Measuring Polarization in Online Communities[*]

using exponential family random graph models and sentiment analysis

William Gerecke

28 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 that ???**. As such, it is 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 that ????**.

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 ahve 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 benifit 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.

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 1 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 <- {
```

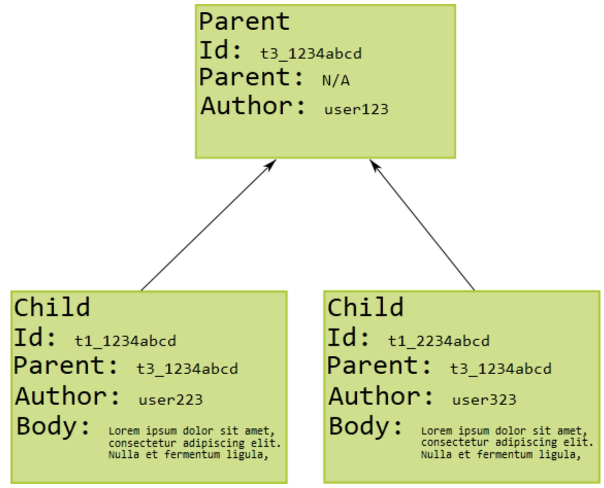
Parent
Id: t3_1234abcd
Parent: N/A
Author: user123
Child
Id: t1_1234abcd
Parent: t3_1234abcd
Author: user223
Body: Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nulla et fermentum ligula,
Child
Id: t1_2234abcd
Parent: t3_1234abcd
Author: user323
Body: Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nulla et fermentum ligula,

Figure 1: 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.

```
        .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 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.

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

5 Discussion

5.1 Findings

5.2 Weaknesses

- scraping algotithm visits some subreddits more often

- scraping didn't expand hidden comments
- computational tractability
  - small dataset

5.3 Future work

6 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/.

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.