Measuring Polarization in Online Communities*

using exponential family random graph models and sentiment analysis

William Gerecke

29 April 2022

Abstract

With the growth of the internet and social media platforms, it is increasingly easy for communities of like-minded people to form. This can be good, but often results in people strengthening beliefs by affirmation rather than by a decision making process. In this paper, I use Exponential Random Graph Models and sentiment analysis to measure polarization in online communities, specifically on the site Reddit. I find some evidence that indicates a diversity of opinions being expressed in several communities. However, it is still important for companies that make social media sites as well as individuals using the sites to consider the impact of forming homogeneous communities in the future.

Keywords: Social networks, Exponential family Random Graph Model, ERGM, Polarization, Sentiment analysis

*The code for this project is hosted on GitHub at wlgfour/social_networks

Contents

1 Introduction

With the popularization of the internet, and increasing ease of access, social media is becoming ubiquitous in modern society. These platforms facilitate instant communication between individuals, or between an individual and large audience. Beyond the ease and availability of such communication, social media platforms are run by highly-engineered algorithms that are designed to find people and communities that are similar to each other. For people like hobbyists, this is good since it allows people from niche groups to find each other easily. It is possible, however, to see that this could have a polarizing effect in a different scenario. For example, people with political beliefs are presented with like-minded opinions from around the world, and such communities strengthen their beliefs by affirmation, rather than by a decision making process. As such, it is important to be able to measure and understand the effect that online communities have on the people that participate in them, and the behaviors exhibited therein

Many social media platforms can be represented as graphs, as they represent an underlying social network of individuals. Graphs are data structures that are represented by a set of nodes V and a set of edges $E \subseteq \{(x, y) \in V^2, x \neq y\}$ which represent a connection between nodes. There are countless phenomena that can be represented by this data structure, such as roads, mathematical operations, supply chains, the internet, social networks, and more. As such, understanding and modelling graphical structures is crucial to understanding the universe that we live in, and has been a topic of much research. The problem with graphical structures is that typical assumptions of independence are violated by many of the scenarios that are represented by graphs. For example, in a social network where people share an edge if they are friends, the probability of sharing an edge is no longer independent. This is because it is reasonable that sharing a mutual friend will influence the probability of a friendship. Statisticians have developed Exponential family Random Graph Models (ERGMs) in order to allow for the representation of these dependent relationships.

In this paper, I look at polarization in online communities, specifically on the social media platform Reddit. For the purpose of this paper, I define polarization as the tendency for people to respond positively in the presence of positive sentiments, and negatively in the presence of negative sentiments. That is, polarization is defined to be when individuals in a community have views and opinions that align. I use a graph to represent interactions within communities on the Reddit, and sentiment analysis to classify individual interactions as positive or negative. I find some evidence that indicates that differing opinions are being expressed on several popular Reddit communities, but highlight the fact that a better construction of the underlying data would yield more conclusive results.

I begin by describing how the data is gathered, cleaned, and parsed into a graphical structure. I proceed to elaborate on the model used for sentiment analysis, and the structure of ERGMs. After that, I discuss the results of applying these models to the data that was gathered in order to measure polarization in online communities. Finally I discuss the impact of the results, weaknesses with the approach presented, and suggest directions for future research.

2 Data

2.1 Software

The R programming language (R Core Team 2021) was used to generate the report, data, and analyses associated with this project. The `purrr`, `dplyr`, `stringr`, and `tidyr` were used for data manipulation Wickham (2021). Data was simulated using the `stringi` and `keyToEnglish` packages Candocia (2021). The `VADER`, which exposed an R API for the VADER NLP model was used to generate sentiment labels for the data (Hutto and Gilbert 2014). `jsonlite` was used to scrape data from Reddit using Reddit's json API (Ooms 2014). The `statnet` package was used to generate network data structures, as well as fit ERGM models Hunter et al. (2008). Finally, the code and data are hosted on GitHub ("Where the World Builds Software," n.d.).

2.2 Reddit

The site Reddit was used to gather data for this project. Reddit is structured such that there are communities that users can subscribe to called subreddits. When someone creates a post, the post is categorized under a subreddit. Once a post is created, users can then comment on a post, as well as other comments. "Has-a" relationships are as follows:

- subreddit: N/A
- post: subreddit, author
- comment: parent (post or comment), subreddit, author

That is, a post has a subreddit and author that it is associated with. This project uses a dataset composed of instances of comments which have a parent, author, and subreddit associated with them, in addition to a comment body (text), and other metadata.

2.3 Gathering

Data was gathered using the Reddit json API. Given any Reddit url `u`, a json string representing that url can be obtained by downloading the url given by `{url}/.json`. The algorithm for scraping comments is as follows:

1. Initialize `links` to have top posts from some seed communities. Initialize `comments` to be empty
2. Randomly select a `link` from `links`
3. Append all comments from `link` to `comments`
4. If `link` represents a use: add the posts from the users most recent comments to `links`
5. If `link` represents a post: add the users from the posts comments to `links`
6. Remove duplicates in `comments` and `links`. Remove `links` that have already been visited. Remove rows in `comments` that contain `NA` values.
7. Goto 2.

At the start of the algorithm, a cache file is generated, and the algorithm caches `comments` every 10 iterations. For this project, the seed was set to three of the most active subreddits, 'AskReddit,' 'worldnews,' and 'gifs' (reddit 2011).

Statistics for the raw data that was gathered can be seen in table 1. There were 7208 subreddits visited, but the data collected for most of them was fairly sparse. It is important to note that when Reddit returns information, it hides many of the comments in a post by default. It is possible to retrieve the hidden comments by expanding certain links, but I did not do that for this project, which explains why the comment multiplicity is so low, averaging between 1 and 2 in the dataset. We can also see that significantly more comments have been gathered from AskReddit than any other subreddit. This is probably because this was one of the seed communities, and the posts on AskReddit have a high multiplicity. As such, many links were gathered from AskReddit early, making it more likely to gather more links from AskReddit, which in turn makes it more likely to gather even more links from AskReddit. This is an artifact of the algorithm that I used to scrape data, and the report would benefit from designing an algorithm that does not have this weakness.

2.4 Cleaning

Once the data scraping script is run, and there are one or more chunks of raw data, the data are combined, filtered, and labelled with sentiment scores. Due to computational constraints, the amount of data that was processed had to be severely limited. The first computational constraint came from generating sentiment scores, which is a very computationally intensive process. The second is fitting the model, and assessing its

Table 1: The table shows the summary statistics for the 15 subredits with the most comments recorded, and the total summary statistics for all subreddits visited.

Subreddit	Posts	Unique users	Average replies per comment
AskReddit	22,605	3,829	1.66
worldnews	13,992	2,795	1.45
gifs	1,521	1,125	2.76
ukraine	1,492	327	1.15
news	1,163	555	1.17
politics	1,051	366	1.20
AmItheAsshole	875	207	1.18
funny	790	473	1.27
interestingasfuck	788	449	1.15
technology	763	369	1.19
movies	694	360	1.16
Damnthatsinteresting	644	311	1.16
AskMen	632	204	1.25
antiwork	628	310	1.26
mildlyinteresting	539	338	1.17
Total	117,437	44,108	1.02

goodness of fit, which was another source of significant computational burden. Even with ample time, R experienced frequent crashes, indicating that in order to perform analyses on the entire dataset or a larger sample, better software design is important.

The filtering process excluded portions of the data for two reasons. One is because there weren't enough data points, and the second is because there were too many data points for the analyses to be computationally feasible. Data were first gathered and grouped by subreddit. Then data from a subreddit with fewer than 500 recorded comments were dropped. Finally, if a subreddit had more than 1000 recorded comments, 1000 were selected at random.

The sentiment scoring process was the main computational bottleneck in the data preparation process. With better software design, the process cold be parallelized to greatly speed up the labeling process and allow for larger datasets. This, however, would likely need to be done outside of the R language because even when run using only one thread on about 20,000 comments, the R garbage collection process caused the program to crash repeatedly.

In table 2, we can see the information for the 15 subreddits that were included in the cleaned dataset. As we can see, the subreddits have between 500 and 1000 comments each, with some subreddits having significantly more unique users. This is most likely due to the random sample that was taken for groups that had more than 1,000 comments. We can also see that the subreddits that were sampled are significantly sparser than they were before the sampling. A better sampling operation that preserves the structure of the comment chains could be advantageous when fitting models to represent the interactions that are present. We can see that there is one subreddit with only 54 values in the cleaned data. This is because the script that cleaned the data and labeled the sentiments always crashed before labeling all of the sentiments.

Figure 1 shows the average sentiments of the comments that were gathered for each subreddit. We can see that overall, most of the comments are slightly positive, but mostly neutral, with very few negative comments relative to positive comments. Beyond that, we can see that news and EldenRing have the lowest average sentiment scores, with gifs having a much higher average score.

Table 2: The table shows the summary statistics for the 15 subredits with the most comments recorded, and the total summary statistics for all subreddits visited.

Subreddit	Posts	Unique users	Average replies per comment	Avg. sentiment
AskReddit	1,000	771	1.20	0.03
gifs	1,000	783	2.54	0.20
news	1,000	512	1.16	-0.02
AmItheAsshole	875	207	1.18	0.09
funny	790	473	1.27	0.07
interestingasfuck	788	449	1.15	0.06
movies	694	360	1.16	0.12
Damnthatsinteresting	644	311	1.16	0.07
AskMen	631	204	1.25	0.13
antiwork	628	310	1.26	0.04
mildlyinteresting	539	338	1.17	0.06
Eldenring	537	120	1.10	0.00
gaming	515	298	1.11	0.12
memes	503	229	1.10	0.05
politics	54	44	1.04	0.03
Total	10,198	5,409	1.26	0.07

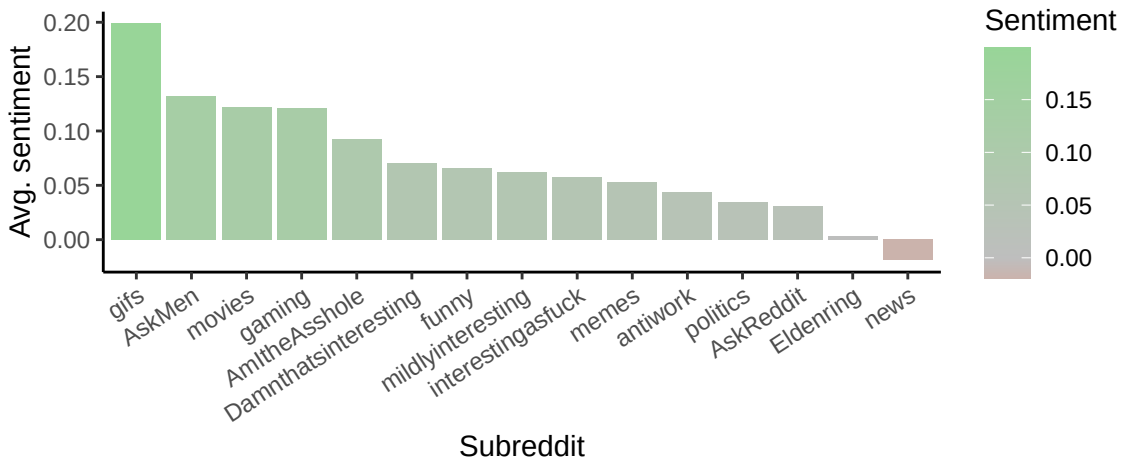


Figure 1: Average sentiment expressed in comments on each subreddit that was included in the cleaned dataset.

2.5 Graph construction

In order to create the graphical representation of the data, it was important to develop a scheme that represented the underlying structure of interactions that are present in the online conversations gathered. Several different approaches were considered, and the one that was selected is outlined below. It is important to note that there is likely a better design that allows for better representations of the data. It is also important to note that the way the graph is constructed directly influences how accurately it represents the network as well as the computational feasibility of the models that operate over the graphical representations.

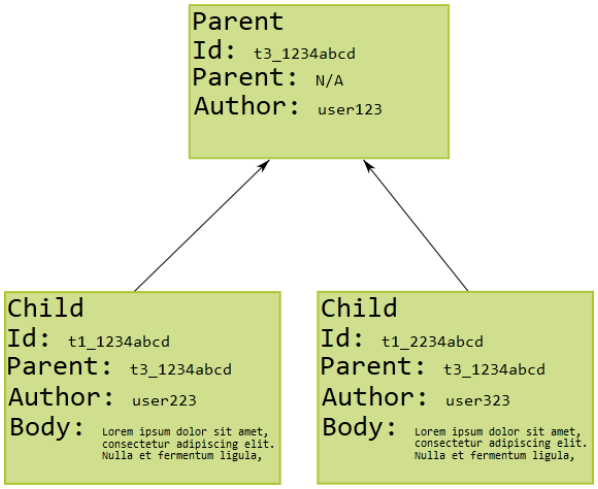


Figure 2: The figure shows the relationships expressed by the graphical construction used by this project. Each square represents a node, and the text contained within represents the attributes associated with each node. The parent in this case represents a post, and would not have a body, but it could also be a comment.

Figure 2 describes the graphical structure that is used in this project. Consider the data examined by this project to be a social network. Social networks can be represented as a graph with nodes V and edges $E \subseteq V \times V$. I decided to use a directed graph where each node represents a post or comment. Nodes have an associated sentiment, author, subreddit, parent, and ID as seen below:

```
node <- {
  .sentiment
  .author
  .subreddit
  .parent
  .id
}
```

There are two constraints on the nodes:

1. $node.sentiment \in [-1, 1]$
2. $\{node.parent | node \in V\} \equiv \{node.id | node \in V\}$

Exceptions to these rules are posts themselves, since top level comments are all direct children of posts, which do not have a parent or sentiment. In this case, $parent.parent = NA$, and is not included in the graph. The result of this is that comment chains will appear as inverted trees.

Since the graph is directional, the existence of an edge $(u, v) \in E$ does not guarantee the existence of edge (v, u). In fact, because comments on Reddit form a tree, and each comment can only be a reply to one

Table 3: The table shows the descriptive statistics of the graphs that represent the data collected for each subreddit.

Subreddit	Nodes	Edges	Density
AmItheAsshole	1,617	875	3e-04
antiwork	1,617	875	3e-04
AskMen	1,617	875	3e-04
AskReddit	1,617	875	3e-04
Damnthatsinteresting	1,617	875	3e-04
Eldenring	1,617	875	3e-04
funny	1,617	875	3e-04
gaming	1,617	875	3e-04
gifs	1,617	875	3e-04
interestingasfuck	1,617	875	3e-04
memes	1,617	875	3e-04
mildlyinteresting	1,617	875	3e-04

parent, this explicitly guarantees that $(v, u) \notin E$. Another side effect of the fact that each node will only originate one edge is that the graph will be very sparse, with $|E| \in \mathcal{O}(V)$.

Furthermore, I define one graph for each subreddit, *sub*, as the graph over the nodes $V_{sub} \equiv \{n | n.subreddit = sub,\ n \in V\}$. Note that there are no edges that need to be severed due to this construction because any comment will explicitly be a reply to some media that was posted in that subreddit. Thus, $G_{sub} = (V_{sub}, E_{sub})$ is a straightforward construction.

Table 3 shows the statistics for the graphs constructed for each subreddit. We can see that the graphs are extremely sparse. This is because of the fact that I did not expand the comments that were hidden when scraping data. This is also due to the fact that I had to sample the data during the cleaning process, and I did it in a way that did not preserve comment chains which resulted in the artifact shown in the table. This is supported by the fact that the subreddits that were subsampled to a lower proportion of the total data collected have a lower density.

3 Model

3.1 Sentiment Analysis

Sentiment analysis was done using the VADER (Valance Aware Dictionary for sEntiment Reasoning) model from NLP that was specifically designed to perform well on content from the internet (Hutto and Gilbert 2014). VADER is capable of identifying polarity and intensity of the sentiment expressed by text and returns a composite sentiment score between -1 and 1, where -1 represents a strong negative sentiment, 0 is neutral, and 1 represents a strong positive sentiment.

VADER uses a dictionary to identify lexical fragments and assign sentiment scores (intensity and polarity) to components of the body of text. The dictionary that maps lexical features to sentiment scores was trained using manually labeled data. VADER is also capable of understanding contextualization such as "I don't like this," as well as non-standard contractions such as "like'nt" by using some simple heuristics that the authors decided on. The scores assigned to lexical components are then averaged, and normalized.

3.2 ERGM

The model used to understand the graphical structure of the data which is examined in this report is the ERGM. This is because typical statistical models assume that the observations that they operate over are independent of each other. For example, if two people are friends on a social network, they are more likely to interact with each others' media, hence increasing the probability of an edge between them, and invalidating the assumption that they are independent of each other. In order to account for the dependence between the observations, we need to use a model that takes into account the underlying representation of the dataset. The ERGM is analogous to the generalized linear model, but takes into account the underlying structure of the graph, which is why it's appropriate for this application.

Given an observed network, the ERGM estimates the parameters of an expontential family model that takes the form of a log-linear combination of feature weights (Wyatt, Choudhury, and Bilmes 2009):

$$p(\mathbf{Y} = y) = \frac{1}{Z_\eta} e^{\eta^\top} \phi(y)$$

Where:

- $\mathbf{Y}$ are weights representing the edges of the graph
- ϕ defines the features over y
- η is a vector of weights
- Z_η is a normalizing constant

Typically, features account for the structural dependencies int he graph, allowing the model to more intuitively reason over the graphical structure of the data. The problem with using these models in practice, though, is that models are highly prone to degeneracy. In order to assess model degeneracy, it is important to examine the goodness of fit (GOF) which uses the generative nature of the estimated ERGM to find simulated networks. The simulated networks are then used to provide estimates for features such as node degree, edgewise shared partners, and geodesic distance. If the estimated networks align with the observed network, the model can be said to be robust.

4 Results

In order to assess polarity in the dataset that I gathered, I fit an ERGM model to the graph that represents each subreddit. For model features, I used the sentiment covariances, as well as basic geometric predictors that are present in the graph, such as edges which is equivalent to using the mean in a linear model. The idea between using the covariance associated with the sentiment score is that if the sentiment score covariance is found to have a positive impact on the formation of an edge, then this indicates that comments are likely to be replies to like-minded comments. That is, comments are more likely to be replies to comments that express the same sentiment. Similarly, a negative impact on edge formation signifies that comments are more likely to disagree with each other. In other words, if a comment expresses a positive sentiment, it is likely to receive a negative reply.

Table 4 describes the model statistics that were fit to the different subreddits. In Appendix A, we can see an example of the GOF assessment of the gifs community. In both Table 4 we can see that the models we see are fairly weak, and this is likely due to the fact that they are heavily influenced by the construction that was used to build the graphs. First of all, we can see that the base log odds of an edge is extremely low, which is because of the fact that the graphs are so sparse. Furthermore, in Appendix A, we see more evidence of this in the GOF plot for geodesic distance, as it indicates that the vast majority of nodes have no path between each other.

With this in mind, we can see that the sentiment covariance is negative, indicating a lack of polarity, or a diversity of comments. However, we can see that the magnitude of these estimates are fairly small relative

Table 4: The table shows the estimated parameters for the models fit to each subreddit graph as well as their confidence levels.

Subreddit	Edges	Pr(edges>|z|)	Sentiment cov.	Pr(cov(sentiment)>|z|)
AmItheAsshole	-8.00	0	-0.02	0.706
antiwork	-7.61	0	-0.02	0.821
AskMen	-7.61	0	-0.06	0.445
AskReddit	-8.11	0	-0.02	0.791
Damnthatsinteresting	-7.70	0	-0.02	0.772
Eldenring	-7.58	0	0.00	0.994
funny	-7.82	0	-0.05	0.556
gaming	-7.52	0	-0.03	0.753
gifs	-7.21	0	-0.33	0.000
interestingasfuck	-7.91	0	-0.02	0.762
memes	-7.51	0	-0.02	0.876
mildlyinteresting	-7.52	0	-0.03	0.758

to the estimates for edges, indicating that perhaps the sentiment of a post has little impact on the sentiment of the posts around it. These observations are mostly invalidated, however, by the fact that the confidence levels associated with the covariance term are extremely low.

5 Discussion

5.1 Findings

As seen in the results section, I find that there is little polarity in several popular reddit communities, and in fact, some evidence that indicates a diversity of opinions. The results, however, are not very conclusive because of the way the underlying structure of the data gathered was changed during the cleaning process.

5.2 Weaknesses

The approach described in this report has many weaknesses that severely affect the quality of the results. Most importantly is the fact that the structure of the data being used to fit the model is different from the structure of the data that was gathered, which is different from the structure of the data as it appears on the internet. These weaknesses come first from the fact that the data scraping script did not expand comments that were hidden by the initial API call, as well as the fact that the obtained data were subsampled in order to make the labeling and model fitting computationally tractable. This greatly affects the impact of the model results, as they no longer correspond to the information that is contained on the original social media site.

Second is that some communities are significantly over represented in the dataset, as seen specifically in the case of AskReddit. This is because of the way the data scraping script was constructed. Consider the links visited as a tree where each link is a node, and its children are the links that it adds to the pool of links. The scraping script selected a leaf at random, with equal likelihood of selecting a leaf regardless of its depth, however, this caused the artifact described above. In order to fix this, the leafs should be selected with a probability that is inversely proportionate to their depth in the tree.

Finally, the model that was used to fit the data was very simple, and there are many more features that could explain relationships in the dataset. An example of this is the directional covariance of the sentiment scores of comments. For example, the sentiment of a node's parent without the sentiment of the node itself

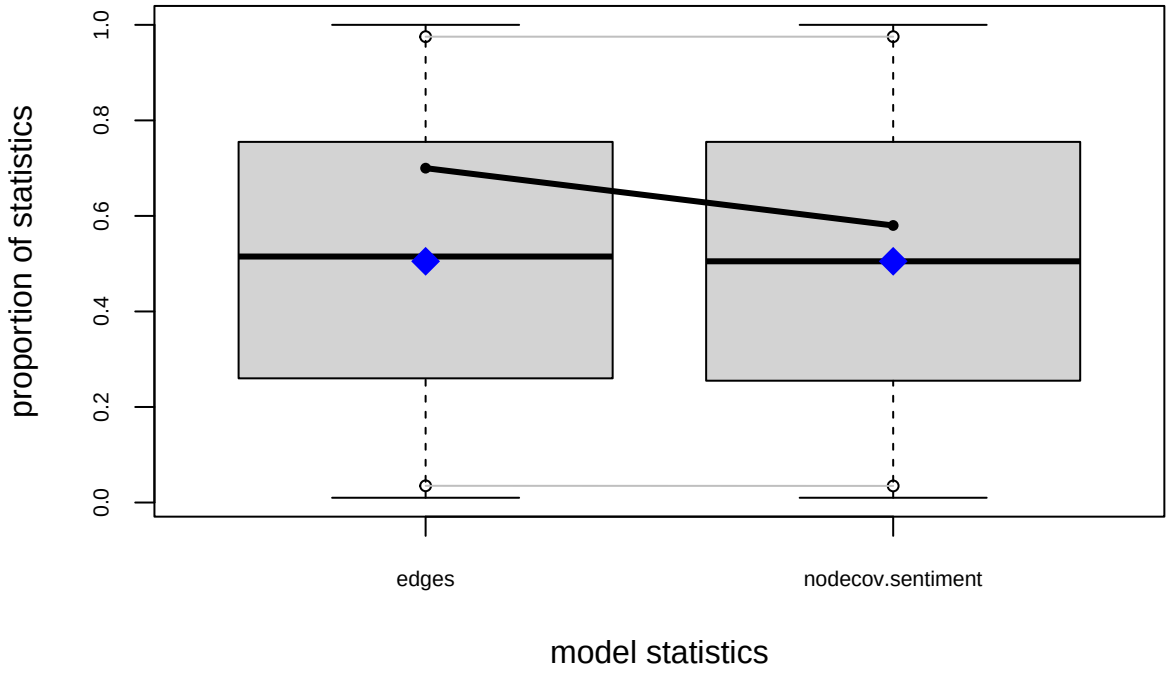
could be a predictor for edge formation. Another example is that the upvotes of a post, or popularity, could affect its visibility, and therefore the likelihood of an edge being formed.

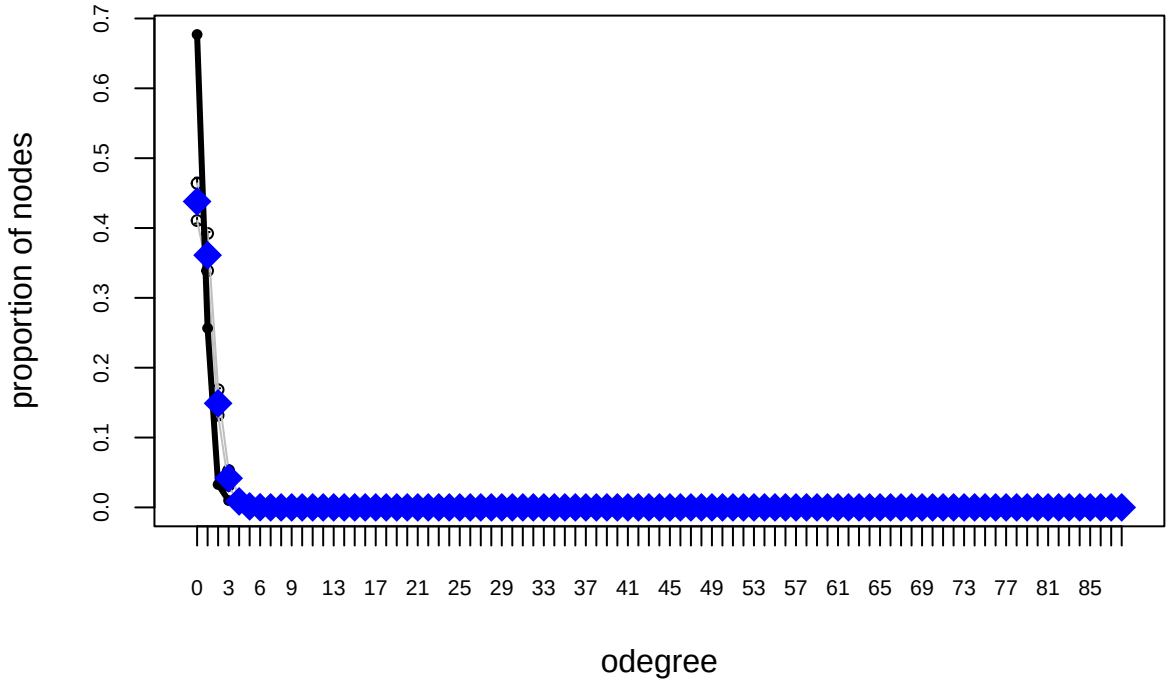
5.3 Future work

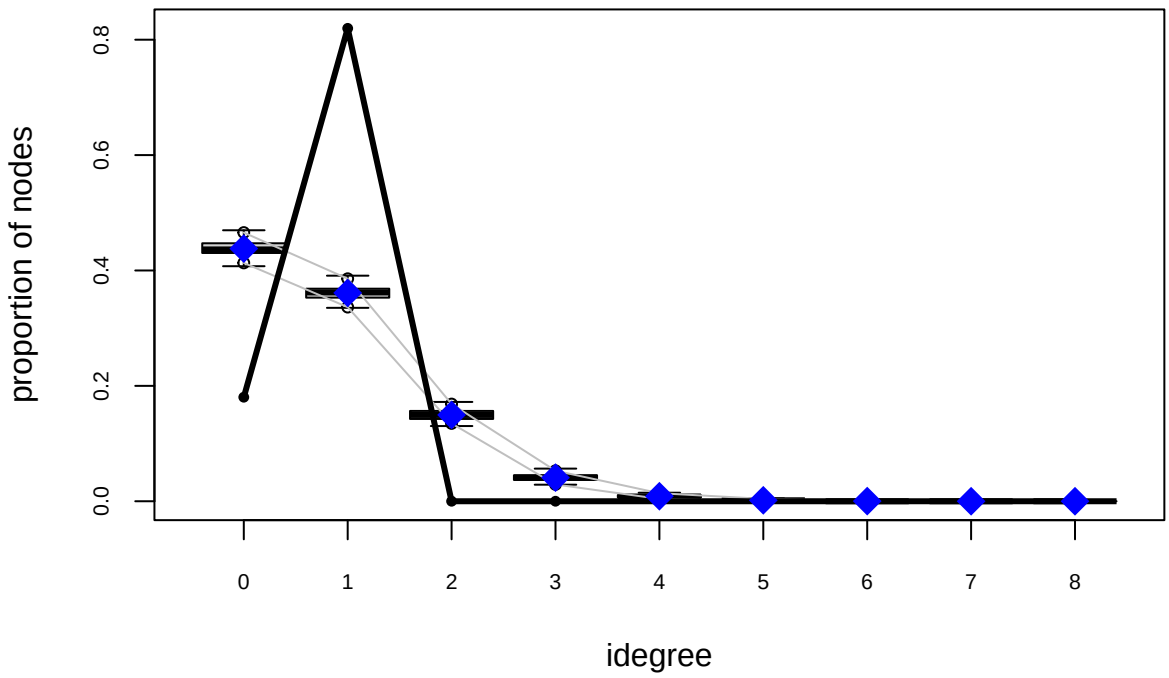
It is very important to understand the effect that social media has on the general population. One of the most impact ways that social media algorithms affect peoples opinions is the people that it connects with each other. Connecting like minded people and not connecting people with different opinions could easily lead to increased polarization in communities. As such, it is important to pay attention to the effect that automatically grouping people into communities has, and develop methods that allow people to examine these effects.

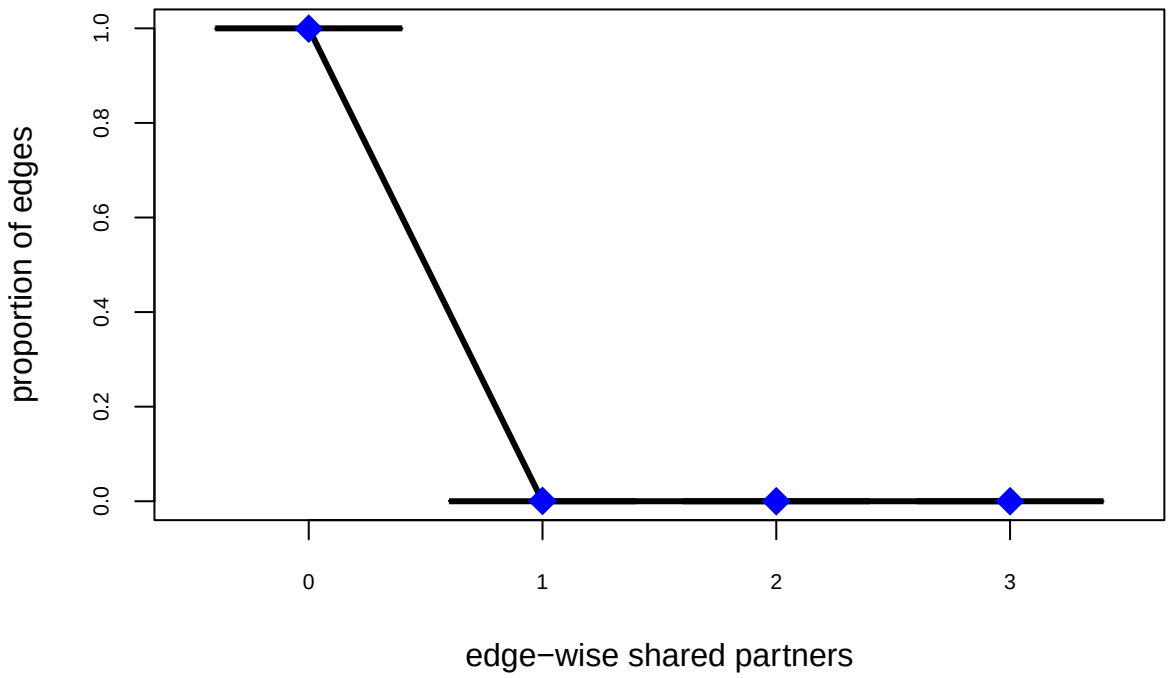
With respect to the work done in this project, there are three major points for future work. The first is developing a scraping algorithm that gathers a more representative sample of the data contained on social media sites. The second is finding a graphical construction that allows for better representations of online conversations, as well as more computational efficient representations of online environments. Finally, creating a model that better fits the data is important in order to fully understand the patterns that occur in online environments.

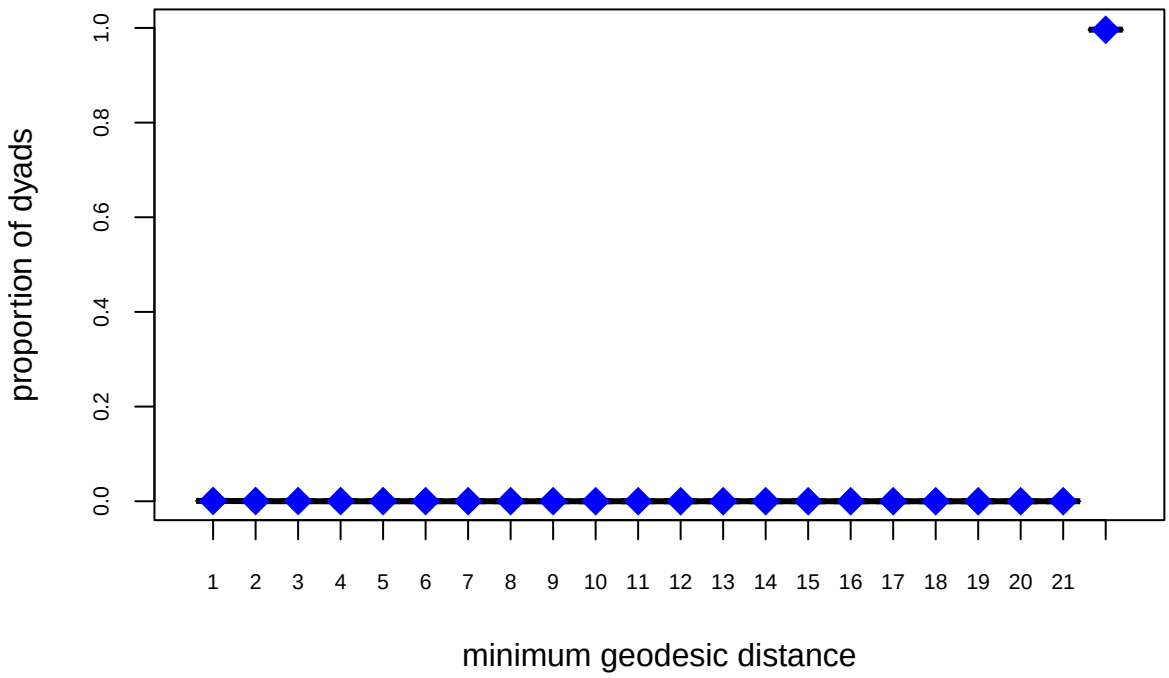
6 Appendix A: GOF plots for r/gifs






proportion of nodes
0.0
0.2
0.4
0.6
0.8
0
1
2
3
4
5
6
7
8
idegree


proportion of edges
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
edge−wise shared partners


Goodness−of−fit diagnostics
proportion of dyads
0.0
0.2
0.4
0.6
0.8
1.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
minimum geodesic distance

7 Appendix B: Data sheet

Extract of the questions from Gebru et al. (2021)

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
    - The dataset was created in order to examine polarity in the online social media site, Reddit. It was constructed with sentiment labels obtained via Vader (Hutto and Gilbert 2014), and with the goal of reporesenting comments as a social network.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - William Gerecke.
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
    - There was no funding for the project.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*
    - Instances of the dataset represent comments on the social media site Reddit.
2. *How many instances are there in total (of each type, if appropriate)?*
    - The size of the dataset varies, but is on the order of 100,000 instances.
3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*
    - This is an extremely small sample, as Reddit has over 1 million comments per day (reddit 2011). The size of the sample was limited by computational constraints. The sample was gathered by randomly performing a breadth first search over the social network. Links to posts were gathered by collecting the most recent activity of users on a given post, and comments were gathered from the posts. Since the posts have high multiplicity with respect to linked posts, nodes in the network were added to a list, and selected at random when visiting a post.
4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*
    - Each instance of the dataset contains:
        - `subreddit`: the community that it was posted under
        - `author`: the user id of the account that posted the comment
        - `id`: the id assigned to the specific comment
        - `to_id`: the id of the comment or post that the comment replied to
        - `permalink`: the url to the post that the comment was posted under
        - `timestamp`: the UTC timestamp that the comment was created at
        - `body`: the raw text of the comment

  - `sentiment`: a score between -1 and 1 that represents the sentiment of the comment as generated by Hutto and Gilbert (2014)

5. *Is there a label or target associated with each instance? If so, please provide a description.*
   - There is no explicit label, the dataset is meant to provide a sample of the structure of online interactions.
6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*
   - No data is missing from individual instances, as incomplete instances were dropped from the dataset. This was due to information appearing as `NULL` when scraped, or in some cases, http errors.
7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*
   - Yes, each instance in the dataset contains the id of the media that it is a reply to. That is, each instance of the dataset represents an edge in a social network.
8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*
   - There is no recommended splits, as the model used in this project had no need for partitioning the data.
   - If the data is split, it is important to respect the underlying structure of the data, and a random split would most likely lead to inaccurate model metrics. For example, it is important not to include comments from the same comment chain in the training and test data sets.
9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*
   - Redundancies were removed in the cleaning process, but it is important to note that the sentiment is most likely fairly noisy. While sentiment prediction models are accurate when operating over text that strictly adheres to language standards, online communication is much more nuanced. Vader was specifically designed to work well in online scenarios, but still has challenges (Hutto and Gilbert 2014).
10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*
    - The dataset is self contains and can be viewed of a snapshot of the Reddit taken at the time that the dataset was created. It contains links to media hosted by Reddit that may not necessarily be valid in the future.
11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*
    - The dataset does not intentionally contain confidential data.
12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

- Yes, Reddit is a social media site that has a wide variety of communities, including communities that are offensive. There are also communities oriented towards viewers 18+ which contain explicit material. As such, comments contained in this dataset may be offensie or explicit.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

    - The only sub populations identified by the dataset are the communities on the site Reddit, subreddits. Each comment in the dataset belongs to a subreddit, and is explicitly labelled.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - No. While accounts are recorded, the accounts are anoymous unless someone intentionally published their personal information.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

    - Yes. There are many different subreddits, some of which are oriented towards sensitive topics. As such, the comments may contain sensitive material.

16. *Any other comments?*

    - There are no other comments.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

    - The data was directly acquired from http://reddit.com, and sentiment abels were generated using Vader.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

    - The data was collected using a script that visited links to posts, and scraped comments. The script used a useful API published by Reddit that allows json representation of any content on the site. It was assumed that the data returned by the API is accurate.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

    - The sampling strategy is analogous to randomly advancing a breadth first search over the social network. If each post is considered a node in a social network, a random node was chosen, comments were gathered from that node, and new nodes were discovered. New nodes were discovered by gathering recent posts that users had commented on for each user who commented onthe post being scraped.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

    - I was the only person involved in the data collection process and I was not compensated.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*
    - The timeframe that data was collected over was very narrow, because of computational limitations. The analyses performed in this report became computationally intractable even with datasets around 100,000 elements long. As such, only a small amount of data was gathered.
6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*
    - There was no ethical review process.
7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*
    - The data was obtained via the Reddit website.
8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*
    - There was no notification given about the data collection.
9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*
    - Consent for data collection was not give, however the data was publicly available.
10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*
    - N/A
11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*
    - No.
12. *Any other comments?*
    - No other comments.

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*
    - Labelling was done using the Vader sentiment model.
2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*
    - Yes, the raw data was preserved, and can be found in the project repositroy.
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

- The software is available in the project repository under scripts.

4. *Any other comments?*
    - No other comments.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - No, the dataset has not been used for any other tasks.
2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*
    - N/A
3. *What (other) tasks could the dataset be used for?*
    - The dataset could be used for a variety of tasks associated with sentiment expression in online communities, such as assessing overall sentiment in communities, or toxicity.
4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*
    - No.
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
    - The dataset should not be used to attempt to identify individuals from their online presence.
6. *Any other comments?*
    - No other comments.

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*
    - It will be publicly available on GitHub.
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
    - The dataset will be available on GitHub, but does not have a DOI.
3. *When will the dataset be distributed?*
    - The dataset will be available immediately.
4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*
    - The dataset will be available under the MIT licence.
5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

- No.

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*
   - No.
7. *Any other comments?*
   - No other comments.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
   - There will be no supporting maintenance.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
   - I can be contacted via email at wlgfour@gmail.com, or via the GitHub associated with this project.
3. *Is there an erratum? If so, please provide a link or other access point.*
   - No.
4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*
   - The dataset will not be updated.
5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*
   - N/A
6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*
   - Yes, data files will not be deleted, and are available on GitHub which keep a history of all files committed.
7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*
   - Contributors are welcome to fork the associated repository on GitHub, and use the code or data presented in the repository.
8. *Any other comments?*
   - No other comments.

References

Candocia, Max. 2021. *keyToEnglish: Convert Data to Memorable Phrases*. https://CRAN.R-project.org/package=keyToEnglish.

Gagolewski, Marek. 2021. *Stringi: Fast and Portable Character String Processing in r*. https://stringi.gagolewski.com/.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.

Handcock, Mark S., David R. Hunter, Carter T. Butts, Steven M. Goodreau, Pavel N. Krivitsky, and Martina Morris. 2018. *Ergm: Fit, Simulate and Diagnose Exponential-Family Models for Networks*. The Statnet Project (http://www.statnet.org). https://CRAN.R-project.org/package=ergm.

Henry, Lionel, and Hadley Wickham. 2020. *Purrr: Functional Programming Tools*. https://CRAN.R-project.org/package=purrr.

Hunter, David R., Mark S. Handcock, Carter T. Butts, Steven M. Goodreau, and Martina Morris. 2008. "Ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks." *Journal of Statistical Software* 24 (3): 1–29.

Hutto, Clayton, and Eric Gilbert. 2014. "Vader: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text." In *Proceedings of the International AAAI Conference on Web and Social Media*, 8:216–25. 1.

Ooms, Jeroen. 2014. "The Jsonlite Package: A Practical and Consistent Mapping Between JSON Data and r Objects." *arXiv:1403.2805 [Stat.CO]*. https://arxiv.org/abs/1403.2805.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

reddit. 2011. "Reddit." http://reddit.com.

"Where the World Builds Software." n.d. *GitHub*. https://github.com/.

Wickham, Hadley. 2019. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://CRAN.R-project.org/package=stringr.

———. 2021. *Tidyr: Tidy Messy Data*. https://CRAN.R-project.org/package=tidyr.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wyatt, Danny, Tanzeem Choudhury, and Jeff Bilmes. 2009. "Dynamic Multi-Valued Network Models for Predicting Face-to-Face Conversations." In *NIPS Workshop on Analyzing Networks and Learning with Graphs*. Citeseer.