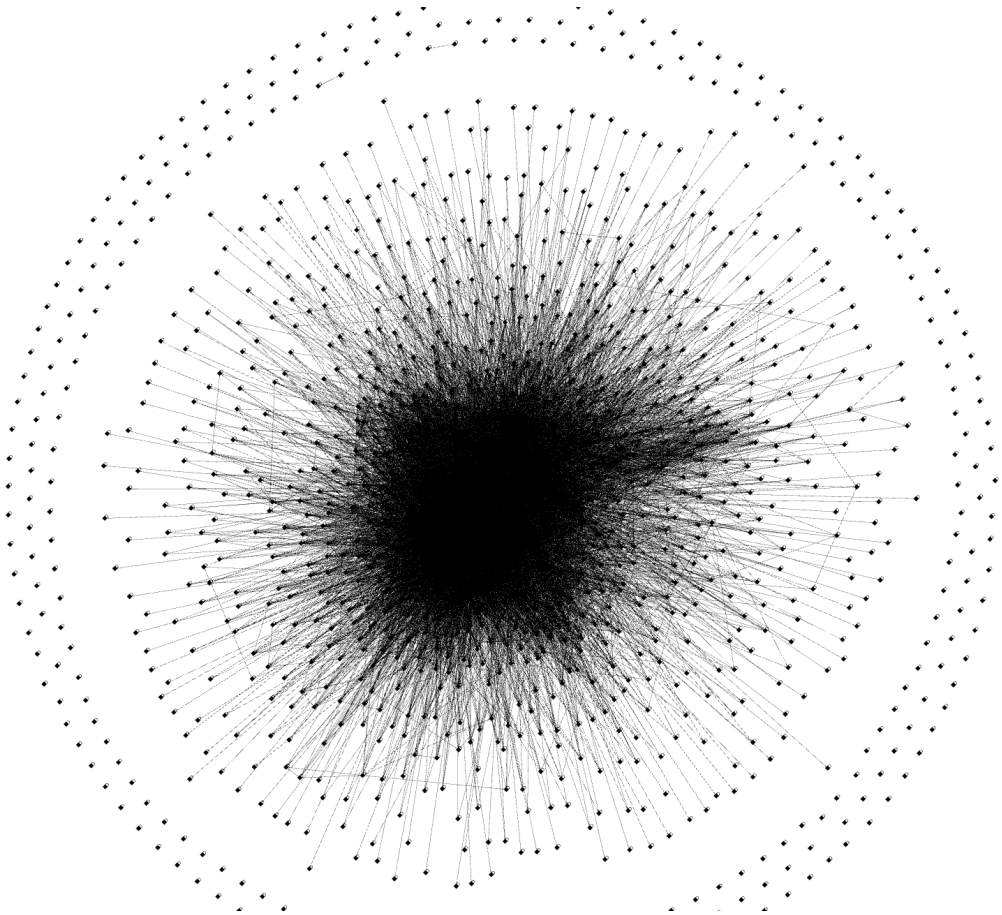


Fig 7: Graph with edges users connected who rate similar notes

It can be seen that the graph is more or less uni-modal, representing the fact that users across birdwatch do not tend to form clusters based on which notes they review.

Discussion

Twitter's Birdwatch data reveals that significant polarization is present between Birdwatch users. My first hypothesis was that birdwatch users would form clusters based on similar note ratings. The data show that users form clear bi-modal networks, which likely represent their political ideology. While this polarization exists for the birdwatch community at large, it is especially clear cut within the most active users. Second, I hypothesized that different topics would result in different levels of polarization. I found through text analysis that the most common topics of tweets that were flagged were political in nature. For these tweets there is a demonstrated difference in how groups of users rate them. For non-political tweets, I was not able to find a topic or group of topics where Birdwatch users were clearly not polarized. This could very well be due to the low prevalence of non-political tweets in the birdwatch dataset. Third, I hypothesized that users would not form clusters based on which tweets or notes they chose to evaluate. Twitter has explicitly designed the platform to avoid this type of clustering, and the data supports that it does not occur.

When attempting to analyze how similarly users rate notes, choice of metric is important. Creating a bipartite network allows users to be connected through the notes that they have reviewed. While a simple count of how many times users agreed on the helpfulness of a note provided some insight, I found that taking the difference of when they agreed and when they did not provided a clearer picture. As such, I argue that dot-product-connectedness is an effective measure of ideological similarity which could be used in studying other similar networks. This metric produces a connectedness score which is consistent with real world beliefs about ideological similarity, and which does not a priori give more weight to active users. Graphs

and network analysis in general allow for hard to quantify trends to be easily visualized. Importantly, when problems relate to relationships between individuals, network analysis is a very important tool.

Limitations and Further Research

Birdwatch offers a unique opportunity, which is that all data is freely available and of a size which is workable. However, because Birdwatch is only a test program and contains a relatively small base of users compared with the general population on twitter, it is prone to selection bias. This bias could be a driving force for the stark polarization found in this paper. However, as one of the largest crowd sourced content moderation programs to date, Birdwatch data analysis can inform both future policy choices for Twitter and for other social media platforms in a similar position. What is more, as users continue to join, we will get a fuller picture of both the practicality and challenges associated with Birdwatch.

While graphs can provide a visual insight into the structure of networks, more statistical tests need to be done to verify the concreteness of these relationships. An interesting path would be to use community detection algorithms and centrality measures to further analyze the networks. Unrelated to networks, predicting the usefulness of notes or measuring the effectiveness of users in creating notes given available data would also be interesting.

Conclusion:

Crowdsourced content moderation offers the possibility of both accurate and large scale review of online content. However, research shows that the process by which people give their input can greatly shape their output. Through an analysis of Twitter's Birdwatch Data, I find that there exists significant polarization that could impact the efficacy of Birdwatch's effort. While it may not be impossible to utilize such polarized opinions–a process which necescitates collaboration and consensus building could be more effective.

References

Do digital echo chambers exist? (2019). BBC News. https://www.bbc.com/news/entertainment-arts-47447633.

Gorwa R., Binns R. & Katzenbach C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society 7 (1): 2053951719897945. https://doi.org/10.1177/2053951719897945.

Guilbeault D., Becker J. & Centola D. (2018). Social learning and partisan bias in the interpretation of climate trends. Proceedings of the National Academy of Sciences 115 (39): 9714–9719. https://doi.org/10.1073/pnas.1722664115.

Hansen D.L., Shneiderman B., Smith M.A. & Himelboim I. (2020). Chapter 4 - installation, orientation, and layout. In: Hansen D.L., Shneiderman B., Smith M.A. & Himelboim I. (eds.). Analyzing social media networks with NodeXL (second edition). Morgan Kaufmann, p. 55–66. https://doi.org/10.1016/B978-0-12-817756-3.00004-2.

How social media firms moderate their content. Knowledge at wharton (s.d.).. https://knowledge.wharton.upenn.edu/article/social-media-firms-moderate-content/ (accessed April 30, 2022).

Liu J. & Ram S. (2011). Who does what: Collaboration patterns in the wikipedia and their impact on article quality. ACM Transactions on Management Information Systems 2 (2): 1–23. https://doi.org/10.1145/1985347.1985352.

Pröllochs N. (2021). Community-based fact-checking on twitter's birdwatch platform. arXiv:2104.07175 [cs]. http://arxiv.org/abs/2104.07175.

Yasseri T. & Menczer F. (2022). Can crowdsourcing rescue the social marketplace of ideas? arXiv:2104.13754 [physics]. http://arxiv.org/abs/2104.13754.