How strong is the echo chamber? Examining the online discussion around the Irish Marriage Referendum

Luke Murray Kearney
Department of Mathematics and Statistics
University of Limerick

Supervisor
Dr David O'Sullivan

Interim Report for a Final Year Project
BSc in Mathematical Sciences
August 17, 2025

Abstract

We provide an examination of the mechanisms of Twitters social discourse on the 2015 Irish marriage referendum using sentiment and network analysis. Using a data set of 408,201 unique tweets this study aimed to use sentiment as an agent for homophily and to unearth political affiliation. Homophily being the tendency of members to interact at a disproportionate rate with similiar individuals. A reciprocated network, where every link in the network has at least one corresponding inversely directed link, was created and found to be much more useful for analysis of dialogue because of it's density of active users and reduction in the large inward focused hubs. We calculated sentiment scores for each tweet using a lexicographical approach to quantify emotion in text, giving a slightly positive overall sentiment centered at 1.03. Aggregation of user in and out sentiments provided us with quantitative measures of overall user emotion. Monte-Carlo simulations use single variable randomisations within a fixed topology to create a null model of random sentiments to compare our results to, which allowed us to test the significance of the empirical observations. The observations chosen are the Pearson correlation between users in and out sentiment and the fraction of both ends of a link having positive nodes in the network. Both results lie outside our confidence interval showing there is a strong positive correlation in user sentiment and the fraction of positive ends indicates homophily. Community detection was implemented on the network to get a deeper understanding into the topological structure of the network. Further grouping of these communities using k-means clustering gives a singular partition into two clusters based on the communities respective aggregated sentiment. Two testing methods were used to quantify the relationship between the clustering and user affiliation. The first method was keyword density, finding a definite variation in relative keyword density between clusters allowing us to classify each cluster as yes or no. To evaluate the accuracy of our classification a random sample is taken of 358 voters who have been manually assigned their true voting preference, to see if our model matches the ground truth, giving an 89% balanced accuracy. A re-scaling of the variables with their respective community size is implemented. This re-scaling increased polarisation in keyword density scores and balanced accuracy to 95%. If the communities contained ideologically heterogeneous users this type of ground truth testing wouldn't be accurate. This homogeneity shows political homophily inside communities and the formation of strong echo chambers of ideologically similar individuals.

Contents

Chapter 1

Introduction

Ireland held a referendum in the summer of 2015 to legalize same-sex marriage. The vote passed with a 62% majority [1],an exceptional point in Irish history, contrasting Ireland's once culturally conservative nature, given the legal enforcement of forbidding homosexual acts until 1993 [2, 3]. The pragmatic and polarising nature of the vote saw it have the highest total turnout of any Irish referendum [4]. The electorate's eagerness could be seen on social media, with Twitter being a hub for this activity [5].

Twitter is a micro blogging service where users can upload messages known as tweets of up to 280 characters [6]. Users of Twitter can subscribe to see each others tweets (follow), these users may also interact with each other through tagging other user's profiles in their tweets. Many of these relationships can be asymmetric, creating a parasocial relationship where one member is following and is interacting with another without any reciprocation [7, 8].

Twitter has become a favoured setting for the propagation of knowledge and beliefs, and facilitation of public discourse on many issues [9–16]. Network science attempts to gain understanding of complex systems, like those found in Twitter dialogue. Researchers analyse these discussions to acquire more understanding of the phenomena which control these social interactions [17]. We see these complex systems in everyday life and the set of Twitter users can be thought of as one, with over 500 million users [18]. Network Science uses mathematical tools to attempt to gain an understanding of the interactions within the system [19].

Sentiment Analysis is contextual mining of conversations [20]. In the case of Twitter data it can be used to understand the social sentiment of a certain event or product [21–23]. It has been used to predict film box office performances [24], stock market fluctuations [25–27] and election results [28–30]. Some of these study's results have been met with considerable skepticism due

to a lack of rigour and accepted repeatable processes [31]. Regardless the use of Twitter data to attain an understanding of the mechanisms of social interaction is an encouraging field of study. Homophily is the proclivity for similar people to be connected to each other at a higher rate than dissimilar people. This tendency follows the proverb "birds of a feather flock together" [32]. It is one of the most robust empirical recurrences in community behaviour, a focal point of studies from 1954 to the present [33].

Community detection on a network can reveal the fundamental functions by which our network operates [34, 35]. Communities are areas of high link density within graphs [36], which have been evident to researchers in social networks for decades [37–39]. After gaining an understanding of community structure clustering algorithms can be a method of finding homophily in communities. Clustering algorithms group points such that some clustering criteria is minimised [40]. A partition can then be used to test our hypothesis against the ground truth.

Using a data set of every tweet containing the hashtags "`#marref`" and "`#marriageref`" this report attempts to analyse the interactions between users in the lead up to same-sex marriage referendum. The analysis will attempt to use a combination of network science tools and sentiment analysis to answer the following:

- Does the sentiment of a user cluster over network homophily?
- Can you use social structure and sentiment to find yes/no voters and sentiment?

In answering these questions, I define the mathematical methods to be used in the analysis (§2). I filter the data set and use the defined mathematical tools to gain an understanding of the interactions between end users (§3). I calculated sentiment scores for each tweet in the data set and integrated it into the network of users (§4). I performed Monte-Carlo simulations and compared the random networks created to the network of interest (§5). I implemented community detection and k-means clustering to the network (§6). For the original clusters and a scaled cluster I used keyword densities to assign my clusters affiliation and tested this affiliation using manually assigned random sampling (§7).

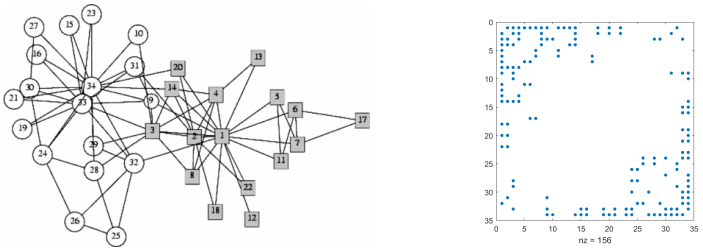
Chapter 2

Mathematical Tools

In this section I introduce some of the fundamental tools and mathematical concepts that I studied as part of network science, that will aid analysis of my network in later chapters when building and analysing networks.

2.1 Adjacency Matrix

An Adjacency Matrix provides a complete list of the links present in a network. It is a very common representation of a network [41] as it provides a complete picture of the connectivity patterns. It provides an easily readable and mathematically useful representation of these links for directed and undirected networks. For a directed network like the one we will be using in our analysis, with N nodes, the adjacency matrix A has N rows and N columns. If there is a link pointing from node i to node j, $A_{ij} = 1$. If there is no link pointing from i and j $A_{ij} = 0$. For an undirected network the adjacency matrix (A) is a symmetric matrix. If there is a link from node i to j then there is a link from j to i. $A_{ij} = A_{ji}$ as long as $i \neq j$ [36]. As an example of a network and adjacency network I will show Zachary's karate club, a famous complex network which will be discussed later in this project,



(a) (b)

Figure 2.1: (a) Zachary karate club network [42] and (b) corresponding sparse adjacency matrix, each dot = 1, empty space = 0 .

As we can see a complex network is a collection of nodes, with links carrying information between these nodes. The adjacency matrix can carry this information in an easily understood and computationally useful matrix. In many cases this representation of a network is sparse and can allow one to calculate the degree of nodes, and the total number of links in a network, along with many other fundamental quantities, which we will discuss later.

2.2 Clustering Coefficient

2.2.1 Local Clustering Coefficient

The local clustering coefficient (C_i) is a measure of the link density within the neighbourhood of the node in question, giving an indication for focal nodes in a network. Which can be easily visualized by the common phrase 'my friends, friend is also a friend of mine.' The degree (k_i) is a measure of the number of links connected to the node i, calculation of the degree (k_i) to which neighbours of a node link to each other and L_i is the number of links connecting the neighbours of i. The equation for the local clustering coefficient is [43]

$$C_i = \frac{2L_i}{k_i(k_i - 1)}. \tag{2.1}$$

This gives $0 < C_i < 1$, as the measure of the network's local link density around the node i.

- $C_i = 0$ if none of the neighbours of node i link to each other, 2 neighbours of the given node cannot be linked.
- $C_i = 1$ if all of the neighbours of node i link to each other, a complete graph. There is a 100% chance that 2 neighbours of a node are linked.
- $C_i \neq 0, C_i \neq 1$ gives the probability of 2 neighbours of a node being linked to each other. (i.e., if $C_i = 0.25$ there is a 25% chance that 2 neighbouring nodes of node i are linked)

2.2.2 Global Clustering Coefficient

The global clustering coefficient (C_Δ), which has been referred to as transitivity in some publications [44–46] is a measure of the number of closed triangles of nodes in a network compared to the number of connected triples of nodes. This is very important in our case. Social networks

(a) (b)

Figure 2.2: (a) Closed Triangle, (b) Connected triple

usually produce a very high degree of clustering, as many of their neighbourhoods form communities with dense internal connections [47]. The equation for the global clustering coefficient is [36]

$$C_\Delta = \frac{3 \times \text{Number of Triangles}}{\text{Number of Connected Triples}}, \tag{2.2}$$

giving C_Δ to be a number between 0 and 1, the closer to 1 it is the higher the amount of strongly linked triads there are in the network. The global and local clustering coefficients were designed for different tasks, the global trying to understand overall clustering in a network using triangles and the local giving a measure of the embeddedness of a node. The beginnings of the use of the global clustering coefficient can be found in the 1940's [37,48] and it has been an invaluable tool for network scientists ever since.

2.3 Shortest Path length

A path between two nodes is the links that are travelled going from the starting node to the target node. In real social networks these links are often short, with Milgram famously showing in 1967 that every two people in the United States are connected by around six intermediary acquaintances [49]. This is known as the "six degrees of separation" phenomena [50]. Consequential studies on online social networks have shown this average to be even smaller [51] in our increasingly interconnected world. The shortest path between two nodes (i and j) is the path with the fewest number of links possible within the network. It is denoted by d_{ij} or d.

It is possible to have multiple shortest paths between two nodes in the same network and the shortest path never contains cycles.

- For undirected networks, $d_{ij} = d_{ji}$ is equivalent.
- For directed networks, like twitter mention networks, often $d_{ij} \neq d_{ji}$ and if a path exists from node i to node j it does not guarantee the existence of a path from j to i [36].

The average shortest path length is the mean of shortest path length for every possible pair of nodes in the network. It is a metric of the efficiency of information movement within a network [52] and was used to find Milgram's six degrees of separation. it is defined as [52]

$$ASPL = \frac{\sum_{i,j}^{n} d_{ij}}{\sum_{i,j}^{n} path_{ij}}, \tag{2.3}$$

where

$$d_{ij} = \text{ the shortest path length between nodes } i \text{ and } j, \tag{2.4}$$

and

$$path_{ij} = \begin{cases} 0, & \text{if there is no path from } i \text{ to } j \text{ or } i = j, \\ 1, & \text{if there is a path from } i \text{to } j. \end{cases} \tag{2.5}$$

2.4 Assortativtiy

Assortativity is the correlation coefficient of degree between pairs of adjacent nodes, introduced by Newman in 2002 [53]. A measure of the rate with which nodes of different connectivity patterns connect to each other.

- The formula for Assortativity is [36]

$$r = \sum_{jk} \frac{jk(e_{jk} - q_j q_k)}{\sigma^2} \tag{2.6}$$

where

$$\sigma = \sum_{k} k^2 q_k - \left[\sum_{k} k q_k\right]^2. \tag{2.7}$$

- Assortativity (r) is the Pearson Correlation coefficient calculated for the degrees present at each end of an edge and r ranges from -1 to 1.

- $r > 0$ means this network has a tendency for nodes of similar degrees to be adjacent to each other. If this is the case the graph is assortative.
- $r \approx 0$ means this networks tendency for nodes of similar degrees to be adjacent is close to random. This is known as non-assortative.
- $r < 0$ means this network has a tendency for nodes of similar degree to not be connected to each other. This is known as dissortative.

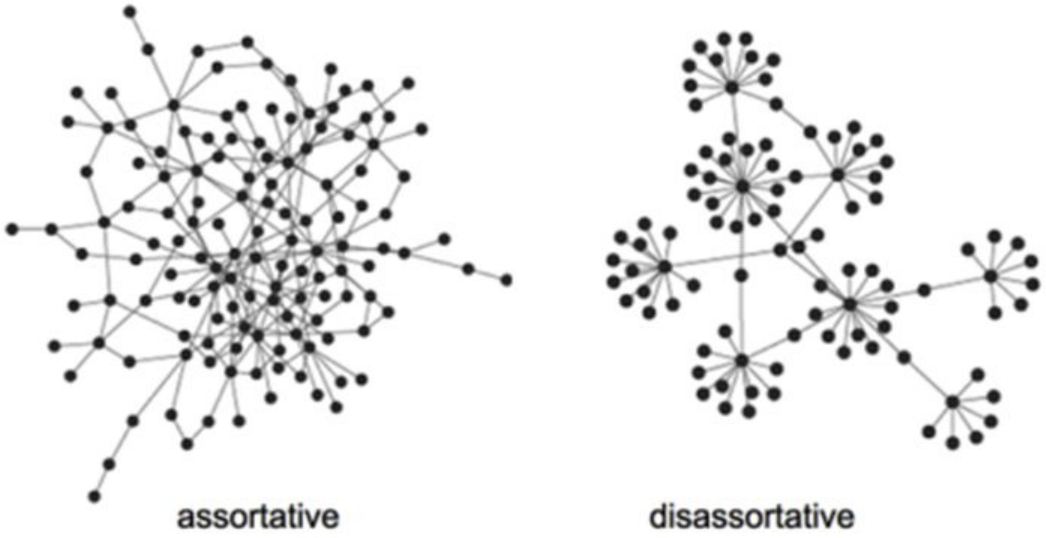


Figure 2.3: Assortivity vs Dissortavity, figure taken from [54].

As can be seen in figure 2.3 assortativity creates dense areas of nodes which have a tendency to attach to nodes of similar degree, closed triads are also visible in the graph which is indicative of a high global clustering coefficient. Whereas in the dissortative graph, wheel and spoke formation can be seen due to the hubs (nodes of a relatively high degree) being connected to nodes of a low degree, no closed triad formation can be seen and the path lengths are visibly longer. Assortative mixing (*i.e.* , An assortative graph) is thought of as one of the essential characteristics of a real social network [55], whereas biological networks tend to be dissortative [53].
These important mathematical tools will help in exploratory analysis of the Irish Marriage Referendum data set in the next chapter.

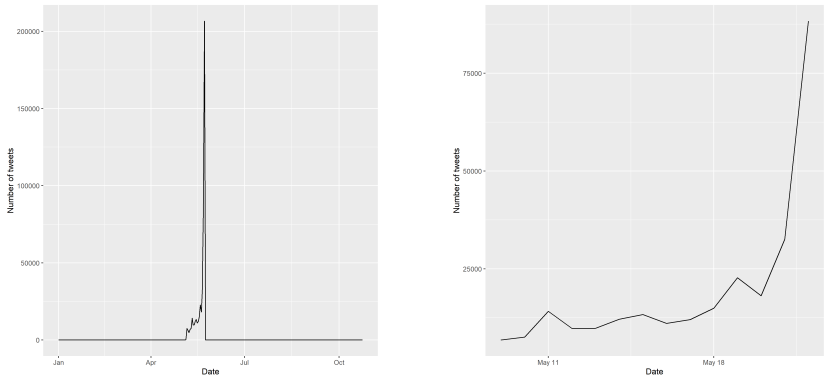
Chapter 3

The Irish Marriage Referendum Data

In this chapter I will begin exploratory analysis on on the data set of tweets to gain an understanding of this social discourse.

3.1 The Data Set

To begin the project, I was given the data set for this project by my supervisor. This is a collection of all of the tweets made with the hashtags "`#marref`" and "`#marriageref`" collected from the 1^{st} of January 2015 to the 24^{th} of October 2015, encompassing the referendum which took place on the 22^{nd} of May 2015. In this data set the majority of the tweets were made in the weeks prior to the referendum.



(a) (b)

Figure 3.1: (a) Tweets per day , (b) Tweets per day 2 weeks before referendum

Figure 3.1(a) shows the spike in tweets around on the date of the referendum, with over 80,000 tweets coming in one day and a steady increase in tweet volume in the days before the referendum. When dealing with this data later we will create a directed network of nodes (users) connected by links (tweets). This project is interested in understanding dialogue between users leading up to the day of the referendum, so we reduce the scope of the analysis to the 8^{th} of May to the 22^{rd} of May, the day after the referendum has the highest tweet frequency with over 200,000 but many of these are post commentary and are therefore not included in this analysis. Figure 3.1(b) shows the steady increase in interactions in the build up to the referendum, coming to a crescendo on the day after the referendum.

Each row of the data in figure 3.2 represents a tweet, with each row holding a piece of information on this tweet. The first-row "`created_at`" was used to create figure 3.1 as it holds the time and date each tweet was made. The second-row "`text`" holds the body of the tweet which we will use for sentiment analysis later in the project. The other rows ("`screen_name`", "`mentions`", "`hashtags`" and "`type`") hold the name of the users who made the tweet, the mentions in that tweet, the hashtags and the type of tweet it is (*i.e.* an original tweet or a retweet). There are 480,201 tweets with 142,672 unique tweeters, 135,370 (28.2%) of these are original tweets with 344,831 (71.8%) of them being retweets. A data set of this size requires the aforementioned mathematical tools to get an understanding of the network characteristics, unlike smaller data sets, which may allow for an initial gain of understanding via inspection.

	created_at	text	type	mentions	hashtags	screen_name
1	2015-01-01 22:33:55	Cathy & Eleanor took time out from their civil partners...	O	@yesequality2015	c("#lgbt", "#marref")	yesequality2015
2	2015-01-23 20:12:26	This is what you will all be voting on in May. #MarRef #Marr...	O	character(0)	c("#marref", "#marriageequality")	yesequality2015
3	2015-01-29 16:55:36	Encouraging words from An Taoiseach @EndaKennyTD #Ma...	O	@endakennytd	c("#marref", "#marriageequality")	yesequality2015
4	2015-02-09 22:35:07	The man himself, Gay Byrne speaks in support of #Marriage...	O	character(0)	c("#marriageequality", "#marref")	yesequality2015
5	2015-02-15 22:52:13	Submitting a postal vote for the referendum? Why not use ...	O	character(0)	c("#stampgate", "#marref")	dailigh
6	2015-02-20 01:42:15	Calling all Irish abroad. Join us in making the journey home ...	O	character(0)	#marref.	gettheboat2vote
7	2015-02-20 09:27:33	Irish abroad? Tempted to come home to vote yes in the #M...	O	@gettheboat2vote	#marref	joeykavanagh_
8	2015-02-20 11:30:47	Save the date! #MarRef http://t.co/uNRcR1xE2j	O	character(0)	#marref	yesequality2015
9	2015-02-21 17:02:00	"They will still contract cancers earlier in life". Handed this le...	O	character(0)	c("#adfam", "#marref")	graceyosmiley
10	2015-03-01 09:33:02	Time to build more equal society - vote for #marref.	O	character(0)	#marref.	form_architect
11	2015-03-01 21:52:40	Fair play to the Ireland team. Lets stand together and make ...	O	character(0)	c("#marref", "#ireveng", "#coybig")	yesequality2015
12	2015-03-02 19:14:19	"The Commission added, though that there was perhaps a t...	O	character(0)	#marref	limerick1914
13	2015-03-03 23:09:36	Who is this Patrick Tracey? And why does he support social ...	O	character(0)	#marref	bearaboi
14	2015-03-08 14:59:26	Happy Intetnational Women's Day. #IWD15 #MarRef http://...	O	character(0)	c("#iwd15", "#marref")	yesequality2015
15	2015-03-09 11:07:11	Guests beginning to arrive for our launch! @mannixflynn #...	O	@mannixflynn	#marref	yesequality2015
16	2015-03-09 11:17:42	More guests arriving at our launch this morning! Very exciti...	O	character(0)	#marref	yesequality2015

Figure 3.2: Snapshot of the data set used.

3.1.1 Reciprocated Network

In most real world social networks there is a high denisty of strong internal connections due to the high level of clustering and community behaviour [47], which creates a high level of reciprocity. In online social networks like Twitter this is not always the case. Twitter communities

with interpersonal interests have a more reciprocated network than those of a community with political interests [56]. This denser political network creates a paradigm of reliance on single members of the group. This alters the network topology creating parasocial interactions with hubs that may not reciprocate the interaction. The creation of a solely reciprocated data set may help rectify this reliance on hubs and allow us to get a greater understanding of the actual public discourse on the referendum.

To create the reciprocated data set, we make a subset of the 2 weeks prior to referendum data set with the following conditions on arbitrary users A, B in the data set,

$$\begin{cases} \forall \text{ tweets from user } A \ \exists \text{ a user } B \text{ mentioned in the tweet,} \\ \forall \text{ tweets from } A \to B \ \exists \text{ a tweet from } A \to B, \text{ unless } A = B. \end{cases}$$

Every retweet in the data mentions the original user so there is the possibility for these to remain in the data set if there is a corresponding reply from the original tweeter. This further filtered data set will also be used for data analysis in the next section.

3.2 Exploratory Data Analysis

We use Data and Network Analysis tools to try and understand the explanatory mechanisms prevalent and produce hypotheses on how network users are connected to each other [57]. Before undertaking this analysis we must create igraph objects in R of the 2 filtered data sets, this is done using the `graph.data.frame` function.

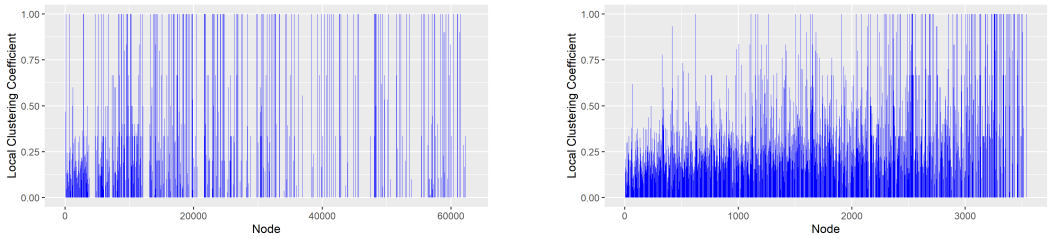
3.2.1 Clustering

The global clustering coefficient (C_Δ) can also be referred to as transitivity and it is found using the `transitivity()` function. C_Δ is 0.015 for the non-reciprocated network and 0.54 for the reciprocated network. These values are surprisingly small for a social network, indicating that closed triads are very uncommon.

$$\frac{C_\Delta(reciprocated)}{C_\Delta(original)} = 3.6$$

The filtering of the network's non-reciprocated nodes increases the global clustering coefficient. This increase in loop density increases the prevalence of link density and triad closure. Even with this increase in C_Δ the reciprocated network still shows the low clustering coefficient indicative of hub and spoke formation seen with politically charged Twitter social networks [47].

The local clustering coefficient of each node reveals a obvious sparsity in the original network which is drastically decreased in the reciprocated network. The local clustering calculation



(a) Original Graph (b) Reciprocated Graph

Figure 3.3: Local clustering coefficient for each node.

shows the presence of complete triangles in the graph, indicating a high level of embeddedness for certain nodes. Clusters are present in the original graph, but these complete subgraphs are much more prevalent in the reciprocated network.

3.2.2 Path length

The diameter is a measure of the longest of the shortest paths between any 2 vertices in a network. For the original and reciprocated networks the diameter is 14. This is relatively small for data sets of this size with low clustering coefficients. The average shortest path length is 4.62 for the original network and 4.60 for the reciprocated network. Higher than expected in our increasingly connected online world where Milgram's '6 degrees of separation' is becoming too high. As Tarr *et al.* [58] found this to be much lower at 3.57 in their analysis of the 1.59 billion people on Facebook [58], whereas older studies found this number to be 3.74 [51,59] and some even found it to be as low as 2.97 [60]. These small diameters and relatively large average shortest path length indicate the presence of dominant hubs, connecting many different users and acting as facilitators of discourse in both networks [57].

The low clustering coefficients deviate from the usual characteristics of real social networks, but have been found in other studies of Twitter data [61]. These may be correlated with the directed nature of a Twitter network, but the shortest path lengths aren't significantly larger than those found in recent studies.

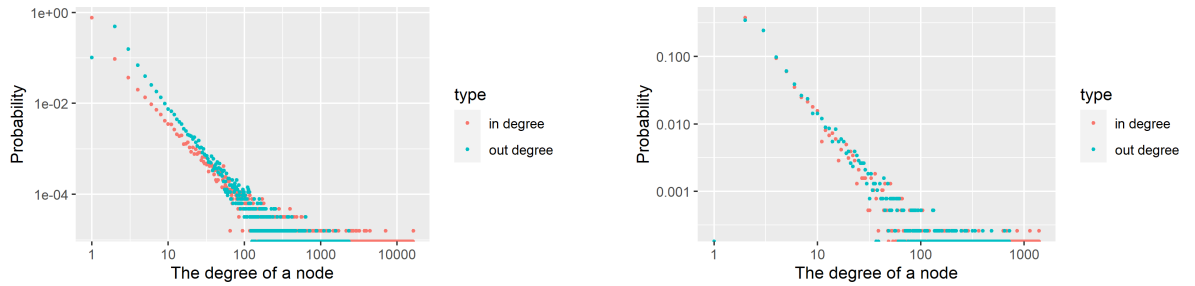
3.2.3 Assortativity

The assortativity for the original network is 0.018 and is 0.082 for the reciprocated network. This shows that both networks show little proclivity for nodes of similar degrees to be adjacent to each other, another exemplar of the lack of the innate characteristics of a social network. It tells us that both networks are neutral, the number of links between the hubs corresponds

with what we would expect by chance [36]. The micro-blogging site differs by it's directed nature from Milgram's famous 'six degree's of separation' experiment [49], links on Twitter don't require mutual agreement. The characteristics discovered do not follow the 'small world' topology outlined by Watts and Strogatz [43]. The small world topology is depicted as a highly clustered network with small path lengths very similar to Milgram's definition of six degrees of sepration [43, 49]. We are beginning to understand that our found characteristics must be thought of separately from our conventional social network structures. The reciprocated network shows a large improvement on the original networks low assortativity and it has better emulated the conventional structure so far.

3.2.4 Degree Distributions

The degree distribution is a measure of the probability for any node to have a certain degree.



(a) Original Network Degree Distribution

(b) Reciprocated Network Degree Distribution

Figure 3.4: In and out degree distributions

Both of these degree distributions follow the shape found in the Pak *et al.* sentiment analysis of Twitter data [62]. The visible inverse relationship between rank and frequency follows Zipf's law, which in this case is between probability and degree.

$$Probability \propto \frac{1}{Degree}$$

As can be seen in Figure 3.4, the probabilities and degrees are negatively correlated in both plots. These plots are vertically and horizontally scaled by log_{10} to allow them to be more easily understood, this logarithmic scale decreases the distance from the smallest values to the biggest. For these plots it creates an approximate linear decrease in probability, meaning these variables follow the power rule as most social network do with an exponent between 2 and 3 [61]. The power law describes a situation where if a node has more links it is more likely to get new links, in creating a network we can think of this as the rich get richer phenomena [63]. This mathematically creates a system where the ratio of nodes with n connections is proportional to $n^{-k}, k = 3$ in Barabási's book Linked [64], but k is only exact as the number of nodes tends to infinity. We also see increased heteroscedacity as the degree becomes very large.

In figure 3.4(a) the out-degree has a stronger negative relationship between the probability and the degree of the node, where as the reciprocated network follows nearly the exact same slope for both in and out, which is due to the removal of many of the low degree in-active nodes in the original network which skewed the data. The very large outliers in the in-degreee of the original network also emphasize this skew.

Both graphs have quite similar shapes and the original network has a very high amount of

nodes with in-degree of 1, having 76%, and the reciprocated network having no nodes with an in-degree of 1. This shows the majority of the networks nodes only receiving one tweet which gives more of a sense of why network efficiency is so low in the original network. With 91.1% of nodes having in-degree less than 5 and 85.7% having an out degree less than 5 in the original network. The reciprocated network provides more active and socially connected nodes with a lower number of nodes with very low degree. Many of these low degree nodes provide noise to the actual political discourse.

One can also see that many of the large outliers in the degree distribution are in the in-degree category. The maximum in-degree of a node in the original graph is 16,243 whereas the maximum out-degree is only 2,329, with many other in-degree nodes much larger than the largest out-degree. This nearly 7 fold difference is due to these hubs of inward tweets that facilitate conversation on the issue from the public [57]. In the reciprocated network, the max in-degree is 1,284 and the the max out-degree is 692, which is a large reduction in hub activity from the original network.

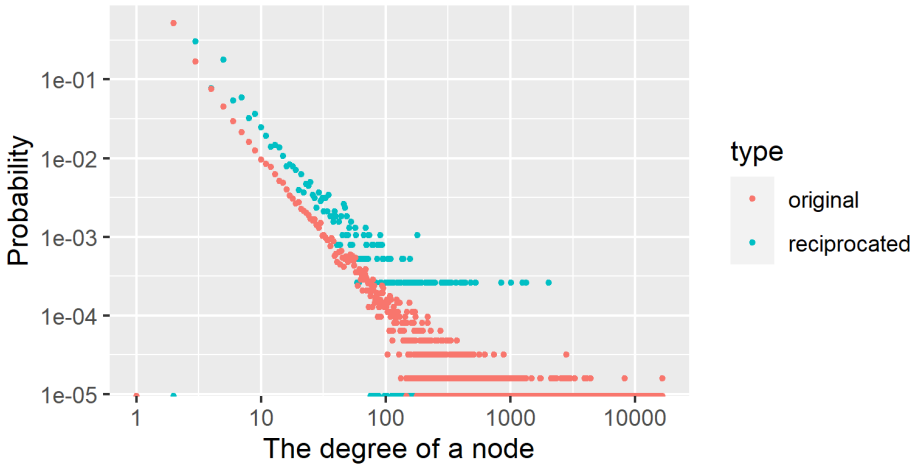


Figure 3.5: Total (sum of in and out) degree distribution of original network vs reciprocated network

As we can see in figure 3.5 the original has a high percentage of very low degree nodes (2) whereas the reciprocated network doesn't have any low degree nodes because many of the inactive nodes were removed in the reciprocity testing process. After these first values, the reciprocated network has higher percentages of medium to high degree nodes, so these active and well connected nodes are more prominent. The reciprocated distribution can be seen to tail off earlier than the original data set due to its smaller data set size and the reduction in maximum values (outliers). The distribution shape, however, is very similar for both networks.

The mode of both networks is a degree of 2, but for a one tailed distribution this does not

tell us much, the mean of the networks is a degree of 8.8 and 14.6 for the reciprocated network. The higher mean for the reciprocated network exemplifies the less negative slope of its degree distribution and shows the removal of many low degree nodes that occurred. The standard deviations for these graphs are 25.8 and 26.13 for the reciprocated network. The slight increase in standard deviation is due to the removal of a large number of low degree nodes but is kept at bay by the reduction in outlier size. This increase is minimal and the reciprocated network will show us more active nodes without the massive pull of outliers skewing the results which is promising.

With all of this analytical understanding of both networks from the summary statistics and graphs, it can be seen that the reciprocated network will be more efficient for understanding the mechanisms of this discourse with less noise from in-active users, so we will continue the analysis with this network.

3.3 Summary Statistics

Summary Statistics		
	Original Network	Reciprocated Network
Number of users	62,622	3,546
Number of tweets	480,201	28,056
Global Clustering Coefficient	0.015	0.054
Diameter	14	14
Average Shortest Path Length	4.62	4.6
Assortativity	0.018	0.08
Mean of Degree Distribution	8.8	14.6
Standard Deviation	25.8	26.13

Table 3.1: Table of summary statistics for original and reciprocated network

In the next section I will begin sentiment analysis on the reciprocated network.

Chapter 4

Sentiment Analysis

In this section we will work towards weighting the links of the reciprocated network by the sentiment of each tweet that link represents. This will allow us to analyse a weighted network, to see if sentiment (how positive or negative the words a person uses) clusters over the network.

4.1 Afinn Method

Method Selection

When approaching sentiment analysis the first thing one must do is decide what method to use to quantify the positive or negative feeling in the text. There are many methods to perform sentiment analysis, including supervised-machine learning and lexical based methods. SentiStrength is an example of a machine-learning based method, which uses classifiers which have been proposed in literature to identify sentiment gradients in data [65–67]. Gonçalves' comparative analysis of Sentiment algorithms showed a very strong correlation between SentiStrength's results and the ground truth of twitter discourse [68]. Although, I chose a method which defines a sentiment value who's magnitude signifies how positive or negative the language conveyed in the text is [69] using a lexicographical methodology, which has been shown to be one of the best methods to execute Twitter sentiment analysis, on par with SentiStrength [70]. The lexicographical approach can also be much more easily discernible to the human eye and they do not need a large training data set to learn from [71].

For this, I used the afinn method in R, developed and manually labelled by Finn Årup Nielsen [72]. It is a lexicon of 2,477 words each with an associated score, each accompanying score relevant to a word is added when they are present in the tweet, making an overall sentiment score for each tweet. This emotional valence is evaluated using an integer from negative three to

positive three for each word inside the dictionary [73]. Scores allow us to discern the emotional intent of the sentence in question [74] which can be used to analyse interpersonal relationships between users that meets the reader's expectations [75]. They allow us to gain an understanding of overall sentiment in sentences in most cases, although it can not exhibit many of the nuances of positive and negative emotion like disgust or serenity or sarcasm. The limitations of sentiment analysis requires us to be very careful and general when interpreting these results.

The afinn method has been used to develop a character to character sentiment analysis in a selection of Shakespeare's plays where it demonstrated that thorough emotion dynamics can be extrapolated [75]. I think the afinn method will be even more useful on the corpus I am using as it is designed for a modern dialect. It has also been used to understand positive or negative polarisation in Indian social media using direct translation of the data to English and then appropriate normalization [76]. The afinn method has been used to analyse disparities in journal articles and mainstream media reporting on chronic traumatic encephalopathy (CTE), in order to unearth any reporting bias on the issue [73]. Lastly it has been used to develop multi-level models to identify the main aspects by which customers rate a restaurant [77]. Afinn's extensive use in scholarly research gives plenty of evidence for its versatility and practicality to aid my attempt to perform a robust sentiment analysis.

	word	value
1	abandon	-2
2	abandoned	-2
3	abandons	-2
4	abducted	-2
5	abduction	-2
6	abductions	-2
7	abhor	-3
8	abhorred	-3
9	abhorrent	-3
10	abhors	-3
11	abilities	2
12	ability	2

Figure 4.1: Example of afinn's sentiment lexicon

Extracting sentiment is very difficult, afinn's lexicon is one of the most simplistic ways of approaching the task of quantifying emotive language but has been seen to be very effective and isn't as computationally expensive as more complex approaches. It also allows us to understand

the reason each tweet receives their designated score without any ambiguity. Given that sentiment is a notoriously hard signal to extract from data, we will aggregate multiple tweets in §4.4 to get an idea of a users average tweet behavior. In the next section I will prepare my data for sentiment analysis.

4.2 Data Cleaning

With very large data sets running the sentiment algorithm can take some time computationally due to the fact that the program must run through the corpus of tweets and it must check every word to see if it is in the lexicon. Removal of these non-discriminant words in the data set known as data cleaning can approve the algorithms efficiency . The ever growing scale of these data sets has increased the need for correct and efficient data cleaning, which has caused it to become a subject of great relevance in the academic community [78–80]. Although it is costly in terms of the analysts time keeping and effort [81] it allows for less time spent waiting for computation completion and is repeatable. This cleaning involves removing special characters, punctuation, numbers and 'stop words'. Stop words include abbreviations and words in an irregular form [82], in this case they are removed using a preset list of words, with some examples highlighted in Figure 4.2.

```
> stopwords("english")
  [1] "i"          "me"         "my"         "myself"     "we"         "our"
  [7] "ours"       "ourselves"  "you"        "your"       "yours"      "yourself"
 [13] "yourselves" "he"         "him"        "his"        "himself"    "she"
 [19] "her"        "hers"       "herself"    "it"         "its"        "itself"
 [25] "they"       "them"       "their"      "theirs"     "themselves" "what"
 [31] "which"      "who"        "whom"       "this"       "that"       "these"
 [37] "those"      "am"         "is"         "are"        "was"        "were"
 [43] "be"         "been"       "being"      "have"       "has"        "had"
 [49] "having"     "do"         "does"       "did"        "doing"      "would"
```

Figure 4.2: Example of 'stop words' used.

The removal of stop words is important in Twitter data because of the substantial volume of abbreviations and colloquialisms that are used and cannot be deciphered by the sentiment algorithm causes data sparsity which is computationally ineffective. There has been some criticism of this method of blanket and preset 'stop word' removal as it reduces the feature space and these words may hold emotional context in some cases [82]. This does not apply to our situation as these issues are only apparent in more complex sentiment analysis mechanisms, like machine learning methods that take into account more complex sentence structure (like bi-grams). Therefore, these non-discriminant words cannot possibly effect the afinn method, showing the imprecise nature of these methods, but aggregated over large data sets these discrepancies become less apparent and allow an overall view of emotion. With our methods limitations in mind we can now proceed to implementation.

4.3 Sentiment Calculation

After prepping the data we can begin sentiment calculation using the `get_sentiment()` function in the `syuzhet` package. The computed sentiment distribution is found to have a range from -11 to 14. The range of these values is positively weighted and the average sentiment of this network is 1.03 with a standard deviation of 1.87, which shows that the tweets of the reciprocated network have a tendency to be positive.

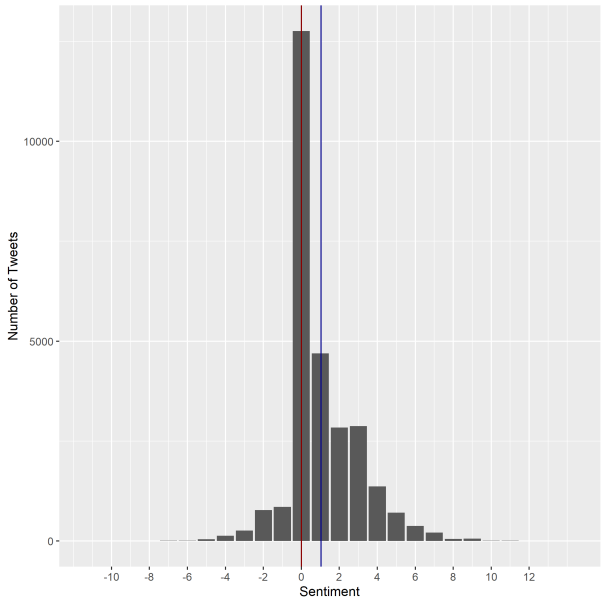


Figure 4.3: Sentiment Distribution for all Tweets,
red line = median @ 0,
blue line = mean @ 1.03.

The calculated mean and median can be seen in figure 4.3, red being the median of 0 and blue being the mean. We see more evidence that the graph is positively skewed as the mean is greater than the median with the peak at 0, just under half of the tweets having a sentiment score of 0 (45%). This positive skew centered at 0 has been seen in other sentiment analyses of Twitter data [83, 84] and even in general natural language analysis [85]. These studies saw a slightly less positively skewed distribution, but this may have to do with the polarizing nature of the marriage referendum and the prevalence of young and liberal social media users [86] providing the more positive discourse.

	Tweet text examples	Score
(a)	how about you **stop** being a **liar stop** being a **dishonest** person you grow up and get a **bloody** life vote **yes**	-9
(b)	well tough you are well able to be rude and **obnoxious**	-3
(c)	turnout in ashbourne 65 at 9 35pm	0
(d)	absolutel trojan work done by you today to spread the word and **good** will **thank** you	5
(e)	aisling **lovely** to meet you and **great** to hear your **support** for yesequality2015 **good luck** with show today	14

Table 4.1: Table of example tweets and their acccompanying sentiment scores
Red words = Negative words, Green words = Positive words.

The table above shows some examples of tweets and their associated sentiment score, these examples range from the most negative to positive tweets. This gives us a feel for how the afinn method works and how these scores relate to emotional intention. These tweets may be difficult to read due to their lack of punctuation but this is done in the previous section during data cleaning to assist the sentiment analysis. In table 4.1(a) we can see a very negative tweet with a score of -9, from visual examination without context we can see that it uses strong unhappy language. The afinn method can also discern this by characterisation of "`stop`", "`liar`", "`dishonest`", "`bloody`" and "`yes`" and initializing each with a sentiment score.

$$\text{Sentiment score } = -9 \begin{cases} -1 & \text{, “stop”} \times 2 \\ -3 & \text{, “liar”} \\ -2 & \text{, “dishonest”} \\ -3 & \text{,“bloody”} \\ +1 & \text{,“yes”} \end{cases}$$

The many disagreeable adjectives add together to give the most negative sentiment score in the corpus and they far outweigh the positive score "`yes`" provides. The entry in 4.1(b) is very similar to 4.1(a) in its unambiguously malevolent portrayal but it provides less adjectives for the afinn method to discern. The only word in the text present in the afinn lexicon is "`obnoxious`" with an associated score of -3. In 4.1(b) we see another limitation of the afinn lexicon, it's incompleteness, it does not contain all of the emotional words in the dictionary. "`rude`" is a word we would perceive as negative but the procedure has initialised it to be neutral, because it has no score saved in its lexicon for this word. This is another possibility for deviation from the ground truth of emotion in the language. The score relative to 4.1(a) seems fair as it is slightly less voracious in its complaining.

In the tweet 4.1(c) the afinn method cannot determine any discriminatory words from its lexi-

con and grants this tweet a score of 0. This seems accurate to the reader as the user seems to be stating a fact with no emotional intention. Nearly half of all tweets display this emotional neutrality or at least have equal positive and negative words.

In the tweets 4.1(d) and 4.1(e) we see only positive adjectives, these solely positive tweets have very different scores though with (d) receiving a score of 5 and (e) receiving 14. The more complex sentence structure of (d) is not understood as positive by the lexicon. "`absolutel(y) trojan work done by you`" can easily be seen as a compliment with positive emotion to a reader but it does not have any common adjectives which would be in the lexicon and so that section of the tweet is given an emotionally neutral score. The rest of the tweet is quantified by "`good`" (+3) and "`thank`" (+2) to give the overall score of 5. A more complex analysis method may be able to recognize these nuances of emotion more accurately. The last sentence (e) reads nearly like an extensive list of positive descriptors and is understandably rated so highly. The emotional identifiers "`lovely`" (+3), "`great`" (+3), "`support`" (+2), "`good`" (+3) and "`luck`" (+3) make this the algorithms most positive tweet with a score of 14. As we can see in this section sentiment analysis may not be a perfect representation for each tweet individually but aggregated over all interactions this may give a more accurate depiction of user sentiment.

4.4 User Sentiment Analysis

We begin analysis of user sentiment by graphing the spread of user sentiment,

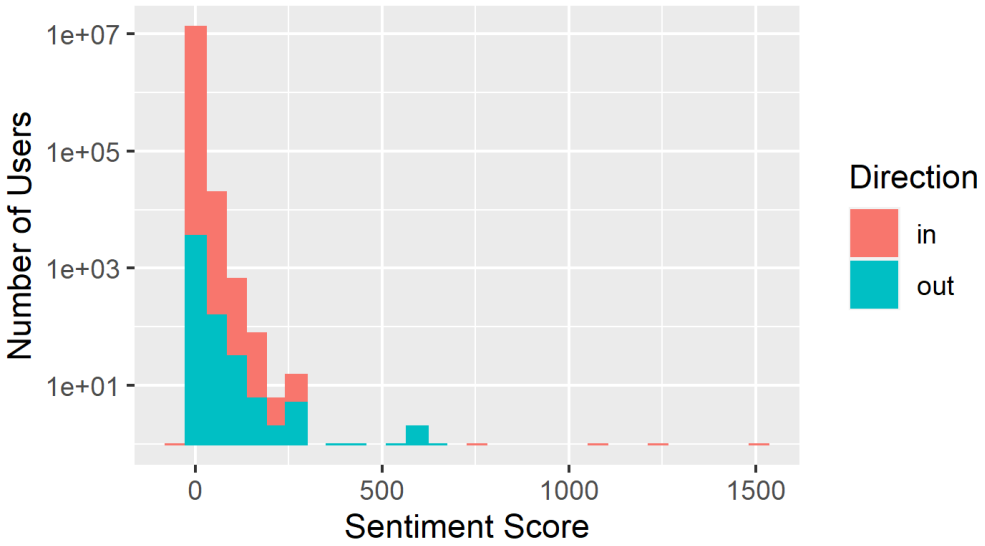


Figure 4.4: user in and out sentiment scores.

In figure 4.4 of the nodal aggregation of user sentiment, we see a very similar spread for in and out sentiment scores with a high density between 0 and 50. Large outliers are more prominent in the in direction as both the maximum and minimum sentiments in the graph relate to an inward sentiment score. The maximum in-sentiment is 1530 and the minimum is -34 and the maximum and minimum out sentiments are 673 and -15, these large ranges are heavily skewed towards a positive sentiment score. The overall positive skew of the network leads to the aggregation of scores in most cases being positive for in and out sentiment of nodes.

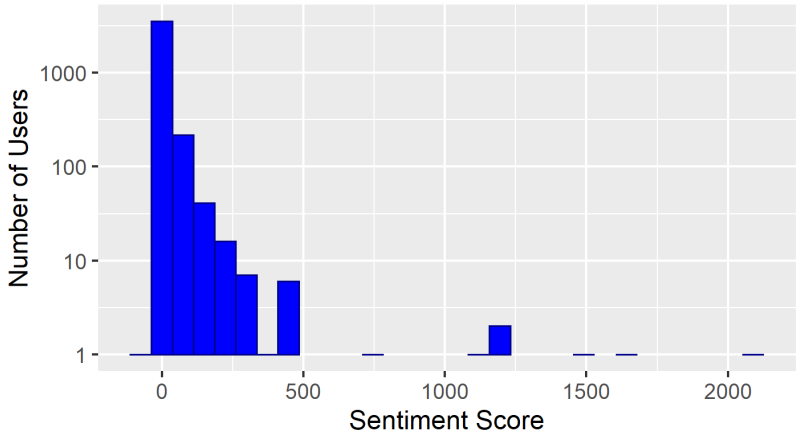


Figure 4.5: Total user sentiment

Figure 4.5 has a very similar distribution to the in/out user sentiment graph 4.4, which provides evidence for the presence of a correlation between the aggregated inward sentiment score and the outward score. The upper bound of this distribution is 2103 and the lower bound is -42. These scores are both of a higher magnitude than the upper and lower bounds of figure 4.4. Therefore the minimum total sentiment is a combination of 2 highly negative sentiment scores and the maximal total user sentiment must be a combination of the highest inward and outward score. These show more clues that there is a positive relationship between inward and outward user sentiment.

The mean of the distribution in figure 4.5 is 15.06. This strong positive average can be seen in the density of nodes around 10. The median is only 4 showing this distribution has a strong positive skew caused by the very large positive outliers.

This aggregated nodal sentiment will be very useful in the next section (§5) where we use Monte-Carlo simulations to test the interactions between nodes and gain a better understanding of how sentiment plays a part in these interactions.

Chapter 5

Monte-Carlo Simulation for Sentiment as a Proxy for Homophily

In this chapter we will test for the presence of homophily in the Irish Referendum Twitter discourse using Monte-Carlo simulations of sentiment alignment as an intermediary for homophily.

5.1 Monte-Carlo Simulations Background

After calculating the sentiment corresponding to each tweet we must begin analysing the interactions between users more comprehensively using total in and out sentiment for each user. This analysis hopes to understand the level of homophily in the network which indicates if these users follow the proverb "birds of a feather flock together" [32]. This well researched, robust regularity in societal behavior will be tested using Monte-Carlo simulations of the sentiment scores for this network topology. Monte-Carlo simulations are used in problems that are analytically unmanageable and experimentation is impractical [87]. The method of simulation uses statistical modeling of generated random samples to estimate complex systems and relies on the fact that if a qualitative property has an effect on the system it will create a definite variation from chance. My rationale to use sentiment conformity as a proxy for hompohily is that users with comparable inclinations towards the referendum may communicate with similar attitudes. For example, yes voters may use more positively charged expressions like "`vote yes`" and other optimistic phrases which can culminate in an elevated user sentiment scores [5], because "`yes`" gives a sentiment increment of +1.

Monte-Carlo testing is required to negate the effect of random variables in testing the probability of arriving at the obtained results by chance. Unfortunately Monte-Carlo methods are not standard in R and must be coded by the analyst. The Monte-Carlo method of "psuedo-

sampling" can be very computationally expensive [88]. To understand the importance of a result of network analysis attained in previous sections we must undertake a comparison of the empirically measured network against the randomized networks to attain a Z value which will tell us the significance of our results using a test statistic [89]. One of the most common issues in Monte-Carlo simulations is knowing which values to randomize and which to keep constant before beginning pseudo-sampling. In our case the choice is obvious as we are trying to use sentiment alignment as a proxy for Homphily, so sentiment randomization is required.

To begin simulating we must have an empirical data set [89], which in our case is the reciprocated network topology found earlier. This empirical network's topology must be kept but weights are to be removed, we then randomize the link weights of this network, which were defined as the sentiment scores of tweets, within the reasonable bounds of the original model and run tests on each randomly weighted network to create a distribution of results which can be analyzed against our obtained results. In the next section we will begin simulations.

5.2 Testing Pearson Correlation between average user sentiment in and out

In this section, I begin testing the empirical network by finding the Pearson's correlation between the aggregated in and out sentiment scores. This gives an understanding of the relationship between emotional language users use in their tweets and what they receive. This seems like a vital insight into these interpersonal relationships and a step closer to understanding the level of homophily present. This measure of correlation doesn't mean a great deal without being verified to be different from random given the current topology, showing the need for these randomization simulations.

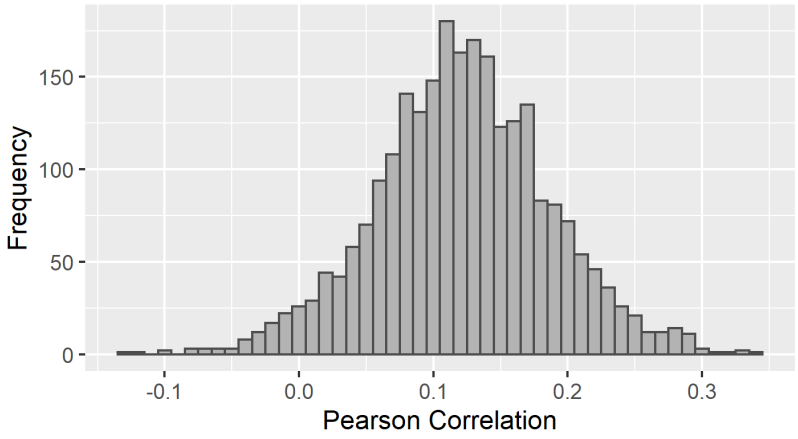


Figure 5.1: Distribution of Randomized networks Pearson correlations

Pearson's correlation coefficient quantifies the strength of linear relation between two variables [90]. The calculated correlation of 0.804 for the empirical network is a very strong positive association, indicating that as the average inward sentiment score increases so does the outward score and vice versa.

Some Monte-Carlo simulations of large networks can have 25,000 randomization trials [88] but due to my lack of computational power I capped trial number at ≈ 10% of total number of nodes, 2500 trials. This weight randomization was made and for each randomization the Pearson correlation of user sentiment in and out was calculated. This distribution of Pearson correlations was then collated and presented in figure 5.1, seen to be normally distributed with a mean near 0.1. Monte-Carlo simulations can produce many desirable distributions like this, exponential, Chi-Squared and many more, it is a useful tool when regression assumptions may be broken otherwise [88]. The randomized samples show a mostly positive correlation also but this is not very strong in any cases.

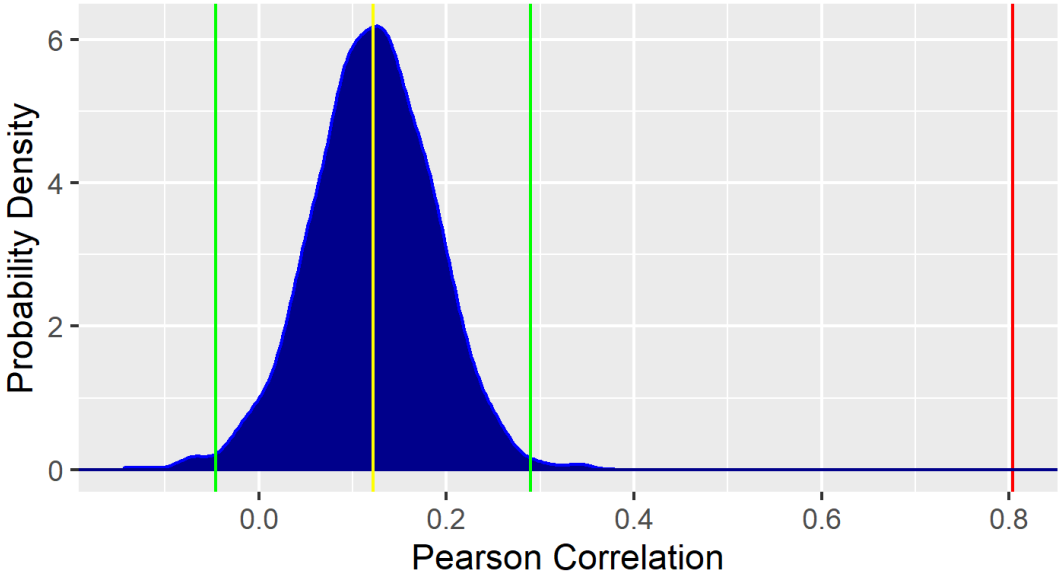


Figure 5.2: Probability Density Monte-Carlo Simulation of Pearson Correlation
yellow line = mean
green lines = ±2.58 standard deviations (99% confidence interval)
red line = correlation of empirical network

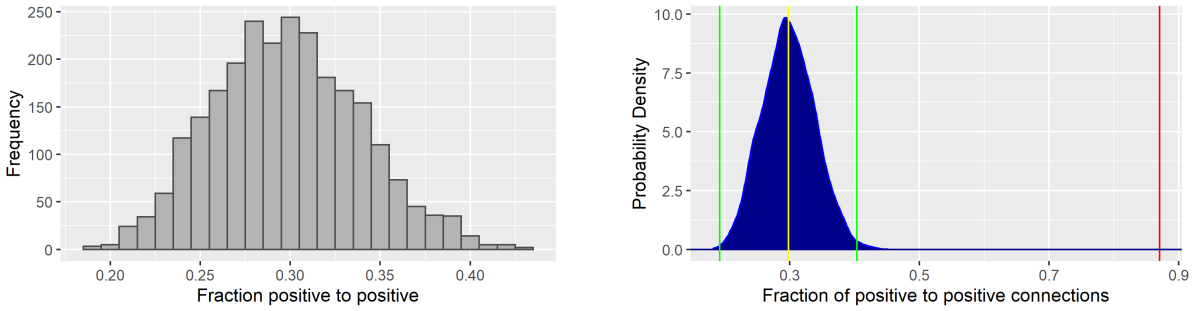
Figure 5.2 shows a more detailed look at the randomized sample distribution, giving the point of its mean at 0.122 and the standard deviation of 0.065 gives us the 99% confidence interval at ±2.58 standard deviations. This slightly positive mean shows a very slight positive association between the aggregate sentiment in and out of each node in the random network. If network topology had no effect on the relationship between in and out correlation the distribution would

be centered around 0, we can see from visual inspection that the correlation distribution doesn't seem to be centered at 0, exemplifying the increase in positive association caused by network topology. Although, the Pearson correlation of 0.804 is much more positive than the center mean of the null model. It is outside the 99% confidence interval of the randomized distribution, showing the significance of the relationship between the user's in and out sentiment compared to the null model.

We have shown a non-trivial association between what users tweet out and what they receive and in the next section we will explore the the relationship between positive to positive links and negative to negative links.

5.2.1 Simulation to test the fraction of positive to positive links in the network

In this section I will test the significance of the fraction of positive to positive links present in the empirical network compared to the null model. A positive to positive link is one which has a overall positively weighted node on both ends of its link. This can give us an understanding of the level of homophily present in the network as it is a measure of the proclivity of nodes with similar sentiment to link to each other.



(a) (b)

Figure 5.3: (a) Bar plot of the fraction of positive to positive links in the null network (b) Probability density of the fraction of positive to positive links in the null network, yellow line = mean(null network), green lines = ± 2.58 standard deviations (99% confidence interval), red line = fraction of positive to positive links in emipirical network

Figure 5.3(a) shows that the fraction of positive to positive links in the null model follows a normal distribution, centered at 0.298 with a standard deviation of 0.041. This distribution is close to a value expected when randomizing with 3 categorical variables positive, neutral and negative. The fraction for the empirical network is much higher than this at 0.870, outside the 99% confidence interval. This high level of significance shows that the communication between

positive nodes is significantly different to the null model, demonstrating a non-trivial result for the presence of homophily in positive users in our network.

In later sections we will delve deeper into the topic of homophily and try to uncover the presence of community behaviour in the network.

Chapter 6

Analysis of the existence of communities

Community formation is of utmost importance because it reveals the internal configuration of the nodes in real networks [91], they are also believed to play a fundamental function in the operative properties of complex networks [34, 35]. For analysts understanding network community behaviour can give you a powerful visual aid when graphing and often times these communities can be treated as if they were themselves nodes in a substantially smaller network with their links re-scaled by inter-community connections for ease of analysis [92]. These techniques can provide visually optimal network lay outs. The presence of communities is easily understood and evident in social networks, scholars have noticed their presence for decades [37–39, 93, 94], and these organizational patterns of humans have been studied in sociology thoroughly [95–97]. Classifying community structures isn't only important in sociology it has been an important activity in bibliometrics [98], biochemistry [99] and computer science [100]. In the development of efficient search engines community detection has been utilized to help users find what they are looking for and avoid other content on the world wide web information system [101].

We expect the presence of community behaviour in this network but we must do further analysis to achieve quantitative results using a robust community detection algorithm to confirm these community interactions.

6.1 The Prevalence of Communities in Social Networks

Community behaviour is a staple of the characteristics of social networks. This can be observed in our everyday lives as we can see there is a visible interconnectedness between people working in the same job or with similar hobbies. This clearly evident phenomena has been shown analyt-

ically, as an example individuals working in a certain company are more likely to interact with individuals employed by a different company [94]. An easily understood example of community detection in a social network was shown in the seminal paper by Zachary on the members of a karate club [39]. The club contained 34 members so it can be very easily understood by inspection, links in the network were added if the members had regular interactions outside of the club. What makes this network so interesting is the definite known community structure caused by a conflict between the clubs president and a trainer, where half of the members followed the president and the other half followed the trainer splitting the club in two [36]. Zachary's

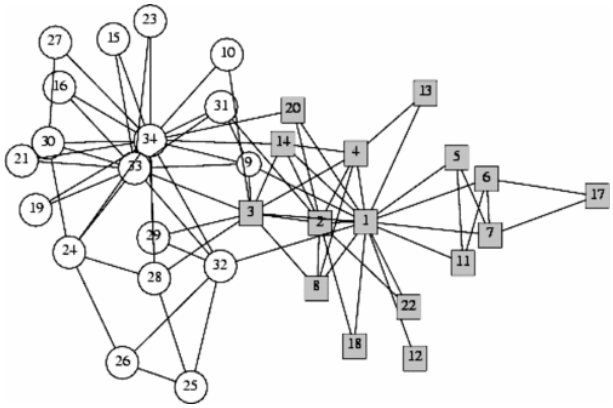


Figure 6.1: Community Structure of Zachary's karate club [42].

karate club paper was relatively untouched from its inception in 1977 until Girvan and Newman's important 2002 paper [34] made use of the naturally formed and obvious community structure Zachary had uncovered. This hence-fourth became the most common benchmark test for community detection algorithms and exploded in popularity. This gives an understanding of what we are trying to achieve in community detection, a deeper understanding of the topology and qualitative characteristics of the network, which we cannot achieve from inspection of our network because of its size (3546 nodes, 94039 edges)

6.2 Community detection algorithms

In this section we will gain an understanding of how and why community detection algorithms work.

6.2.1 Defining Communities

A community is formally known as a locally dense connected subgraph of a network [36]. The connectedness requirement means all nodes in a community must be possibly reached through a link from another member of the community and the density requirement defines a community member should have a higher probability to link to other members of that community than

to members of other communities. One of the earliest papers written on community structure defined a community as an accumulation of members who all have mutual links with each other [37]. This definition is very strict and is no longer being used in research to allow for more flexibility in community detection, these maximum link density subgraphs are now known as cliques or complete subgraphs. Although, a rigorous conception of community structure is still ambiguous, in general, a complex system's community structure incorporates a collection of hierarchical and modular components [92]. This lack of an accepted and quantitative universal definition of community to be implemented into algorithms leads to inherent problems in the translation of the output if there is no non-topological information provided [102]. Thankfully in our data set we have copious amounts of non-topological edge and vertex attributes to help with understanding of our algorithmic outputs, which will effectively provide a ground truth for comparison (whether members are voting yes or no).

6.2.2 Modularity

Modularity is a measure developed a posteriori which can begin to give us a more quantitative understanding of how to detect communities and the apparent effect they have on the network topology in turn making the concept of community detection more clear-cut and useful [99]. The measure introduced by Newman and Girvan [103] does this through being a quantitative measure of the community structure appended to the given network, there have been other quantities proposed for this purpose [35, 102, 104] without the same universal use. The modularity is inspired by the fact that in a topologically random network the connection pattern of edges is expected to be uniform, irrespective of the degree distribution. Therefore, random networks are not predicted to demonstrate systematic density variations which could be construed as community structures, giving us the ability to define a measure of the deviation from a random configuration to measure the quality of these partitions against a null model [36]. To define modularity we must define some parameters of the system first, consider a network which has already partitioned into n_c communities with L_c edges and N_c nodes in the selected community C_c. If L_c is greater than expected from the degree distribution, then we have evidence to suggest that the partition C_c could be part of a real community. Therefore, we measure the modularity using the difference between the network's real wiring diagram A_{ij} and the expected number of links between i and j if the network is randomly wired p_{ij}, [36]

$$M_c = \frac{1}{2L} \sum_{(i,j)\in C_c} (A_{ij} - p_{ij}), \tag{6.1}$$

p_{ij} is calculated using the degree preserving null model,

$$p_{ij} = \frac{k_i k_j}{2L}. \tag{6.2}$$

The degree preserving null model is possible because in randomisation of the link layout the expected degree of each node remains unchanged, allowing for easy calculation of expected value which is highly correlated with the relative degree of the nodes, following our previous assertion on the increased likely hood for hubs to attach to each other. We can see that this measure essentially quantifies the differnce in the number of links inside a community with the expected value of a graph of the same number of nodes and degree sequence but a random layout [105]. If the community modularity is positive the C_c has more links than expected by chance, consequently it provides evidence for the presence of a potential community. If $M_c = 0$, the connectivity between the nodes in our partitioned community is fully described by the degree distribution and is random. If M_c is negative, the nodes of C_c do not follow our definition of a community. The definition of community modularity can be simplified to,

$$M_c = \frac{L_c}{L} - \left(\frac{k_c}{2L}\right)^2 \tag{6.3}$$

through analysis of the effect of merging communities, with k_c being the total degree of nodes in the community and L being the total number of links in the network. The total modularity of the partition is then achieved by summing over all of the community modularities [106],

$$M = \sum_{c=1}^{n_c} \left[\frac{L_c}{L} - \left(\frac{k_c}{2L}\right)^2\right]. \tag{6.4}$$

The modularity of a partition is a value between -1 and 1, in our search for the best community detection in our network we try to achieve a high modularity, i.e an optimal partition. The main algorithm we will use for our analysis is a modularity optimization algorithm using this measure systematically to try and uncover the actual partitions of the data.

6.2.3 Limitations of Modularity

Modularity has its own limitations which has caused scholars to search for other measures of partition quality [35,102,104], the first of these is resolution limit. The modularity maximisation of a network forces relatively small communities into larger ones [107] when they meet a certain criterion. If we suppose A and B are two distinct communities in a given network the modularity maximisation should allow them to remain distinct, although when,

$$\frac{k_A k_B}{2L} < 1 \tag{6.5}$$

the modularity change is $\Delta M_{AB} > 0$, if there is one or more links between the two communities ($l_{AB} \geq 1$). In a simple case, if $k_a \sim k_b = k$ the modularity will be maximised by merging the two communities when

$$k \leq \sqrt{2L}. \tag{6.6}$$

The premise of this phenomena is that if k_A and k_B are under the threshold 6.5 or 6.6 the expected number of links between them is smaller than one. Even one shared link between A and B will then break this expectation causing merging of possibly distinct communities in modularity maximisation. Real networks have numerous small communities [108–110], which are forced into larger communities, creating a misleading characterisation of the community structure. To avoid the resolution limit large communities can be subdivided after modularity optimisation [105, 111, 112].

Finally as we choose modularity as our relevant quality function in community detection of the reciprocated marriage referendum data set, our problem of community detection becomes commensurate to modularity optimisation. Unfortunately, modularity optimisation is not computationally cheap as the number of possible partitions in a network increases at least exponentially with the size of the network [105]. We must therefore use a heuristic search method to reduce the search area while preserving the goal of optimisation. The fastest available procedures avail of greedy techniques [111, 113] and extremal optimisation [114–116], whose reduction in computational cost is made necessary by our large network size. In many cases the efficacy of these optimised algorithms is verified using the aforementioned Zachary karate club [99, 111, 114, 117] or against computer generated graphs that have a established accepted community structure [34, 114]. Although, our case we will attempt to reveal the users political leaning using an modularity optimisation method.

6.2.4 The Louvain Algorithm

We begin looking at community detection algorithms with the Louvain algorithm, introduced by Blondel has been shown to outperform all other community detection algorithms before it in terms of computation time [118]. This efficiency makes it very scalable to large networks, which is above 2000 nodes as a rule of thumb. Its method consists of 2 steps repeated iteratively, before the iteration begins the method assigns each node to a different community. The Louvain method has two distinct steps which involve maximising the modularity change by moving a node into the optimal neighbouring community and then creating a new network who's nodes correspond to each community and their link weights are the sum of the links between communities, this network is then optimised in the same way as the first step.

Step 1

For each node we measure the gain in modularity from the reallocation of this node i into the community of one of its neighbours j. This modularity change is defined by [118]:

$$\Delta M = \left[\frac{\sum_{in}+2k_{i,in}}{2W} - \left(\frac{\sum_{tot}+k_i}{2W}\right)^2\right] - \left[\frac{\sum_i n}{2W} - \left(\frac{\sum_t ot}{2W}\right)^2 - \left(\frac{k_i}{2W}\right)^2\right]. \tag{6.7}$$

We check all neighbouring communities and then move node i into the community which increases the modularity the most. If no positive modularity increase can be found, i stays in its original community. This process is repeated for all nodes until no further improvements can be found.

Step 2

In step 2 Louvain's algorithm a new network is constructed whose nodes are the communities which have been identified in step 1. The weights between each node in the new graph is the sum of the weight of the links between the nodes in the corresponding communities. The links between nodes in communities in the original network are summed to create weighted self loops of each node.

When step 2 is completed we continue to iterate the procedures until a pass through the procedure causes no changes, i.e., maximum modularity has been attained. On each pass the number of communities decreases. The computational complexity of the algorithm is at worst $\mathcal{O}(L)$, meaning the number of computations scale linearly with the number of links L in the first iteration of procedure and then with diminishing number of links on subsequent passes.

6.3 Community Detection

To begin community detection I implement the Louvain method on the reciprocated network using the igraph package. Resulting in a partition into 13 communities, with a wide spread in community sizes from 5 members to 844. The partitions overall modularity was 0.257, which is quite high for a network of this size that we just postulate to have a hierarchical structure. This is of comparable scale to the Zachary karate club example with a maximal possible modularity of 0.412, a network with a known and evident partition.

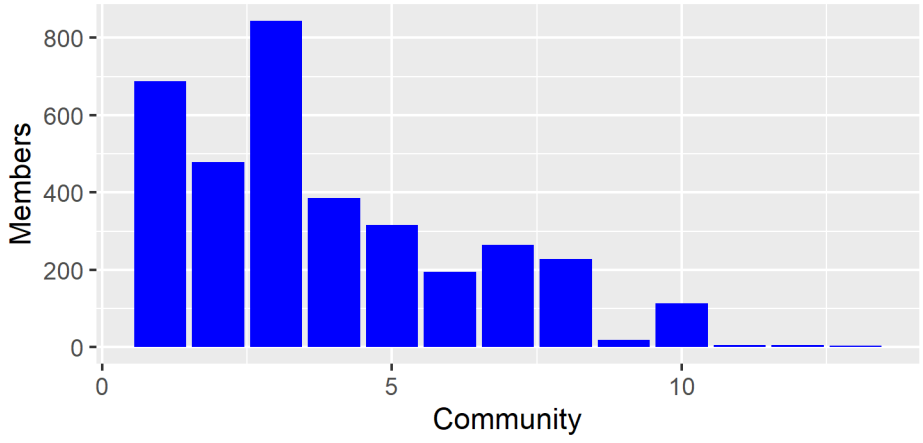


Figure 6.2: Community sizes

This distribution of sizes show some large communities that are dominating and some very small communities which contain less than 20 members. Many studies have reported similar results of small communities coexisting with a few very large ones [102, 108, 111, 113, 119]. Plotting community detection in a small graph like the Zachary karate club 6.1 shows an obvious structure that is present. Although our large network of 3546 nodes isn't as visually fourthcoming with its qualitative properties.

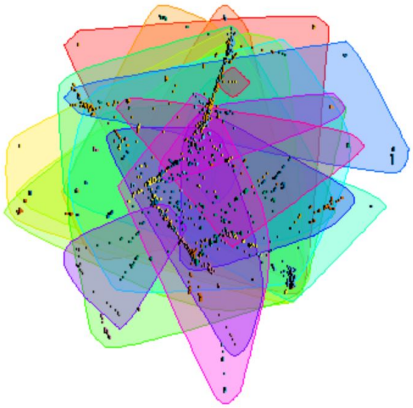


Figure 6.3: Graph of community structure without edges

Figure 6.3 doesn't tell us very much about the community structure present, the overlapping communities makes our representation very messy. To try clarify the properties of this image we must look at the edge weights of the links, which we have previously referred to as the sentiment of the tweets. This information did not play a role in the partitioning of the network by the Louvain algorithm, the purely topological features of an unweighted version of the network were used. As these communities arise from conversations between users the communities tweet sentiment can be used to further categorise members.

6.4 Community Grouping using Sentiment

To create a more tangible understanding of our communities, we attempt to group communities in the partition. Grouping is instigated by sentiment scores, as we have considered sentiment a proxy for homophily between users, setting a precedent for sentiment to illustrate similarity between communities [5]. We begin by assigning the edge weights in the network and aggregating these over the amount of sentiment entering a node in a community and leaving a node in that community.

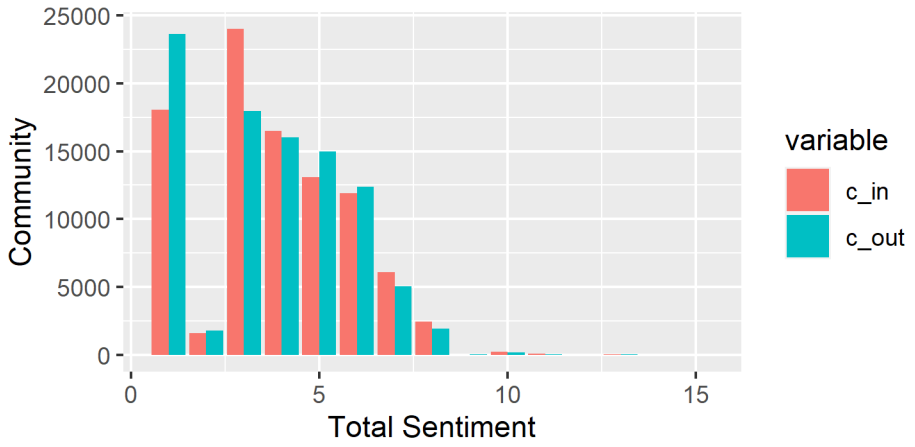


Figure 6.4: Total Sentiment in and out of each community

On it's own, this measure of sentiment between communities doesn't tell us much as it is not scaled with the variation in community size. Although, the smaller communities are expected to tell us very little because of the noise present in sentiment scores of tweets [120, 121], which reduces our abilities to produce rigorous statistical analysis of these communities. The random effects of noise will become less important in large communities.

6.4.1 K-means Clustering

Our grouping method uses the k-means clustering algorithm, which takes our previously calculated total in and out for each community as one of its inputs. K-means is the most popular and one of the simplest forms of clustering algorithm, which aims to partition the data into disjoint subsets known as clusters such that the clustering error criteria is minimised [40]. We must also provide the number of centroids the algorithm should use, where centroids define the number of clusters in the data. The number of centroids can be found using the elbow method:

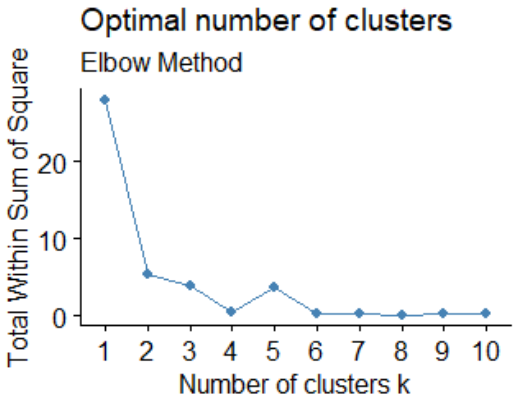


Figure 6.5: Results of elbow method.

The elbow method takes the sum of squared euclidean distances between the points denoted by Sentiment in, Sentiment out of each community and the centroid in that cluster. The maximum possible value is always at K=1, but we want to find when there is a drastic change in slope [122]. A sharp increase in slope will look like an elbow in the graph and in our case this is present at two, giving an optimal output when there are two centroids.

To begin this algorithm the centroids are placed at random locations and then undertake an iterative process until convergence or the user set limit of iterations is reached.

Step 1

For each point, $x_i, i = 1, ..., n$, n being the number of communities, we find the nearest centroid, c_j, using the euclidean distance between x_i and each cluster center, c_j,

$$arg\min_j D(x_i, c_j). \tag{6.8}$$

The cluster j providing it refers to the minimum distance is then assigned to the point x_i, completing step one.

Step 2

For each cluster, $c_j, j = 1, ..., K$, K being the calculated number of clusters, we calculate the mean of all points in the given cluster for each dimension and assign this mean to the centroid:

$$\begin{aligned} c_j(a) &= \frac{1}{n_j} \sum_{x_i \to c_j} x_i(a), \\ a &= 1, ..., d \end{aligned} \tag{6.9}$$

d in this case refers to the number of dimensions in the system, two in our case. The new centroid, c_j, is now used in the previous step until either the method causes no change in the cluster layout (convergence) or the maximum designated iterations has been reached.
The computational time to reach convergence is $\mathcal{O}(\#iterations * \#clusters * \#instances * \#dimensions)$, which in our case reduces to $\mathcal{O}(2 * 13 * 2 * \#iterations)$. This is invariably just $\mathcal{O}(n)$, where n is the number of iterations, meaning this method has an incredibly fast rate of computation versus other clustering algorithms. However, k-means' simplicity comes with its own drawbacks as it suffers from the fact that its performance is dependant on the initial starting conditions [123]. To attempt to correct this problem, many other procedures have been developed that employ stochastic global optimisation procedures, unlike k-means' local search method. Examples of these methods include simulated annealing and genetic algorithms. These alternative techniques described have also become the basis for many community detection algorithms [124–126]. Although these techniques are not widely accepted, k-means with multiple

restarts for variation in initial conditions is much more widely instituted [127], in our case the relative simplicity of the 13 node k-means test in two dimensions does not require multiple restarts.

6.4.2 K-means Results

Applying k-means to our data has partitioned the data into 2 clusters depending on each communities total sentiment in and total sentiment out. The first partition contains five communities, with 2634 members and it has an overall more positive sentiment than the second cluster which contains 8 communities and 929 members. Giving the partition:

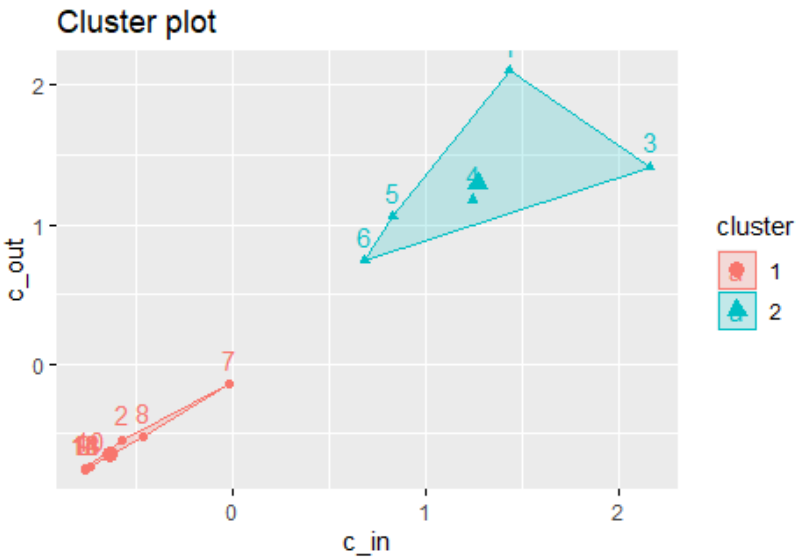


Figure 6.6: Results of k-means clustering

The evident split in the graph shows a jump in sentiment between the 2 clusters, with the negative cluster showing a much smaller variance between in and out values compared to the positive cluster. Now we try to understand if this clustering shares a relationship with the tweeter's views on the referendum.

Chapter 7

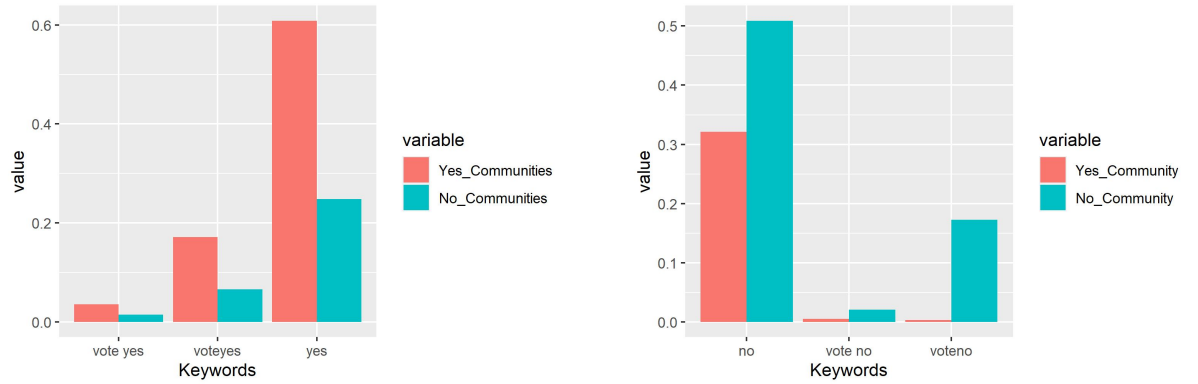
Analysing Yes and No Affiliation

To begin analysing the community members political leaning we begin by trying to understand if the clustering is in any way representative of yes and no supporters. We use two different methods of testing this hypothesis by calculating the density of keywords and through stratified random sampling to see if we can identify actual yes and no affiliation with our model.

7.1 Keyword Density

We begin analysis of our partitions by calculating the prevalence of some of the most important tags used by members of the opposing viewpoints. The polarising effect of this vote causes the users to use the hashtags "`voteyes`" and "`voteno`" which identifies their political leaning in their tweets. We also expect that campaigning tweets would contain appeals to "vote yes" or "vote no" depending on their stance. This allows us to identify keywords for each side of the debate which we can analyse as indicators for each cluster. This method of keyword density analysis has been used in the visualisation of theme progression in novels [128] and to assess student learning in school [129]. The main use for this concept is in search engine optimisation, where it effects a site's visibility in search engine results pages [130]. We calculate this density by dividing the frequency of keywords in the cluster by the total number of tweets in the cluster.

The keyword density in 7.1(a) shows a very large disparity between the two cluster. The positively skewed cluster showed a more than twice as large density of the keywords "`voteyes`" and "`vote yes`", which gives evidence for the increased likelihood of yes voters in this cluster relative to the other. Figure 7.1(b) follows the same pattern, with the negative community displaying much higher rates of the negative keywords "`voteno`" and "`vote no`". The density of "`voteno`" is 64 times greater in the negative list than the positive and the "`voteyes`" density is 2.5 times



(a) (b)

Figure 7.1: Keyword densities of both communities with (a) analysing positive keywords and (b) analysing negative keywords

larger for the positive community. The prevalence of these hashtags is a great indicator for sentiment to be dependant on political leaning. These findings allow us to denote the positive community as the yes leaning community and the negative as the no community. We can now test this allocation further using a more rigorous sampling method.

7.2 Ground Truth Testing

To further study this relationship between sentiment and affiliation, we take a random sample of 358 members of the reciprocated network. These were manually classified as yes or no affiliated to test our claim that the positively skewed cluster contains yes leaning voters and the negatively skewed cluster contains no voters. This is a sizeable representation (10%) of our data set so it should achieve a good representation of the overall structure. To classify each user, their Twitter biography and their catalogue of tweets in our data set were analysed. If an account had no obvious affiliation discernible from the information we have they are placed in unaligned and not tested against our partition. In many cases the accounts with no obvious leaning are impartial journalists or institutional accounts which do not give explicit opinions on the events leading up to the referendum, instead they become a vector for discourse between the polarised public. The removal of seemingly unaffiliated members of the network reduced the sample size to 308 members, 8.7% of the network.

The left-leaning nature of twitter can be seen in our analysis with only 26 out of 308 members being designated as no voters. With these classifications we examine if they match the community cluster layout of the network using a confusion matrix.

	Predicted Yes	Prediced No	Actual Total
Actual Yes	219	63	282
Actual No	0	26	26
Predicted Total	219	89	308

Figure 7.2: Confusion diagram of sampling classifications versus clustering classifications

We can see from the confusion diagram that our clustering configuration leads to a very effective identification of no voters in the population, with all of the no voters in the sample having been identified correctly. Although, the sample of yes voters does not have the same level of accuracy with 63 out of 282, 22% of yes leaning tweeters misclassified. Overall 245 out of our 308 sample points have been classified correctly, giving an 80% overall classification accuracy, which is worse than the null-error 91.6%. The null-error is calculated by assigning every variable to the dominating characteristic, yes alignment. These metrics for testing give too much power to the dominating yes voters so we calculate a balanced accuracy which can give us a more uniform understanding of our classification. The balanced accuracy of our classification is:

$$\frac{1}{2}\left(\frac{219}{282}+\frac{26}{26}\right)=0.89 \tag{7.1}$$

The balanced accuracy gives a better outlook on our model but it still under performs the null-error slightly. This is because the no-side is being over represented in our clustering. This could be due to the method used in k-means clustering, our clustering is based solely on the sentiment in and out of the communities. This sentiment is not scaled to the relative size of each community, which would not be an issue if the mean sentiment was zero, but there is a slight positive skew in the data giving a mean sentiment ≈ 1. This overall positive causes sheer volumes of people to outweigh the relative sentiment of these communities meaning communities relative size has a larger impact on their cluster placement than their tweet sentiment in some cases. To rectify this I propose a re-scaling of the system.

7.3 Re-Scaling

To begin re-scaling I divide each total in and out community sentiment value by their relative community size to achieve an average per node in the community. These average values are then used for the k-means process and cluster testing previously implemented.

Scaled k-means

The first procedure is to ensure there is no change in the required number of clusters from this scaling of the variables. The elbow point at two is slightly less pronounced after scaling but

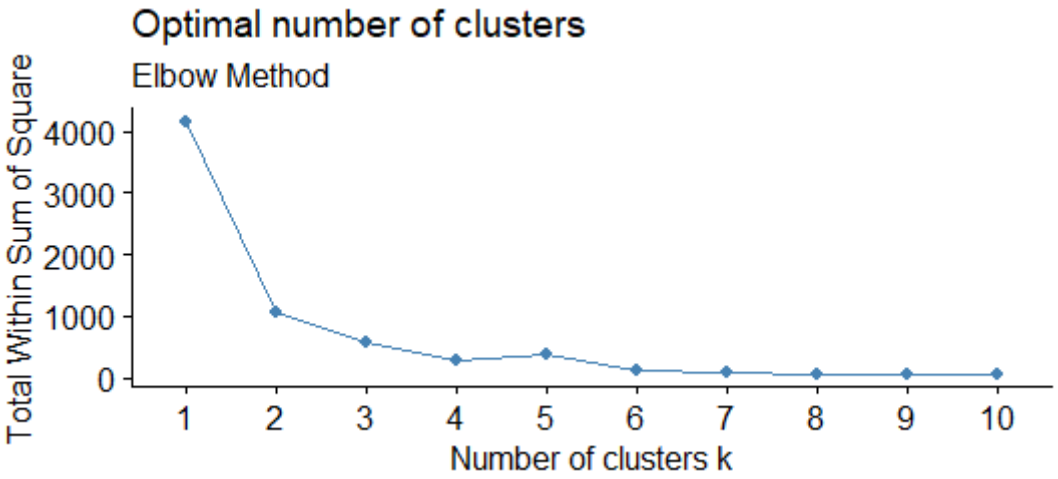


Figure 7.3: Elbow method after scaling

the optimal cluster number is still the same as previous iteration. After running the k-means

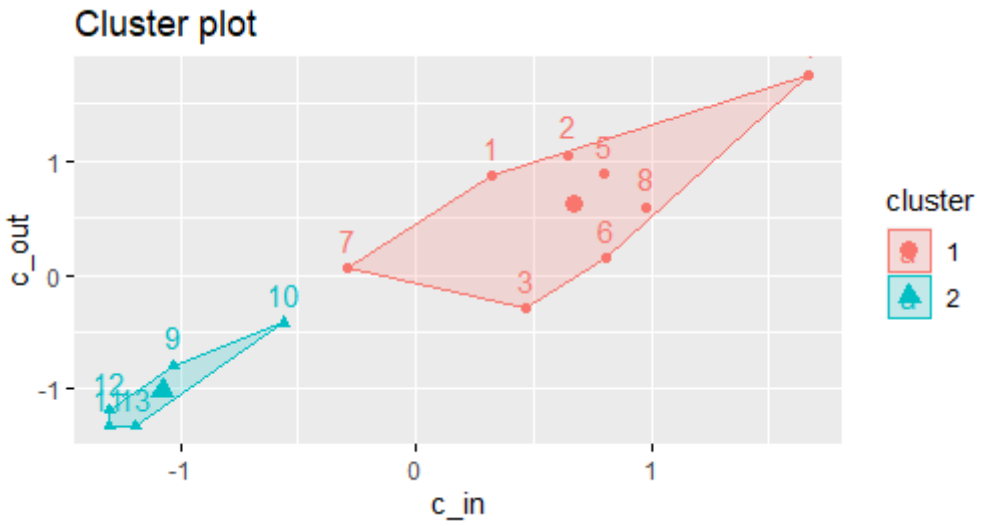


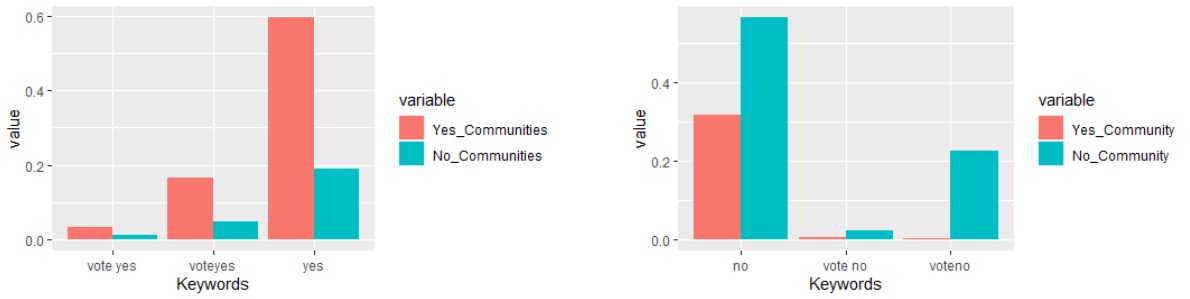
Figure 7.4: K-means after scaling

algorithm the new positive cluster contains more communities which is more indicative of the size difference between the yes and no side seen in random sampling. There is still an issue with smaller communities being dragged into the no group because of their relative inactivity in comparison with the bigger clusters. The statistical noise of sentiment scores [5] also make it so these small clusters give us very little information irrespective of their affiliation. Next we will

see how this scaling affects how well our model mirrors the sampling data.

Keyword Analysis

The new model brings a change in the keyword density, but keeps the overall qualitative structure of the positive partition having a much higher density of the yes keywords and vice versa. Our



(a) (b)

Figure 7.5: Keyword densities of both communities with (a) analysing positive keywords and (b) analysing negative keywords

previously seen disparities between keyword density have increased with the positive cluster "`voteyes`" density going from 2.5 to 3.5 times larger than the negative cluster and for the "`voteno`" density the negative cluster contains 94 times the positive cluster. This increase in keyword discrepancy shows a more definite difference between these communities which we can further inspect using random sampling.

Ground Truth Testing

We hope that this increase in variation between clusters in keyword density will also help our modeling of sample data.

	Predicted Yes	Prediced No	Actual Total
Actual Yes	254	28	282
Actual No	0	26	26
Predicted Total	254	54	308

Figure 7.6: Confusion matrix after scaling

From our testing of sample data it can be seen that our new model retained its accuracy in detecting all of the no leaning members of the sample. There are also improvements in prediction of positive community members with 35 more tweeters being correctly classified. This improves

our total accuracy to 91%, and the balanced accuracy to 95%. The new balanced accuracy is an improvement on the null error and shows our model using Sentiment clustering and community detection gives an improved model on assuming all members are of the dominant yes side, showing the efficacy of our model.

7.4 The Echo Chamber

One possible issue with twitter discourse on the overall view of the conversation around same-sex marriage referendum is the under-representation of the no side in this space. The referendum received a 40% no vote [5], but our definite alignments in sampling only contained 8.4% no voters and 91.6% yes. This disparity is caused by the relatively left-leaning populous of twitter and the low average age, who were much more likely to vote yes. As well as this effect we see ideological clustering in the data set perpetuated in our community structure and clustering procedure. This ideological clustering is common on political twitter [131–134], meaning cross ideological discourse is uncommon. In our network we find 3.9% of tweets are across yes and no ideology, which further expresses this political homophily and topical polarisation. Although, In Conover et al. their user-user mention network opposes these findings as it creates a single politically heterogeneous mention network cluster, which contains ideologically-opposed individuals [132]. The marriage referendum does not directly oppose these findings, as it is an amalgamation of the two network types used in this paper, the user-user and the retweet network. Our referendum data is dominated by retweets, with only 19% of its edges being user-user messages, the rest being retweets. The network's retweet prevalence leads to it following a segregated partisan structure, with little inter-connectivity between yes and no voters.

This behaviour creates echo chambers for both sides, where opposed ideologies are likely to further inflate their partisan loyalty [133]. An echo chamber is an environment where users share a narrative which gets reinforced through repeated interactions with people or information of a similar attitude to oneself, which most users have a tendency to favour [135–140]. This can be done through diffusion between like minded peers of bias affirming information in retweets and limiting exposure to diverse perspectives reinforcing their narrative [141]. Selective exposure [142] and confirmation bias [143] may begin to explain how these echo chambers have come about on social media platforms [137, 140, 144, 145]. After the echo chamber's creation it can act as a facilitator to reinforce an existing ideology, and as a result, develop the members thoughts to a more extreme position, known as group polarisation theory [146]. The presence of these echo chambers in our data has allowed us to use mathematical techniques to get an understand of political leaning of users in discourse by link weights and sentiment in the network through their clustering which we would not have been able to otherwise. Drawing parallels to what has been seen at the poles, where referendums have been seen to exacerbate this idea of "us

against them", facilitating polarisation [147, 148]. This polarisation has been seen to allow like minded individuals to possess the contextual information to translate their values into support of a particular side [149, 150] and join the rally for the cause.

Chapter 8

Conclusion

We have analysed the inner workings of a data set of tweets made in relation to the 2015 Irish Marriage referendum, to investigate the discourse surrounding the referendum and understand the dissemination of information. From a total of 408,201 tweets, the data was reduced to 94,039 reciprocated tweets between 3546 unique users, to better understand the interpersonal political discourse and negate the directed nature of tweets. Using the reduced data I constructed a directed reciprocated mention network whose link weights correspond to the sentiment of the tweet between the 2 users. The sentiment was calculated using the manually assigned Afinn lexicon. The lexicon cannot interpret sarcasm or tone, making the sentiment scores inherently inconsistent in a vacuum, but aggregated over thousands of tweets the effect of statistical noise diminishes.

We performed statistical testing to understand how the sentiment of users tweets affected their rate of interaction. Monte Carlo simulations were used to provide a topologically identical null model to analyse our system against, giving a significant correlation between the positivity with which a user tweets and the emotion of the response they receive. Monte Carlo simulations were implemented to analyse the ratio of links with positive ends between null and empirical system, showing users who perpetuate positive emotions are more likely to interact with similar people. We use this sentiment similarity as a proxy for homophily between users in the network and see that both of these hypothesis are robust to randomization testing.

Using the Louvain algorithm, 13 communities were detected in the network of varying sizes with a modularity of 0.257. The total in and out sentiment was attached to each community and k-means clustering used to classify these communities into two partitions. After keyword analysis to both ideologies was implemented it is evident that the more positive of the groups of communities can be referred to as the yes cluster and the negative being the no cluster. Manu-

ally assigned sampling data was used to test the efficacy of these communities and it finds that the model over reports no affiliation due to the small communities being pulled into the negative cluster. The k-means algorithm used has no community size scale and labels small communities as negative because their total sentiment can't be as large. A scaling of community sentiment with their relative size before k-means is introduced which leads to a more accurate sampling test and pronounced polarisation in the keyword analysis.

Our ability to predict ideology through clustering and community detection with a 95% accuracy shows a polarised discourse of homophilic communities, that have a very low level of links between clusters (3.9%). Polarisation is expected during a referendum, but on social media it can lead to an echo chamber which can further increase radicalisation and polarisation for its members. The dominance of retweets in our network may further exacerbate this view of polarisation, with a reduction to just user-user tweets this ideological intermingling may become more apparent as it has in other papers [132]. User-user analysis coupled with more qualitative features of users could give a more robust understanding of these mechanisms in further research and allow us to draw a better qualitative picture of this referendum, or any twitter discussion topic.

Bibliography

[1] Kath Browne and Catherine Jean Nash. Resisting marriage equalities: The complexities of religious opposition to same sex marriage. In *Spaces of Spirituality*, pages 37–53. Routledge, 2018.

[2] Government of Ireland. Criminal law (sexual offences) act 1993, 1993.

[3] Claire Edwards. Emplacing disabled bodies/minds in criminal law: Regulating sex and sexual consent in ireland's criminal law (sexual offences) act 1993. In *Disability, spaces and places of policy exclusion*, pages 45–61. Routledge, 2014.

[4] Yvonne Murphy. The marriage equality referendum 2015. *Irish Political Studies*, 31(2):315–330, 2016.

[5] David JP O'Sullivan, Guillermo Garduño-Hernández, James P Gleeson, and Mariano Beguerisse-Díaz. Integrating sentiment and social structure to determine preference alignments: the irish marriage referendum. *Royal Society open science*, 4(7):170154, 2017.

[6] Mohammed Al-Sarem, Wadii Boulila, Muna Al-Harby, Junaid Qadir, and Abdullah Al-saeedi. Deep learning-based rumor detection on microblogging platforms: a systematic review. *IEEE Access*, 7:152788–152812, 2019.

[7] Jihyun Kim and Hayeon Song. Celebrity's self-disclosure on twitter and parasocial relationships: A mediating role of social presence. *Computers in Human Behavior*, 62:570–577, 2016.

[8] Michael G Blight, Erin K Ruppel, and Kelsea V Schoenbauer. Sense of community on twitter and instagram: Exploring the roles of motives and parasocial relationships. *Cyberpsychology, Behavior, and Social Networking*, 20(5):314–319, 2017.

[9] Nicholas A Diakopoulos and David A Shamma. Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of the SIGCHI conference on human factors*

in computing systems, pages 1195–1198, 2010.

[10] Damian Trilling. Two different debates? investigating the relationship between a political debate on tv and simultaneous comments on twitter. *Social science computer review*, 33(3):259–276, 2015.

[11] Stefanie Walter, Ines Lörcher, and Michael Brüggemann. Scientific networks on twitter: Analyzing scientists' interactions in the climate change debate. *Public Understanding of Science*, 28(6):696–712, 2019.

[12] Paul Baker and Tony McEnery. Who benefits when discourse gets democratised? analysing a twitter corpus around the british benefits street debate. In *Corpora and discourse studies*, pages 244–265. Springer, 2015.

[13] Mirko Lai, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. Stance evolution and twitter interactions in an italian political debate. In *International Conference on Applications of Natural Language to Information Systems*, pages 15–27. Springer, 2018.

[14] Joan Balcells and Albert Padró-Solanet. Crossing lines in the twitter debate on catalonia's independence. *The International Journal of Press/Politics*, 25(1):28–52, 2020.

[15] Iina Hellsten, Sandra Jacobs, and Anke Wonneberger. Active and passive stakeholders in issue arenas: A communication network approach to the bird flu debate on twitter. *Public Relations Review*, 45(1):35–48, 2019.

[16] Xabier Martínez-Rolán, Teresa Piñeiro-Otero, et al. The use of memes in the discourse of political parties on twitter: analysing the 2015 state of the nation debate. *Communication & Society*, 29(1):145–160, 2016.

[17] Peter Cihon and Taha Yasseri. A biased review of biases in twitter studies on political collective action. *Frontiers in Physics*, 4:34, 2016.

[18] Ghadir Abdulhakim Abdo Abdullah Alselwia and Sümeyye Kaynakb. Sentiment analysis on covid-19 vaccines tweets.

[19] Zsolt Bederna and Tamás Szádeczky. Modelling computer networks for further security research. *Security and Defence Quarterly*, 2021.

[20] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113, 2014.

[21] Chang Liu and Yueli Xu. Consumer sentiment involvement in big data analytics and its

impact on product design innovation. *Sustainability*, 13(21):1–12, 2021.

[22] Minara P Anto, Mejo Antony, KM Muhsina, Nivya Johny, Vinay James, and Aswathy Wilson. Product rating using sentiment analysis. In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, pages 3458–3462. IEEE, 2016.

[23] Sayan Unankard, Xue Li, Mohamed Sharaf, Jiang Zhong, and Xueming Li. Predicting elections from social networks based on sub-event detection and sentiment analysis. In *International Conference on Web Information Systems Engineering*, pages 1–16. Springer, 2014.

[24] Sitaram Asur and Bernardo A Huberman. Predicting the future with social media. In *2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology*, volume 1, pages 492–499. IEEE, 2010.

[25] Anshul Mittal and Arpit Goel. Stock prediction using twitter sentiment analysis. *Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf)*, 15, 2012.

[26] Joshi Kalyani, Prof Bharathi, Prof Jyothi, et al. Stock trend prediction using news sentiment analysis. *arXiv preprint arXiv:1607.01958*, 2016.

[27] Venkata Sasank Pagolu, Kamal Nayan Reddy, Ganapati Panda, and Babita Majhi. Sentiment analysis of twitter data for predicting stock market movements. In *2016 international conference on signal processing, communication, power and embedded system (SCOPES)*, pages 1345–1350. IEEE, 2016.

[28] Widodo Budiharto and Meiliana Meiliana. Prediction and analysis of indonesia presidential election from twitter using sentiment analysis. *Journal of Big data*, 5(1):1–10, 2018.

[29] Parul Sharma and Teng-Sheng Moh. Prediction of indian election using sentiment analysis on hindi twitter. In *2016 IEEE international conference on big data (big data)*, pages 1966–1971. IEEE, 2016.

[30] Murphy Choy, Michelle LF Cheong, Ma Nang Laik, and Koo Ping Shung. A sentiment analysis of singapore presidential election 2011 using twitter data with census correction. *arXiv preprint arXiv:1108.5520*, 2011.

[31] Daniel Gayo-Avello. " i wanted to predict elections with twitter and all i got was this lousy paper"–a balanced survey on election prediction using twitter data. *arXiv preprint arXiv:1204.6441*, 2012.

[32] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

[33] Gueorgi Kossinets and Duncan J Watts. Origins of homophily in an evolving social network. *American journal of sociology*, 115(2):405–450, 2009.

[34] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[35] Roger Guimera, Marta Sales-Pardo, and Luís A Nunes Amaral. Modularity from fluctuations in random graphs and complex networks. *Physical Review E*, 70(2):025101, 2004.

[36] Albert-László Barabási. Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1987):20120375, 2013.

[37] R Duncan Luce and Albert D Perry. A method of matrix analysis of group structure. *Psychometrika*, 14(2):95–116, 1949.

[38] Robert S Weiss and Eugene Jacobson. A method for the analysis of the structure of complex organizations. *American Sociological Review*, 20(6):661–668, 1955.

[39] Wayne W Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, 33(4):452–473, 1977.

[40] Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. The global k-means clustering algorithm. *Pattern recognition*, 36(2):451–461, 2003.

[41] Hailei Jiang, Rohit Patwardhan, and Sirish L Shah. Root cause diagnosis of plant-wide oscillations using the concept of adjacency matrix. *Journal of Process Control*, 19(8):1347–1354, 2009.

[42] James P Bagrow and Erik M Bollt. Local method for detecting communities. *Physical Review E*, 72(4):046108, 2005.

[43] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.

[44] Hugh Louch. Personal network integration: transitivity and homophily in strong-tie relations. *Social networks*, 22(1):45–64, 2000.

[45] Francis J Flynn, Ray E Reagans, and Lucia Guillory. Do you two know each other? transitivity, homophily, and the need for (network) closure. *Journal of personality and social psychology*, 99(5):855, 2010.

[46] Bat Batjargal. Network triads: Transitivity, referral and venture capital decisions in china and russia. *Journal of International Business Studies*, 38(6):998–1012, 2007.

[47] Riitta Toivonen, Jukka-Pekka Onnela, Jari Saramäki, Jörkki Hyvönen, and Kimmo Kaski. A model for social networks. *Physica A: Statistical Mechanics and its Applications*, 371(2):851–860, 2006.

[48] Stanley Wasserman, Katherine Faust, et al. Social network analysis: Methods and applications. 1994.

[49] Stanley Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967.

[50] Ergin Elmacioglu and Dongwon Lee. On six degrees of separation in dblp-db and more. *ACM SIGMOD Record*, 34(2):33–40, 2005.

[51] Lars Backstrom, Paolo Boldi, Marco Rosa, Johan Ugander, and Sebastiano Vigna. Four degrees of separation. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 33–42, 2012.

[52] Zhiwei Gao, Dexing Kong, Chuanhou Gao, and Michael Chen. Modeling and control of complex dynamic systems 2013, 2013.

[53] Mark EJ Newman. Assortative mixing in networks. *Physical review letters*, 89(20):208701, 2002.

[54] Dapeng Hao and Chuan-Xing Li. The dichotomy in degree correlation of biological networks. *PloS one*, 6:e28322, 12 2011.

[55] Mark EJ Newman and Juyong Park. Why social networks are different from other types of networks. *Physical review E*, 68(3):036122, 2003.

[56] Sujin Choi and Han Woo Park. Networking interest and networked structure: A quantitative analysis of twitter data. *Social Science Computer Review*, 33(2):145–162, 2015.

[57] Theo Lynn, Pierangelo Rosati, Binesh Nair, and Ciáran Mac an Bhaird. An exploratory data analysis of the# crowdfunding network on twitter. *Journal of Open Innovation: Technology, Market, and Complexity*, 6(3):80, 2020.

[58] Sergey Edunov, Carlos Diuk, Ismail Onur Filiz, Smriti Bhagat, and Moira Burke. Three and a half degrees of separation. *Research at Facebook*, 694, 2016.

[59] Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503*, 2011.

[60] Lei Zhang and Wanqing Tu. Six degrees of separation in online society. 2009.

[61] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600, 2010.

[62] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326, 2010.

[63] David Easley, Jon Kleinberg, et al. Power laws and rich-get-richer phenomena. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World. Cambridge University Press*, 2010.

[64] Albert-László Barabási. Linked: The new science of networks, 2003.

[65] Georgios Paltoglou and Mike Thelwall. Twitter, myspace, digg: Unsupervised sentiment analysis in social media. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4):1–19, 2012.

[66] Mike Thelwall. Heart and soul: Sentiment strength detection in the social web with sentistrength, 2017. *Cyberemotions: Collective emotions in cyberspace*, 2014.

[67] Adam Bermingham and Alan F Smeaton. Classifying sentiment in microblogs: is brevity an advantage? In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1833–1836, 2010.

[68] Pollyanna Gonçalves, Matheus Araújo, Fabrício Benevenuto, and Meeyoung Cha. Comparing and combining sentiment analysis methods. In *Proceedings of the first ACM conference on Online social networks*, pages 27–38, 2013.

[69] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *arXiv preprint cs/0506075*, 2005.

[70] Fajri Koto and Mirna Adriani. A comparative study on twitter sentiment analysis: Which features are good? In *International Conference on Applications of natural language to information systems*, pages 453–457. Springer, 2015.

[71] Hailong Zhang, Wenyan Gan, and Bo Jiang. Machine learning and lexicon based methods for sentiment classification: A survey. In *2014 11th web information system and application conference*, pages 262–265. IEEE, 2014.

[72] F Nielsen. Evaluation of a word list for sentiment analysis in microblogs. arxiv 2011. *arXiv*

preprint arXiv:1103.2903.

[73] K Todd, AP Lapointe, and SP Broglio. Sentiment analysis of journal articles and news articles pertaining to cte. *Archives of Clinical Neuropsychology*, 34(5):738–738, 2019.

[74] Julia Silge and David Robinson. *Text mining with R: A tidy approach.* " O'Reilly Media, Inc.", 2017.

[75] Eric T Nalisnick and Henry S Baird. Character-to-character sentiment analysis in shakespeare's plays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 479–483, 2013.

[76] Shashank Sharma, PYKL Srinivas, and Rakesh Chandra Balabantaray. Text normalization of code mix and sentiment analysis. In *2015 international conference on advances in computing, communications and informatics (ICACCI)*, pages 1468–1473. IEEE, 2015.

[77] Qiwei Gan and Yang Yu. Restaurant rating: Industrial standard and word-of-mouth–a text mining and multi-dimensional sentiment analysis. In *2015 48th Hawaii International Conference on System Sciences*, pages 1332–1340. IEEE, 2015.

[78] Xu Chu, John Morcos, Ihab F Ilyas, Mourad Ouzzani, Paolo Papotti, Nan Tang, and Yin Ye. Katara: A data cleaning system powered by knowledge bases and crowdsourcing. In *Proceedings of the 2015 ACM SIGMOD international conference on management of data*, pages 1247–1261, 2015.

[79] Daniel Haas, Sanjay Krishnan, Jiannan Wang, Michael J Franklin, and Eugene Wu. Wisteria: Nurturing scalable data cleaning infrastructure. *Proceedings of the VLDB Endowment*, 8(12):2004–2007, 2015.

[80] Zuhair Khayyat, Ihab F Ilyas, Alekh Jindal, Samuel Madden, Mourad Ouzzani, Paolo Papotti, Jorge-Arnulfo Quiané-Ruiz, Nan Tang, and Si Yin. Bigdansing: A system for big data cleansing. In *Proceedings of the 2015 ACM SIGMOD international conference on management of data*, pages 1215–1230, 2015.

[81] Sanjay Krishnan, Daniel Haas, Michael J Franklin, and Eugene Wu. Towards reliable interactive data cleaning: A user survey and recommendations. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, pages 1–5, 2016.

[82] Hassan Saif, Miriam Fernández, Yulan He, and Harith Alani. On stopwords, filtering and data sparsity for sentiment analysis of twitter. 2014.

[83] Nathaniel Charlton, Colin Singleton, and Danica Vukadinović Greetham. In the mood:

the dynamics of collective sentiments on twitter. *Royal Society Open Science*, 3(6):160162, 2016.

[84] Aliza Sarlan, Chayanit Nadam, and Shuib Basri. Twitter sentiment analysis. In *Proceedings of the 6th International conference on Information Technology and Multimedia*, pages 212–216. IEEE, 2014.

[85] Peter Sheridan Dodds, Eric M Clark, Suma Desu, Morgan R Frank, Andrew J Reagan, Jake Ryland Williams, Lewis Mitchell, Kameron Decker Harris, Isabel M Kloumann, James P Bagrow, et al. Human language reveals a universal positivity bias. *Proceedings of the national academy of sciences*, 112(8):2389–2394, 2015.

[86] Jonathan Mellon and Christopher Prosser. Twitter and facebook are not representative of the general population: Political attitudes and demographics of british social media users. *Research & Politics*, 4(3):2053168017720008, 2017.

[87] Robert L Harrison. Introduction to monte carlo simulation. In *AIP conference proceedings*, volume 1204, pages 17–21. American Institute of Physics, 2010.

[88] Christopher Z Mooney. *Monte carlo simulation*. Number 116. Sage, 1997.

[89] Stefan Krause, Lutz Mattner, Richard James, Tristan Guttridge, Mark J Corcoran, Samuel H Gruber, and Jens Krause. Social network analysis and valid markov chain monte carlo tests of null models. *Behavioral Ecology and Sociobiology*, 63(7):1089–1096, 2009.

[90] Philip Sedgwick. Pearson's correlation coefficient. *Bmj*, 345, 2012.

[91] Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi. Benchmark graphs for testing community detection algorithms. *Physical review E*, 78(4):046110, 2008.

[92] Mason A Porter, Jukka-Pekka Onnela, Peter J Mucha, et al. Communities in networks. *Notices of the AMS*, 56(9):1082–1097, 2009.

[93] Stuart A Rice. The identification of blocs in small political bodies. *American Political Science Review*, 21(3):619–627, 1927.

[94] George C Homans, A Paul Hare, and Richard Brian Polley. *The human group*. Routledge, 2017.

[95] Barry Wellman. The development of social network analysis: A study in the sociology of science. *Contemporary Sociology*, 37(3):221, 2008.

[96] James Moody and Douglas R White. Structural cohesion and embeddedness: A hierarchical concept of social groups. *American sociological review*, pages 103–127, 2003.

[97] James Samuel Coleman et al. Introduction to mathematical sociology. *Introduction to mathematical sociology.*, 1964.

[98] Leo Egghe and Ronald Rousseau. *Introduction to informetrics: Quantitative methods in library, documentation and information science*. Elsevier Science Publishers, 1990.

[99] Luca Donetti and Miguel A Munoz. Detecting network communities: a new systematic and efficient algorithm. *Journal of Statistical Mechanics: Theory and Experiment*, 2004(10):P10012, 2004.

[100] Ingve Simonsen, Kasper Astrup Eriksen, Sergei Maslov, and Kim Sneppen. Diffusion on complex networks: a way to probe their large-scale topological structures. *Physica A: Statistical Mechanics and its Applications*, 336(1-2):163–173, 2004.

[101] Gary William Flake, Steve Lawrence, C Lee Giles, and Frans M Coetzee. Self-organization and identification of web communities. *Computer*, 35(3):66–70, 2002.

[102] Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto, and Domenico Parisi. Defining and identifying communities in networks. *Proceedings of the national academy of sciences*, 101(9):2658–2663, 2004.

[103] Mark EJ Newman and Michelle Girvan. Mixing patterns and community structure in networks. In *Statistical mechanics of complex networks*, pages 66–87. Springer, 2003.

[104] Ulrik Brandes, Marco Gaertler, and Dorothea Wagner. Experiments on graph clustering algorithms. In *European symposium on algorithms*, pages 568–579. Springer, 2003.

[105] Santo Fortunato and Marc Barthelemy. Resolution limit in community detection. *Proceedings of the national academy of sciences*, 104(1):36–41, 2007.

[106] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[107] Jussi M Kumpula, Jari Saramäki, Kimmo Kaski, and János Kertész. Limited resolution in complex network community detection with potts model approach. *The European Physical Journal B*, 56(1):41–45, 2007.

[108] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *nature*, 435(7043):814–

818, 2005.

[109] Roger Guimera, Leon Danon, Albert Diaz-Guilera, Francesc Giralt, and Alex Arenas. Self-similar community structure in a network of human interactions. *Physical review E*, 68(6):065103, 2003.

[110] Leon Danon, Albert Diaz-Guilera, Jordi Duch, and Alex Arenas. Comparing community structure identification. *Journal of statistical mechanics: Theory and experiment*, 2005(09):P09008, 2005.

[111] Mark EJ Newman. Fast algorithm for detecting community structure in networks. *Physical review E*, 69(6):066133, 2004.

[112] Jianhua Ruan and Weixiong Zhang. Identifying network communities with a high resolution. *Physical Review E*, 77(1):016104, 2008.

[113] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.

[114] Jordi Duch and Alex Arenas. Community detection in complex networks using extremal optimization. *Physical review E*, 72(2):027104, 2005.

[115] Stefan Boettcher and Allon G Percus. Optimization with extremal dynamics. *complexity*, 8(2):57–62, 2002.

[116] Stefan Boettcher and Allon G Percus. Extremal optimization for graph partitioning. *Physical Review E*, 64(2):026114, 2001.

[117] Jörg Reichardt and Stefan Bornholdt. Detecting fuzzy community structures in complex networks with a potts model. *Physical review letters*, 93(21):218701, 2004.

[118] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[119] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.

[120] Usman Naseem, Shah Khalid Khan, Imran Razzak, and Ibrahim A Hameed. Hybrid words representation for airlines sentiment analysis. In *Australasian Joint Conference on Artificial Intelligence*, pages 381–392. Springer, 2019.

[121] Aris Tri Jaka Harjanta and Bambang Agus Herlambang. Extraction sentiment analysis using naive bayes algorithm and reducing noise word applied in indonesian language. In

IOP Conference Series: Materials Science and Engineering, volume 835, page 012051. IOP Publishing, 2020.

[122] Purnima Bholowalia and Arvind Kumar. Ebk-means: A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications*, 105(9), 2014.

[123] José M Pena, Jose Antonio Lozano, and Pedro Larranaga. An empirical comparison of four initialization methods for the k-means algorithm. *Pattern recognition letters*, 20(10):1027–1040, 1999.

[124] Andrea Lancichinetti and Santo Fortunato. Community detection algorithms: a comparative analysis. *Physical review E*, 80(5):056117, 2009.

[125] Roger Guimera and Luís A Nunes Amaral. Functional cartography of complex metabolic networks. *nature*, 433(7028):895–900, 2005.

[126] Mursel Tasgin, Amac Herdagdelen, and Haluk Bingol. Community detection in complex networks using genetic algorithms. *arXiv preprint arXiv:0711.0491*, 2007.

[127] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[128] Miyuki Yamada, Yuichi Murai, and Ichiro Kumagai. Story visualization of novels with multi-theme keyword density analysis. *Journal of visualization*, 16(3):247–257, 2013.

[129] Xin Chen and Brook Wu. Assessing student learning through keyword density analysis of online class messages. *AMCIS 2004 Proceedings*, page 362, 2004.

[130] Meenakshi Bansal and Deepak Sharma. Improving webpage visibility in search engines by enhancing keyword density using improved on-page optimization technique. *International Journal of Computer Science and Information Technologies*, 6(6):5347–5352, 2015.

[131] Preethi Lahoti, Kiran Garimella, and Aristides Gionis. Joint non-negative matrix factorization for learning ideological leaning on twitter. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 351–359, 2018.

[132] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. Political polarization on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, pages 89–96, 2011.

[133] Anatoliy Gruzd and Jeffrey Roy. Investigating political polarization on twitter: A canadian perspective. *Policy & internet*, 6(1):28–45, 2014.

[134] Itai Himelboim, Stephen McCreery, and Marc Smith. Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on twitter. *Journal of computer-mediated communication*, 18(2):154–174, 2013.

[135] Kathleen Hall Jamieson and Joseph N Cappella. *Echo chamber: Rush Limbaugh and the conservative media establishment*. Oxford University Press, 2008.

[136] R Kelly Garrett. Echo chambers online?: Politically motivated selective exposure among internet news users. *Journal of computer-mediated communication*, 14(2):265–285, 2009.

[137] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559, 2016.

[138] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 world wide web conference*, pages 913–922, 2018.

[139] Wesley Cota, Silvio C Ferreira, Romualdo Pastor-Satorras, and Michele Starnini. Quantifying echo chamber effects in information spreading over political communication networks. *EPJ Data Science*, 8(1):1–13, 2019.

[140] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. The effect of collective attention on controversial debates on social media. In *Proceedings of the 2017 ACM on Web Science Conference*, pages 43–52, 2017.

[141] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9), 2021.

[142] Joseph T Klapper. The effects of mass communication. 1960.

[143] Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220, 1998.

[144] Michela Del Vicario, Gianna Vivaldo, Alessandro Bessi, Fabiana Zollo, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Echo chambers: Emotional contagion and group polarization on facebook. *Scientific reports*, 6(1):1–12, 2016.

[145] Alessandro Bessi, Mauro Coletto, George Alexandru Davidescu, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Science vs conspiracy: Collective narratives in the age of misinformation. *PloS one*, 10(2):e0118093, 2015.

[146] Cass R Sunstein. The law of group polarization. *University of Chicago Law School, John M. Olin Law & Economics Working Paper*, (91), 1999.

[147] Jane Suiter. Deliberation in action–ireland's abortion referendum. *Political Insight*, 9(3):30–32, 2018.

[148] Bryana Tunder. Same-sex marriage and conservative christian values: A comparison between the republic of ireland and the state of california (us) from a post-legalisation perspective.

[149] John R Zaller et al. *The nature and origins of mass opinion*. Cambridge university press, 1992.

[150] Jane Suiter and Theresa Reidy. Does deliberation help deliver informed electorates: Evidence from irish referendum votes. *Representation*, 56(4):539–557, 2020.