Measuring Polarization in Online Communities*

using exponential family random graph models and sentiment analysis

William Gerecke

07 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 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 wherein.

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

##Software

2.1 Gathering

2.2 Cleaning

2.3 Graph construction

In order to create the graphical representation of the data,

6 References