THE UNIVERSITY of ADELAIDE

Towards Exposing Coordinating Inauthentic Groups on Social Media

Author
Derek Christopher Weber

Supervisor
Prof Frank Neumann

Co-supervisors
A/Prof Lucia Falzon
Prof Michael Webb

A thesis submitted for the degree of
DOCTOR OF PHILOSOPHY
School of Computer Science
The University of Adelaide

September 6, 2022

Abstract

Derek Christopher WEBER

Towards Exposing Coordinating Inauthentic Groups on Social Media

Narratives can influence people on social media, and coordinating their dissemination can amplify their effects, which may result in polarisation between communities. Misinformation can exacerbate this polarisation by causing misunderstandings, potentially encouraging the formation of echo chambers and filter bubbles, further hampering dialogue. Deliberate coordinated inauthentic behaviour (CIB) is a core element of disinformation campaigns and Strategic Information Operations (SIOs) that exploit these online phenomena. CIB has been used for ideological and political reasons to pollute our information environment with biased narratives, misleading and false information and propaganda, intensifying existing societal divisions to the extent that it can threaten national security.

Prior research has focused on detecting and classifying entire campaigns (e.g., spam) and individual social bots (automated accounts that deceive and influence by appearing human) and botnets. The damage that information disorders causes to society is also well established, with real-world effects such as vaccine hesitancy, increased conspiratorial thinking and even mass violence. We seek to detect the groups of accounts coordinating their behaviour as part of SIOs, appealing to and recruiting unwitting users to promote their propaganda. First, however, we need to understand the context that CIB occurs in, which we investigate via two avenues: the information environment and the communication environment.

The information environment consists of commercially encumbered social media data. This presents challenges for research due to a lack of transparency, which causes a trust deficit in the results of social media analyses. Opaque sampling biases result in filtered social media data streams that produce variations in data with identical boundary criteria. We present a novel process to examine these variations and demonstrate the method via systematic case studies, finding significant flow-on effects on social network analyses.

The communication environment is replete with contentious online discussions, which are particularly vulnerable to information disorders. We detect and characterise the communication strategies of two polarised groups in a temporally phased investigation of an Australian bushfire discussion, and observe the effects of the strategies. Then, in a longitudinal study, we explore how multiple polarised groups reappear and align in differently themed discussions, finding the polarisation largely remains though the discussion themes can overlap.

With this knowledge, we present and demonstrate our novel network-based approach to detect coordinating groups, focusing on identifying accounts that appear to cooperate with anomalously high levels of coincidental behaviour, artificially raising the voices of the few above the many. The method is generalised, applicable to major platforms, and is amenable to near real-time applications, which are vital to counter influence campaigns before they take hold. Further, we extensively validate the method with several political Twitter datasets, introducing techniques to move beyond manual inspection, which has been the dominant approach in the literature.

The research presented in this thesis provides a solid foundation for future investigation of CIB, online polarisation, and trust in social media data.

Declaration of Authorship

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

The author acknowledges that copyright of published works contained within the thesis resides with the copyright holder(s) of those works.

I give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Derek Christopher Weber

SEPTEMBER 2022

Acknowledgements

In the same way that coordinated behaviour is only meaningful if it has a community around it to influence, this thesis is the core of a great deal of analysis and writing, most of which lies on the cutting room floor, and much of which would not have been possible except for the community around me.

I acknowledge the organisation for which I work, the Defence Science and Technology Group, which has supported me throughout this endeavour, but, of course, it is the people in the organisation who are responsible. I owe a great debt of gratitude to my current and former supervisors, particularly Dr Aaron Ceglar and Dr Tim Pattison. For reviews, encouragement and navigating bureaucracy, thank you to Dr Martin Wood, Dr Danielle Iannella, and Dr Katie Parsons, and also to Ms Vandra Adderly, our amazing librarian.

On campus, my biggest thanks go to my supervisors, Profs Frank Neumann and Michael Webb, and A/Prof Lucia Falzon. Thank you for your guidance, encouragement and expertise. Thank you also to the CompSci admin staff, the lifeblood of the university, on whom I relied so often.

Special thanks I reserve for Dr Mehwish Nasim, who helped me so greatly and collaborated so generously. I also acknowledge with gratitude my other co-authors, particularly A/Prof Lewis Mitchell, Dr Dennis Assenmacher, and Dr Christian Grimme, who gave me such great opportunities to learn and grow as researcher. Thanks also to Dr Tim Graham for enthusiasm, discussions and data (for Chapter 5).

Although our fields differed, the Optimisation and Logistics research group has been welcoming, friendly, and such a pleasure to work alongside. Thank you especially to Dr Vahid Roostapour, for his cheerful expertise and company.

There are of course many more people who assisted and supported me on this journey and I acknowledge and thank them for our collaborations and interactions.

Thank you to my Mum and Dad for getting me here, and to my parents-in-law, for constant encouragement, understanding, and patience (especially while proofreading my thesis!).

The biggest thanks of all go to my amazing wife and wonderful children. This thesis is dedicated to my family, and to Mum and Dad especially.

Contents

List of Figures

List of Tables

Preface

This preface provides relevant contextual information about this thesis, including publications arising from the thesis research and information regarding ethics protocols used, the location of source code and other analytic materials, and use of open materials.

Publications

The following publications have arisen from the research conducted for this thesis and are referred to herein by their Roman numeral labels.

I Derek Weber, Mehwish Nasim, Lucia Falzon, and Lewis Mitchell (Apr. 2020a). "#ArsonEmergency and Australia's "Black Summer": Polarisation and Misinformation on Social Media". In: *Disinformation in Open Online Media*. MISDOOM '20. Springer, pp. 159–173. DOI: 10.1007/978-3-030-61841-4_11

II Derek Weber and Frank Neumann (Dec. 2020). "Who's in the Gang? Revealing Coordinating Communities in Social Media". In: *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ASONAM '20. IEEE, pp. 89–93. DOI: 10.1109/asonam49781.2020.9381418

III Derek Weber, Mehwish Nasim, Lewis Mitchell, and Lucia Falzon (Dec. 2020c). "A method to evaluate the reliability of social media data for social network analysis". In: *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ASONAM '20. IEEE, pp. 317–321. DOI: 10.1109/asonam49781.2020.9381461

IV Dennis Assenmacher, Derek Weber, Mike Preuss, André Calero Valdez, Alison Bradshaw, Björn Ross, Stefano Cresci, Heike Trautmann, Frank Neumann, and Christian Grimme (May 2021). "Benchmarking Crisis in Social Media Analytics: A Solution for the Data-Sharing Problem". In: *Social Science Computer Review*, pp. 1–27. DOI: 10.1177/08944393211012268

V Derek Weber, Mehwish Nasim, Lewis Mitchell, and Lucia Falzon (July 2021a). "Exploring the effect of streamed social media data variations on social network analysis". In: *Social Network Analysis and Mining* 11.1, 62:1–62:38. DOI: 10.1007/s13278-021-00770-y

VI Derek Weber and Lucia Falzon (July 2021). "Temporal Nuances of Coordination Network Semantics". In: *arXiv preprint*, pp. 1–14. arXiv: 2107.02588v2 [cs.SI]

VII Derek Weber and Frank Neumann (Oct. 2021). "Amplifying influence through coordinated behaviour in social networks". In: *Social Network Analysis and Mining* 11.1, 111:1–111:42. DOI: 10.1007/s13278-021-00815-2

VIII Derek Weber, Lucia Falzon, Lewis Mitchell, and Mehwish Nasim (June 2022). "Promoting and countering misinformation during Australia's 2019–2020 bushfires: A case study of polarisation". In: *Social Network Analysis and Mining* 12.1, 64:1–64:26. DOI: 10.1007/s13278-022-00892-x

IX Mehwish Nasim, Derek Weber, Tobin South, Jonathan Tuke, Nigel Bean, Lucia Falzon, and Lewis Mitchell (Jan. 2022). "Are we always in strife? A longitudinal study of the echo chamber effect in the Australian Twittersphere". In: *arXiv preprint*. arXiv: 2201.09161 [cs.SI]

Author's Contributions

The author's contributions to the publications mentioned above are as follows:

I "#ArsonEmergency and Australia's "Black Summer": Polarisation and Misinformation on Social Media"

The author performed the data collection, the majority of analysis and more than 80% of the writing and editing. The initial retweet network construction and clustering analysis of it, in particular the production of Figure 5.4, were conducted by Dr Nasim.

II "Who's in the Gang? Revealing Coordinating Communities in Social Media"

The author performed the conceptualisation, coding, data collection and analysis, writing and editing. Guidance on the paper's structure and a review of the first draft was provided by the paper's second author.

III "A method to evaluate the reliability of social media data for social network analysis"

The author contributed to the conceptualisation, conducted the collection of one of the parallel datasets, and performed the remaining coding and analysis, and the majority of the writing and editing, revising elements of the initial draft introduction and background provided by the other authors.

IV "Benchmarking Crisis in Social Media Analytics: A Solution for the Data-Sharing Problem"

The author contributed to the initial conceptualisation, provided a background subsection, and conducted extensive proofreading, commenting and some editing of drafts prior to publication.

V “Exploring the effect of streamed social media data variations on social network analysis”

Building on Publication **III**, the author conducted the collection of all bar two of the new parallel datasets, as well as the coding, analysis, writing and the majority of the editing.

VI “Temporal Nuances of Coordination Network Semantics”

The author developed the conceptualisation with the guidance of the paper’s second author, and conducted the coding, analysis, writing and editing.

VII “Amplifying influence through coordinated behaviour in social networks”

Building on Publication **II**, the author conducted the coding, analysis, writing and editing. The sliding window concept was co-developed by both the publication’s authors.

VIII “Promoting and countering misinformation during Australia’s 2019-2020 bushfires: A case study of polarisation”

Building on Publication **I**, the author conducted the coding, analysis, writing and editing, in particular revising the entire narrative and elements of the publication’s structure based on reviewers’ guidance. The other authors reviewed the final draft.

IX “Are we always in strife? A longitudinal study of the echo chamber effect in the Australian Twittersphere”

The initial concept of this research, that of persistent polarisation, and the collection and early analysis of the Same Sex Marriage (SSM) and Election datasets, was developed and led by Dr Mehwish Nasim, conducted by the other authors (not including the thesis author) and presented as talks at the 2019 NetSci (Nasim et al., 2019) and ASNAC conferences (Nasim, 2019). Together with Dr Nasim, the author of this thesis expanded the concept to further relevant datasets and both developed key elements of the results discussion. Apart from the labelling of the SSM tweets, all further comparative analysis with the newly added datasets was conducted by the author of this thesis. The author of this thesis prepared a complete draft of the paper, revising all earlier text, which was then pared back by Dr Nasim in preparation for journal submission. That complete draft is the basis for Chapter 6 and Section 2.5.

Acronyms

The following acronyms and terms are used in this thesis.

AdlWW	Adelaide Writers Week, an annual writers festival in South Australia
AFL	Australian Football League, the professional competition of Australian rules football
API	Application Programming Interface
ARI	Adjusted Rand Index
BLM	Black Lives Matter, an activism campaign for Black rights
CCP	Chinese Communist Party
CIB	Coordinated inauthentic behaviour (Gleicher, 2018)
CCOT	Christian Conservative on Twitter
DoS/DDoS	(Distributed) Denial of service
DL	Deep learning
DNC	Democratic National Convention
HCC	Highly coordinating community
ID	Identifier
IO	Information operation or influence operation
KAG	Keep America Great, a campaign slogan
KHive	An informal online community supporting Kamala Harris, the 49th Vice President of the United States
ISIS	Islamic State of Iraq and Syria, a terrorist group
MAGA	Make America Great Again, a campaign slogan
ML	Machine learning
MSM	Mainstream media
NLP	Natural language processing
NSW	New South Wales
OSN	Online social network or social media platform (e.g., Facebook, Twitter, or Tumblr)
OSoMe	The University of Indiana's Observatory on Social Media project
RAPID	Real-Time Analytics Platform for Interactive Data Mining (Lim et al., 2019)
RNC	Republican National Convention
RT	Retweet
RU-IRA	The Russian Internet Research Agency (Chen, 2015; Mueller, 2018)
SA	South Australia
SIO	Strategic information operation
T&Cs	Terms and conditions
tf-idf	Term frequency–inverse document frequency
TCOT	Top Conservative on Twitter[1]
UI	(Graphical) User interface
UK	The United Kingdom
URL	Uniform resource locator
US	The United States of America
WWW	World Wide Web

[1] https://www.dailydot.com/debug/tcot-trump-maga/. Posted 2019-07-19. Accessed 2022-01-29.

Conventions

Where hashtags containing multiple words, they are presented with each word capitalised for readability, e.g., `#BlackLivesMatter`, unless the hashtag is commonly known by another form (e.g., `#auspol` or `#qanda`). If the hashtag includes a word that is an acronym, only the first letter of the acronym term will be capitalised, e.g., `#NswVotes` referring to New South Wales (NSW) Votes, or `#AflPiesDees` referring to Australia Football League (AFL) game between the Pies (Collingwood Magpies) and the Dees (Melbourne Demons). All hashtag analysis is case-insensitive, however, as this is how they are treated by the social media platforms.

The mention interaction, common to Twitter, Facebook and other platforms, is sometimes written as '@mention' to emphasise that it is the interaction type that is being used.

To improve readability, thousands may be abbreviated with 'k', e.g., 22k is equivalent to 22,000, and millions may be abbreviated with 'm', e.g., 1.3m is equivalent to 1,300,000.

Ethics

All data associated with this PhD project were collected, stored, processed and analysed according to two ethics protocols, H-2018-045 and #170316, approved by the University of Adelaide's human research and ethics committee.

Open Content

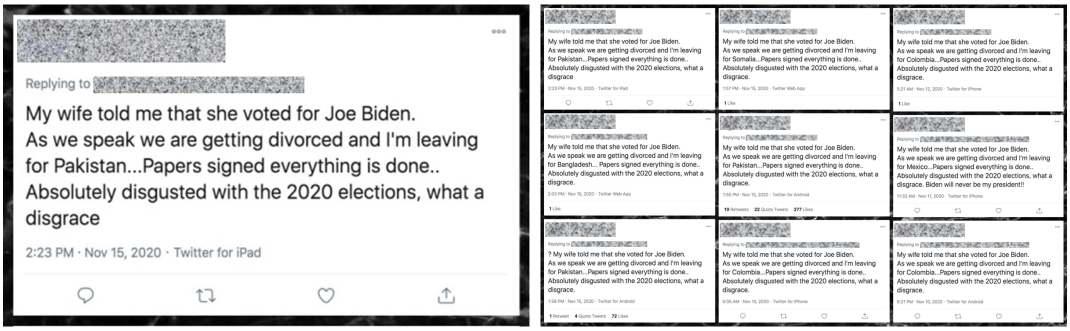
Open content in this thesis was used in accordance with Creative Commons[2] and GPLv3[3] licences.

[2] Attribution-NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0): https://creativecommons.org/licenses/by-nc-nd/3.0/. Accessed 2021-12-20.

[3] The GPLv3 licence used by the DBA GitHub project: https://github.com/fpetitjean/DBA/blob/master/LICENSE. Accessed 2021-12-20.

Chapter 1

Introduction



[redacted]

Replying to [redacted]

My wife told me that she voted for Joe Biden.
As we speak we are getting divorced and I'm leaving for Pakistan...Papers signed everything is done..
Absolutely disgusted with the 2020 elections, what a disgrace

2:23 PM · Nov 15, 2020 · Twitter for iPad

(a) One upset partner.

[redacted]

Replying to [redacted]

My wife told me that she voted for Joe Biden.
As we speak we are getting divorced and I'm leaving for Pakistan...Papers signed everything is done..
Absolutely disgusted with the 2020 elections, what a disgrace

2:23 PM · Nov 15, 2020 · Twitter for iPad

[redacted]

Replying to [redacted]

My wife told me that she voted for Joe Biden.
As we speak we are getting divorced and I'm leaving for Somalia...Papers signed everything is done..
Absolutely disgusted with the 2020 elections, what a disgrace.

1:57 PM · Nov 15, 2020 · Twitter Web App

1 Like

[redacted]

Replying to [redacted]

My wife told me that she voted for Joe Biden.
As we speak we are getting divorced and I'm leaving for Colombia...Papers signed everything is done..
Absolutely disgusted with the 2020 elections, what a disgrace.

5:21 AM · Nov 12, 2020 · Twitter for iPhone

1 Like

[redacted]

Replying to [redacted]

My wife told me that she voted for Joe Biden.
As we speak we are getting divorced and I'm leaving for Bangladesh... Papers signed everything is done..
Absolutely disgusted with the 2020 elections, what a disgrace.

2:03 PM · Nov 15, 2020 · Twitter Web App

1 Like

[redacted]

Replying to [redacted]

My wife told me that she voted for Joe Biden.
As we speak we are getting divorced and I'm leaving for Pakistan...Papers signed everything is done..
Absolutely disgusted with the 2020 elections, what a disgrace.

1:55 PM · Nov 15, 2020 · Twitter for Android

10 Retweets 22 Quote Tweets 277 Likes

[redacted]

Replying to [redacted]

My wife told me that she voted for Joe Biden.
As we speak we are getting divorced and I'm leaving for Mexico...Papers signed everything is done..
Absolutely disgusted with the 2020 elections, what a disgrace. Biden will never be my president!!

11:32 AM · Nov 11, 2020 · Twitter for iPhone

[redacted]

Replying to [redacted]

? My wife told me that she voted for Joe Biden.
As we speak we are getting divorced and I'm leaving for Pakistan...Papers signed everything is done..
Absolutely disgusted with the 2020 elections, what a disgrace.

1:58 PM · Nov 15, 2020 · Twitter for Android

1 Retweet 4 Quote Tweets 72 Likes

[redacted]

Replying to [redacted]

My wife told me that she voted for Joe Biden.
As we speak we are getting divorced and I'm leaving for Colombia...Papers signed everything is done..
Absolutely disgusted with the 2020 elections, what a disgrace.

5:05 AM · Nov 15, 2020 · Twitter for iPhone

[redacted]

Replying to [redacted]

My wife told me that she voted for Joe Biden.
As we speak we are getting divorced and I'm leaving for Colombia...Papers signed everything is done..
Absolutely disgusted with the 2020 elections, what a disgrace.

6:21 PM · Nov 10, 2020 · Twitter for Android

(b) Many upset partners.

FIGURE 1.1. Copypasta tweets noticed in the aftermath of the 2020 US presidential election, which may belie a coordinated campaign to undermine confidence in American society's ability to accept electoral outcomes, or may just be a prank similar to a flashmob.

In the aftermath of the 2020 US presidential election, a data scientist noticed a pattern emerging on Twitter.[1] Figure 1.1a shows a tweet by someone who was so upset with their wife for voting for Joe Biden in the election that they decided to divorce them immediately and move to Pakistan (in the midst of the COVID-19 pandemic). This might seem an extreme reaction, but the interesting thing was that the person was not alone. The researcher identified dozens of similar, but not always identical, tweets by people leaving for other cities but for the same reason (Figure 1.1b). Analysis of these accounts also revealed they were not automated. These posts are coordinated, clearly, but is this "copypasta" directed or emergent?

Another recent example revealed coordination strategies being used. The campaign recruited people to amplify `#Remove1991WorshipAct` in India.[2] It consisted of WhatsApp messages referring people to a Google Drive file of comments to tweet, each of which had a button to create a new pre-written tweet, ready to be sent.

This coordinated pattern of tweeting has been used for more than advocating opinions or political campaigns. It had also been used by ISIS terrorists as they approached the

[1] https://twitter.com/conspirator0/status/1328479128908132358. Posted 2020-11-17. Accessed 2022-01-11.

[2] https://twitter.com/BenDoBrown/status/1383337211832139778. Posted 2021-04-17. Accessed 2022-01-25.

city of Mosul, Iraq, in 2014. By using their Dawn of Glad Tidings app, giving them access to followers' Twitter accounts, they coordinated posts to give the impression of a giant invading army convincing the local forces to abandon their posts. ISIS then occupied Mosul for several years (Brooking and Singer, 2016).

It is unclear whether the first copypasta example is part of a deliberate Strategic Information Operation (SIOs, Starbird et al., 2019), designed to damage trust in the electoral system and ability of Americans to accept the loss of a preferred political party in elections. It could be part of a campaign by an issue-motivated group with the same aims, foreign or domestic, or simply a viral gag by a group of like-minded jokers engaging in a kind of flashmob. The second two examples are clearly deliberate. At the very least, it is important to be able to identify which accounts are core to the activity, and how they are coordinating their actions.

The aim of this thesis is to develop techniques to identify groups engaging in *coordinated inauthentic behaviour* (CIB) on social media, and the context in which CIB is used. CIB can be described as people (or accounts, more precisely) aligning their actions deceptively for political, ideological or commercial gain (Gleicher, 2018). This research, therefore, contributes to the detection and characterisation of SIOs. Groups engaging in CIB aim to manipulate public opinion, sow discord, or otherwise amplify specific narratives and propaganda via the online social networks (OSNs) that have become central to modern life. Evidence has shown that such groups can disrupt communities and exacerbate societal divisions (e.g., CREST, 2017; Keller et al., 2019) including to a level 'beyond reasonable doubt' as judged in courts of law (e.g., Mueller, 2018; Keller et al., 2019). From a national security perspective, they can "interfere with democratic, political and societal processes",[3] are a key element of disinformation operations (Paul and Matthews, 2016; Starbird et al., 2019; Rid, 2020) and have been associated with real-world violence (Scott, 2021; Samuels, 2020; Mackintosh, 2021) and vaccine hesitancy (Broniatowski et al., 2018; Loomba et al., 2021).

CIB can exploit and perpetuate information disorders, such as misinformation and disinformation, which are prevalent online. It has been shown that false information can spread very quickly when it goes viral (Vosoughi et al., 2018) and is very hard to counter once it is anchored in people's minds (Tversky and Kahneman, 1973; Kuran and Sunstein, 1999; Paul and Matthews, 2016). As a result, those tasked with addressing it, such as military and national security and law enforcement agencies, need to keep appraised of online events in near real-time. Despite the plethora of online social network (OSN) data available, or even because of it, the constrained times in which to react can make it difficult to discern malicious information campaigns from genuine grassroots activities. Genuine activities range from amusing fads (e.g.,

[3] Remarks at the Home Affairs Town Hall, National Office, Canberra, made by Michael Pezzullo, Secretary Department of Home Affairs, on 2018-04-19. Source: https://www.homeaffairs.gov.au/news-media/speeches/2018/19-april-home-affairs-town-hall. Posted 2018-04-19. Accessed 2022-01-25.

#RuinAMovieWithOneWord[4]) to persistent activism (e.g., the #BlackLivesMatter or #MeToo movements, Jackson et al., 2020), but even these can be hijacked (e.g., the #StopAsianHate campaign, Zhang, 2021). Modern information operations are 'participatory' activities, appealing to and recruiting unwitting but pliant members of the public to promote preferred narratives (Starbird et al., 2019). Additionally, misinformation and disinformation are often blended in with genuine content to obscure it (Paul and Matthews, 2016; Starbird, 2019; Rid, 2020). As a consequence, what become genuine amplification activities may have, in fact, been instigated by malicious actors.

Increasingly, it appears that no country is immune to SIOs. Since the election-related interference in the US and UK in 2016, protecting Australian elections from interference has been of particular interest to the Australian Government. Its stated Foreign Policy is to "protect the sovereignty, integrity and transparency of our institutions" and "ensure that national decision-making and institutions remain free from foreign interference", specifically with regard to the use of "new media platforms. . . to sow misinformation" (p.76, DFAT, 2017). More recently, the Government has also released a policy on developing resilience to online misinformation and disinformation at the national and international level.[5] As a result, this research contributes to the nation's efforts to protect itself and its interests.

Previous efforts in this space have focused on campaign detection and classification (e.g., Lee et al., 2013; Cao et al., 2015; Varol et al., 2017b; Wu et al., 2018) or the detection of social bots, which are actively deceptive automated accounts (Ferrara et al., 2016; Cresci, 2020). There has been a growing emphasis, however, on the human aspects of running campaigns emerging from the nexus of sociological and computer science research, dubbed computational social science. Effort has shifted to detecting the core groups of accounts behind the activities (e.g., Cao et al., 2014; Yu et al., 2015; Şen et al., 2016; Grimme et al., 2018) rather than the entire campaigns, in which many of the participants may be unwitting recruits. There have been attempts to detect these groups based on specific behaviours (e.g., Vo et al., 2017; Giglietto et al., 2020a; Yu, 2021) but few have proposed generalised approaches.

It is also important to acknowledge that CIB and SIOs do not occur in a vacuum. They work precisely *because* they influence the broader discussion. This context needs to be observed and understood in order to understand how CIB can be identified and what its effects are. In Figure 1.2 we consider a breakdown of the relevant elements of the world, first identifying offline (i.e., real-world) and online communications, and then social media within the online environment, with its various platforms. On social media, there are discussions, which can exhibit polarisation over issues within

[4] https://twitter.com/jimmyfallon/status/1214276112441860098?lang=en. Posted 2020-01-07. Accessed 2022-01-25.

[5] https://www.internationalcybertech.gov.au/our-work/security/disinformation-misinformation. Accessed 2022-01-25.

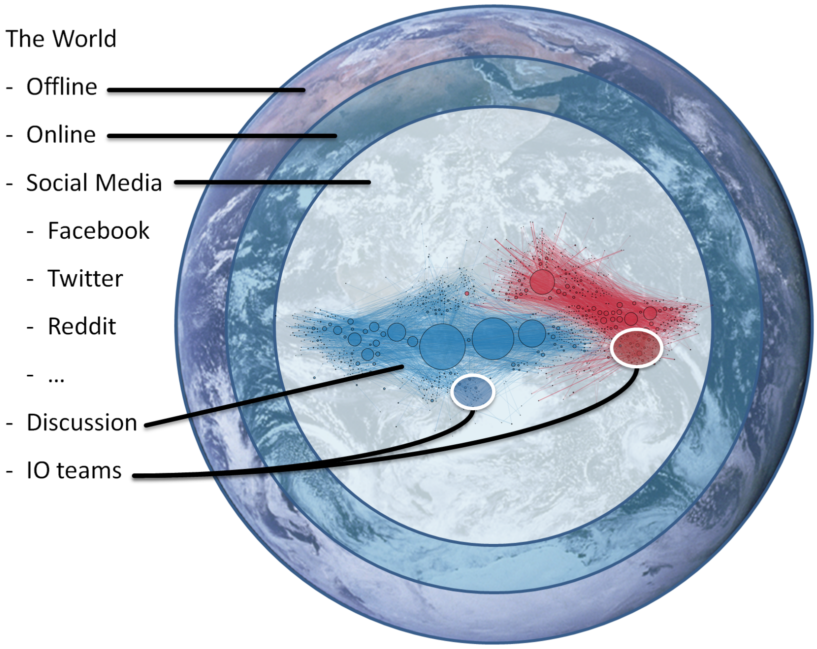


Figure 1.2. Although online activity is often regarded as not part of the 'real world', offline activities affect online behaviour, which then affects offline behaviour.[6] Our focus is on discovering the teams of accounts engaging in IO in contentious discussions on social media, helping drive conflict and polarisation.

the discussion, and ultimately our target is the information operations (IO) teams or groups, which are exacerbating the conflict thereby contributing to the polarisation.

In this thesis, prior to presenting our CIB group detection approach, we examine two elements of the CIB context: the information environment and the communication environment. Here, the information environment refers to the limited data made available by the OSN owners, which are then used to study the communication environment, consisting of the interactions between and content produced by the users. From the perspective of data collection, we rely on OSN data provided via Application Programming Interfaces (APIs). The OSNs are run by commercial entities, with priorities that often conflict with open and transparent data access (in fact, their rich data holdings are the source of their income). In Part I, the effects of this lack of transparency are considered. From the perspective of the communication environment, CIB works by influencing members of the broader discussion and engaging them to willingly disseminate propaganda themselves, giving it the sheen of legitimate opinion. In Part II, we examine the communication environment, contentious online discussions vulnerable to information disorders, and thus manipulation. In these, we identify and characterise polarised groups and the degree to which their relative isolation persists over time and across discussion topics. With those foundations established, in Part III, we present and validate our novel network-based approach for detecting groups engaging in CIB.

[6] The photograph of Earth "The Blue Marble" used in the background was taken on 1972-12-07 by the crew of Apollo 17. Source: https://web.archive.org/web/20160112123725/http://grin.hq.nasa.gov/ABSTRACTS/GPN-2000-001138.html. Accessed 2022-01-21.

1.1 Research Questions

To guide our research, we have organised it according to the following thesis-level research questions (TRQs):

TRQ1 *To what extent can we have trust in social media data and the results of analysis on them?*

This is addressed in our exploration of the information environment through Publications **III**, **IV** and **V** and in Part I.

TRQ2 *How can we identify and characterise polarised communities on social media?*

This is addressed in our analysis of a contentious online discussion relating to climate change, the presence and characteristics of polarised communities in the discussion, and how the discussion changes over time, presented in Publications **I** and **VIII** and in Chapter 5.

TRQ3 *To what extent does polarisation between groups endure over periods of time? Does it only relate to single issues, or does polarisation over one issue consolidate across sets of issues?*

These issues are considered in a longitudinal study of polarised communities, which at times overlap and align depending on the discussion themes, in Publication **IX** and in Chapter 6.

TRQ4 *How can we find groups of accounts that work together to encourage conflict and polarisation?*

Foundations for addressing this question are established with our temporally-aware network-based approach in Publications **II** and **VII** and in Part III.

1.2 Approach

A variety of methods are available to analyse social media data, including natural language processing (NLP), supervised and unsupervised machine learning (ML) techniques, and social network analysis (SNA). We adopt a computational social science approach to our analysis, relying on SNA to examine direct and indirect interaction patterns between accounts and the communities they form, paying particular attention to the temporal aspects of those patterns. Results are confirmed with the examination of the structured elements of social media content, but we avoid the analysis of the free text of social media posts. This approach is justified because:

- social media data is inherently network-based, as the majority of it consists of posts produced by accounts being disseminated to other accounts, where those posts may include links to other accounts and entities (e.g., via URLs, hashtags, and @mentions);

- our primary target is group behaviour based on alignment of account actions, and thus deep analysis of post content is not required; and
- analyses based only on content are limited to the small amounts of highly varying text in the posts themselves, making certain types of analysis challenging;
- analysing the media embedded in posts (e.g., imagery and video content) is computationally prohibitive to process at scale; and
- ML methods, though they can exploit a significant portion of the metadata of posts as part of analysis, are inherently opaque in their operation and require ongoing retraining, and are therefore well-suited to confirmation or triage roles.

The ideal system could exploit a variety of these approaches, as they complement each other to tackle larger aspects of detecting information campaigns, but as our focus is on only one element of those campaigns, identifying groups of accounts disseminating propaganda, we can concentrate on temporally-aware SNA.

1.3 Overview of Main Contributions

In this thesis, we first consider the information environment of social media data and some of the implications for research with such data. We then examine contentious discussions on social media, the communities that form and polarise in them, their behavioural characteristics, and how the discussions shift and communities evolve over time. Finally, we present our novel network-based method for findings groups of accounts that drive narratives and information disorders, encouraging argument and the resulting polarisation.

1.3.1 Part I: The Information Environment

Social media provides an information environment eminently suited to mass distribution of information, partly due to the internet's connectivity and partly due to its popularity. Unfortunately, that information can easily be misinformation or disinformation, resulting in a misinformed public, which can lead to conflict when the information relates to contentious issues. Unlike tracking, say, face-to-face conversations in a schoolyard, tracking social media conversations is a much more tractable task as the information is recorded by the OSNs and elements of it are made available via Application Programming Interfaces (APIs). Though this is a boon for researchers and analysts, there are issues of transparency. It is not clear whether requests to APIs are fulfilled with complete responses, nor what sampling methods are used to decide what is included in a response. Further, the data in the responses is only available for use under OSN-specific terms and conditions (T&Cs).

In Part I, we highlight the implications of this non-transparent information environment on trust in research results. Firstly, we briefly discussing the importance of open unencumbered data for benchmarking, and how social media analytics research may be in the midst of a ‘benchmarking crisis’. Secondly, and more comprehensively, in Chapter 4 we explore the phenomenon of variations in the data provided by Twitter streaming APIs under a variety of conditions. We observe that when using the APIs to collect data at the same time with the same collection criteria, the same results are not always obtained, which has flow on effects on analyses of the data. We establish a systematic methodology for comparing social network analyses of data on such parallel datasets, and demonstrate it through several case studies.

1.3.2 Part II: The Danger of Polarisation

Arguments between communities have occurred for millennia, and may even be identity-forming, but social media has removed previous geographical constraints. Whereas previously, neighbouring villages might disagree over who is responsible for maintaining a bridge over a common river, now communities can argue, having formed by drawing in members from anywhere in the world, with the only limiting feature being language. It may be that arguments on social media are more vociferous and longer lasting, due to the fact that earlier comments can be retrieved and read, whereas in ancient times when someone yelled an epithet across the river it vanished once the recipients forgot about it. Online polarisation presents three primary issues:

- repeated sharing and re-endorsement of limited information and opinions within communities can contribute to the formation of echo chambers and filter bubbles;
- such communities are vulnerable to misinformation and disinformation, and have the potential to be radicalised or self-radicalise; and
- the aggression from this shift to extremism can then spill into the real world, resulting in damage to societal institutions and trust in authority, and even violence.

For these reasons, it is vital to be able to observe and characterise polarised communities on social media, and to analyse their behaviour in the context of real-world events that stimulate their online activity, and the effect it has on the broader discussion. The focus in Chapter 5 is to conduct such an investigation in the context of a contentious discussion relating to the worst Australian bushfires on record and the role of arson and climate change. Having identified polarised communities in one discussion, in Chapter 6, we turn to examine their roles in other contentious online discussions and their degree of overlap with other previously observed polarised communities. The purpose of this is to explore how persistent polarisation can be in the Australian Twittersphere and how that polarisation relates to the topics under debate.

1.3.3 Part III: The Hunt for CIB

As observed during the 2016 US presidential election, existing community divisions were exacerbated through concerted efforts by motivated actors exploiting existing fissures in the fabric of society (Mueller, 2018). Part of those efforts consisted of

Strategic Information Operations (SIOs) on social media, cultivating existing communities and seeding them with disinformation to further inflame tensions and entrench divisions.

In Chapter 7, we present our novel network-based method for identifying groups of accounts engaging in coordinated behaviour, particularly to amplify content, such as URLs, hashtags, or phrases, or direct attacks on other accounts and communities. With specific regard to political discussions, we evaluate our technique against two relevant datasets with comparisons against ground truth and random datasets. In order to move beyond manual inspection of results, we also provide and demonstrate a variety of validation techniques. The method is designed with two key priorities: generalisablity for broad applicability to many common OSNs, and suitability to near real-time processing for practical application in real-world analyst collection systems.

We now provide a detailed background of related fields of research in Chapter 2, which provides context and further motivation. Following this, in Chapter 3, we provide an explanation of the analysis methods we employ, with particular attention to the theory and practice of graphs, networks and SNA.

Chapter 2

Background

The practice of social media analytics lies at the nexus of a great many fields, including computer science, mathematics and statistics, but also sociology, psychology, political science and media studies. Because the data under examination is human-generated, using the lens of only one discipline necessarily misses the context provided by others. In this chapter, we provide a broad overview of a number of these related fields and their concepts in order to provide context for the later chapters, and to illustrate how they are connected in the research landscape.

We begin by exploring the information disorders that afflict the current state of the "'public sphere': the shared space [in which] social issues are discussed and public opinion is formed" (p.50, Wardle and Derakhshan, 2017), such as misinformation and disinformation. Consideration is then given to how the beneficial features of social media, in particular, can be turned to foster information disorders and exacerbate their harm, culminating in a discussion of the concept of computational propaganda. At this point, we emphasise two vital aspects of data accessibility relevant to social media researchers: the importance of benchmarking and how OSN data hampers them, and the reliability of the data provided by the OSNs. A discussion follows of how the concepts of sociology and social network analysis can be applied to data provided by OSNs and some of the challenges faced when doing so. Of particular interest are the concepts contributing to community formation (e.g., *homophily*, McPherson et al., 2001) and conflict (e.g., *polarisation*, Kligler-Vilenchik et al., 2020), and how they are manifested on social media, particularly during times of social, environmental or political importance. Additionally, the concepts of *echo chambers* (Barberá et al., 2015) and *filter bubbles* (Pariser, 2012) provide a basis for academic studies of contentious social, ideological and political discussions. Finally, we examine the literature on the concept of coordinated online behaviour within the context of inauthentic behaviour analysis, including how the concept has developed over time, and how such behaviour can be detected and characterised.

2.1 Information Disorders

Social media use has increased significantly in recent years (notably for political communication) and so the market has followed, with media organisations using it for cheap, wide dissemination and consumers increasingly looking to it for news (Shearer and Grieco, 2019). This has enabled the democratisation of publishing (anyone with an internet connection can be a journalist now, in this age of *citizen journalism*, Gillmor, 2006), removing the traditional intercessor of the news media editors (Woolley and Guilbeault, 2018) and leading to a lack of control over bias and veracity in what is presented as first-person reporting. Coinciding with this have been significant drops in levels of trust not only in political figures but also other authorities such as scientists and technical experts, resulting in people putting more trust in the recommendations of peers, friends and family (i.e., peers, or those they perceive as peers) than in what they observe in the traditionally trusted sources, like the mainstream news media (Kavanagh and Rich, 2018). Media literacy has also been observed lacking, particularly in Australia (Notley et al., 2021), leading to a limited ability for most people to critically assess online information. In this context, in a similar way to a virus taking advantage of suitable environmental conditions to replicate and spread, a number of overlapping information disorders have been documented, many of which have significant real-world effects. These include "fake news", misinformation, disinformation, harmful rumours and conspiracies.

2.1.1 Definitions

In 2017, the phrase "fake news" was voted the American Dialect Society's "word of the year", defined as "disinformation or falsehoods presented as real news" and "actual news that is claimed to be untrue",[1] having been widely popularised in the preceding year's US presidential election campaign. The phrase had been used at times in that year to represent true news that the recipient or subject simply did not like, even to the extent that it might be applied to an entire media organisation,[2] and efforts continue to address its definition (Wardle, 2019a; Starbird, 2019; Chirwa and Manyana, 2021).

Fake news is simply an umbrella description for the information disorders described in the above examples. Wardle (2019b) describes information disorders using a Venn diagram combining falseness and the disseminator's intent to harm, reproduced in Figure 2.1. Similarly, Kumar and Shah (2018) distinguish these disorders based on the information's veracity and the intent of the disseminator. *Disinformation* is information the disseminator knows is false, which they are publishing with a clear intent to cause harm. In contrast, *misinformation* is false information disseminated

[1] https://www.americandialect.org/fake-news-is-2017-american-dialect-society-word-of-the-year. Accessed 2021-11-23

[2] On the basis of reporting he did not like, on 11 January 2017, President-elect Donald Trump remonstrated with a CNN reporter, saying "Your organisation's terrible. . . . You are fake news.". https://www.nytimes.com/video/us/politics/100000004865825/trump-calls-cnn-fake-news.html Accessed 2021-11-23.

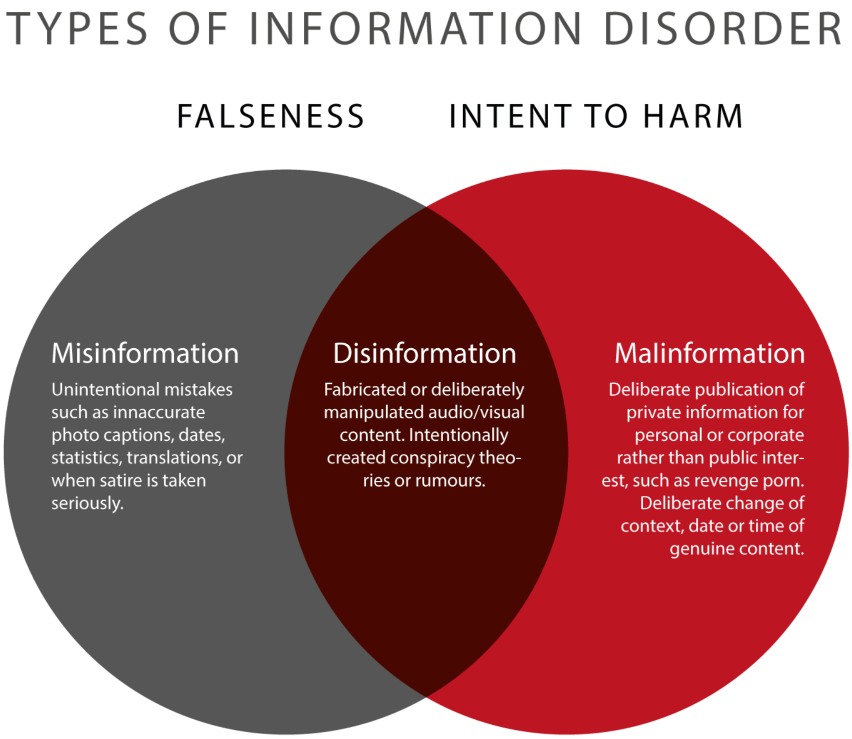

TYPES OF INFORMATION DISORDER
FALSENESS
INTENT TO HARM
Misinformation
Unintentional mistakes such as innaccurate photo captions, dates, statistics, translations, or when satire is taken seriously.
Disinformation
Fabricated or deliberately manipulated audio/visual content. Intentionally created conspiracy theories or rumours.
Malinformation
Deliberate publication of private information for personal or corporate rather than public interest, such as revenge porn. Deliberate change of context, date or time of genuine content.

FIGURE 2.1. Types of information disorder, reproduced from Wardle and Derakhshan (2017) as per Creative Commons licensing (CC BY-NC-ND 3.0). (Updated image obtained from https://medium.com/1st-draft/information-disorder-part-3-useful-graphics-2446c7dbb485 on 2021-11-23.)

by someone who does not realise it is false, or cares sufficiently little to check it, often because it aligns with their worldview (Wardle and Derakhshan, 2017). The final part of the Venn diagram is true information that is distributed with an intent to harm – this is labelled *malinformation* and refers to things like personal information exposed as part of *doxxing*[3] or revenge porn.

The term *propaganda* can be defined as “weaponized speech designed to support one party over another” (p.15, Wardle and Derakhshan, 2017), but also as a synonym for disinformation itself (p.8, Kavanagh and Rich, 2018). It is intentionally disseminated information designed to support one party (e.g., political or nation states) or denigrate the other, whether the information is biased, misleading or entirely false. Its defining trait is that it is designed to promote a point of view, regardless of the means.[4]

[3] The practice of doxxing (revealing people’s personal details into the public to shame or harass), in particular, reveals the ethical element in categorising information disorders, as doxxing is also used to shame those engaged in behaviour damaging to society, as well as by those aiming to harass for personal reasons. After the 6 January 2021 Capitol building riots in Washington D.C., as internet sleuths started using social media and broadcast footage to identify rioters to report them to police, extremism researchers raised concerns with the balance of the activities’ public good with the risk of misidentification and risk to individuals of revenge attacks after exposing dangerous people (Lapowsky, 2021).

[4] https://www.collinsdictionary.com/dictionary/english/propaganda. Accessed 2022-02-02.

Categories of content can also be ranked in terms of the harm they are intended to cause. Wardle (2019b) defines seven types in order of severity: satire or parody; false connections (e.g., misleading headlines); misleading content (i.e., framed or biased presentation of factual material); false context (placing real content in a different context to give false impressions); imposter content (false information presented as reported by reputable sources); manipulated content (e.g., edited videos or images); and outright fabrication aimed to deceive and harm. The point this illustrates is that there is a wide variety of material that could be considered harmful and many people, but not all, will see through the deception. Such techniques have long been used in propaganda and public diplomacy at the nation state level but the internet and social media has vastly expanded the environment in which such techniques can be effective and the pool of people and organisations that can use them (Singer and Brooking, 2019; Rid, 2020).

A few further terms warrant explanation, as they relate to vulnerabilities to information disorders. As part of studying COVID-19 vaccine-related narratives in mid-2020, Smith et al. (2020) highlighted two primary "market failures of the information industry: data deficits and data oversupply" (p.20, Smith et al., 2020). These both relate to the amount of information, particularly credible information, in a discussion relative to the demand. A *data oversupply* results in a crowded information space, where people are easily confused and overloaded by (sometimes contradictory) information, which causes them to disengage. A *data deficit*, in contrast, occurs when there is a lack of credible information about an issue but significant demand for it. Their example of vaccine discussions revealed a lack of understanding of the safety of vaccines and how they work and how necessary they are, but particularly also exposed concerns over the political and economic motivations of those promoting the vaccines, including political leaders, health experts and the health industry. Although data deficits are not deliberately created (experts may not realise what information people require or which require it), it can be vulnerable to the introduction of misinformation and exploited with disinformation. A data deficit can deliberately created, providing an environment in which to build a community around misinformation; the `#ArsonEmergency` discussion discussed in Chapter 5 is an example of this according to Graham and Keller (2020).

2.1.2 Post-dissemination complications

These definitions begin to blur once content has been seeded, however. After disinformation is released, is it still disinformation when an "unwitting agent" (Bittman, 1985, as cited on p.127:4 by Starbird et al., 2019) or "sincere activists" (Starbird and Wilson, 2020) reposts it? Is it misinformation at that point, or simply information to be interpreted by its recipient? Wardle (2017) noted that

> "...social networks allow 'atoms' of propaganda to be directly targeted at users who are more likely to accept and share a particular message. Once

> they inadvertently share a misleading or fabricated article, image, video or meme, the next person who sees it in their social feed probably trusts the original poster, and goes on to share it themselves. These 'atoms' then rocket through the information ecosystem at high speed powered by trusted peer-to-peer networks."

For practical purposes, however, outside discussions of abstract semantics, it is possible to study the spread and effect of misinformation or disinformation campaigns, and to use expert judgement to determine which label to apply. Whether or not a retweeter deep in the retweet chain is aware that the information they are sharing is part of a nation state's foreign influence campaign is irrelevant—the question is how far that information is disseminated and how people react to it that is important.

2.1.3 The vicious cycle of information disorders on society

The effect of such information disorders contributes to a diminishing of trust in official and previously respected sources, resulting in people relying on their social circles for information, which, when that information is incorrect or at least overly biased, can lead to a further loss of trust in official sources, perpetuating the cycle (Kavanagh and Rich, 2018). This new reliance on social circles is an opportunity that online influencers have exploited to financial benefit – by posting regularly about their everyday events, they generate a "constructed friendship"[5] with their followers, a sense of being a close friend without the reciprocation, but that nevertheless results in being trusted commensurately when they offer information, regardless of its veracity (Bruns et al., 2020). Celebrity endorsement of conspiracies can result in the similar unwarranted influence of such ideas (Bruns et al., 2021) as well as provide an information environment with opportunities for proponents of populist politics (Bergmann, 2020). This vicious cycle is presented in Figure 2.2, highlighting the complicated nature of the information space and the progressive nature of the cycle reinforcing distrust in authority figures and the breakdown of trust in societal institutions. The OECD reported in mid-2021 that "in 2020, only 51% of people in OECD countries trusted their government" (p.5, OECD, 2021), and emphasised the need for governments to safeguard trust with transparency and better governance and by reinforcing democracy, particularly while responding to the COVID-19 pandemic.

2.1.4 Real-world impact

Politically- and ideologically-motivated misinformation and disinformation have had significant impacts on society including incidents of violence, such as the `#StopTheSteal` movement culminating the 6 January 2021 storming of the US Capitol Building in Washington, D.C. (Scott, 2021) and widespread anti-lockdown protests around the world (Loucaides et al., 2021; Graham et al., 2020b). In the midst of

[5] https://www.abc.net.au/news/science/2020-12-09/social-media-conspiracy-theorists-5g-covid-19-influencers/12937950. Posted 2020-12-09. Accessed 2021-12-07.

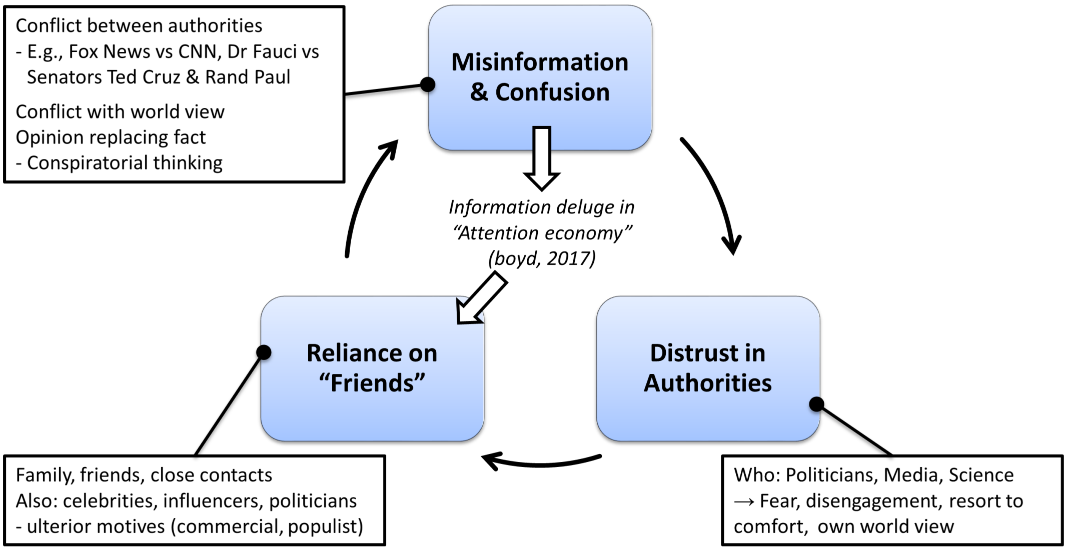

Conflict between authorities
- E.g., Fox News vs CNN, Dr Fauci vs Senators Ted Cruz & Rand Paul
Conflict with world view
Opinion replacing fact
- Conspiratorial thinking
Misinformation & Confusion
Information deluge in "Attention economy" (boyd, 2017)
Reliance on "Friends"
Distrust in Authorities
Family, friends, close contacts
Also: celebrities, influencers, politicians
- ulterior motives (commercial, populist)
Who: Politicians, Media, Science
→ Fear, disengagement, resort to comfort, own world view

FIGURE 2.2. The vicious cycle of misinformation and other information disorders in the information sphere and their effect on society. Misinformation combined with conflict between authorities results in confusion in the population. As a result, their faith in those authority figures and organisations wanes, and fear-based thinking forces them to resort to what they prefer as a source of who they trust. This is typically who they regard as peers, but also extends to their preferred celebrities and politicians and, increasingly, social media influencers, who spend considerable amounts of time gaining the trust of their followers, often to rely on them as a source of income. This reliance on unreliable sources, such as peers and commercially and politically motivated public figures, results in further spread of questionable unreliable information. NB, The lack of a capital in the reference to boyd (2017) is deliberate.

a global pandemic, such as that caused by the COVID-19 coronavirus, which burgeoned in early 2020, the effect of misinformation in this information economy has demonstrably lowered vaccination rates and risked lives as a result (Tasnim et al., 2020; Loomba et al., 2021). Conspiracies have flourished in recent years, whether it be the resurgence of the Flat Earth Society and the increasing rejection of modern science of which it is emblematic (Brazil, 2020), the burning of 5G towers for their role in the COVID-19 pandemic (Bruns et al., 2020), or the QAnon conspiracy that is now regarded as a potential American national security threat (The Soufan Center, 2021a). This puts QAnon on a par with other nationalistic, sovereign citizen and white supremacist movements, which are now referred to by the terms domestic violent extremism (DVE), religiously motivated violent extremism (RMVE) and ideologically-motivated violent extremism (IMVE, ASIO, 2021; DNI, 2021). That said, conspiracy movements, particularly anti-authoritarian ones, have long been a concern for law enforcement and national security agencies (Pitcavage, 2001; Sunstein and Vermeule, 2009), but Bruns et al. (2020) argue that the continued media attention on fringe conspiracies coupled with a lack of effective countering of their narratives by political leaders leaves us vulnerable to them continuing to gain prominence, including through attention in the mainstream media.

2.1.5 Information operations and disinformation campaigns

The vicious cycle presented in Figure 2.2 can be fostered deliberately with propaganda as part of information operations and disinformation campaigns. Historically this has been a (typically covert) element of foreign policy, prominent during the Cold War (Rid, 2020), but recently, it is clear that such techniques, in conjunction with the communication pathways afforded by the internet and social media, can be used by ideological and activist groups or even motivated individuals just as easily, if they are sufficiently well-resourced. Bradshaw et al. (2021) recently found that such activities were conducted in 81 countries in 2020.

Starbird et al. (2019) discuss *Strategic Information Operations* (SIOs), distinguishing them from disinformation operations, making it clear that SIOs are designed as "manipulation efforts", not so much "done to human crowds rather than *something human crowds do*" when appropriately manipulated (p.127:3, Starbird et al., 2019). Features of the internet that support advertising, such as population segmenting and micro-targeting, can also be used to focus political messaging in the same way as commercial messaging—Cambridge Analytica, the political data analytics company, used these techniques during 2016 for both the Brexit referendumand the US presidential election (*Understanding Mass Influence* 2021).Guided by psychological research into personality trait prediction from social media behaviour (Kosinski et al., 2013; Youyou et al., 2015), Cambridge Analytica used personal information fraudulently sourced from Facebook to target political ads at niche community segments (e.g., by geography, demographics and worldview preferences) with the aim (in the US) of convincing Republican-leaning people to vote and dissuading Democrat-leaning people from voting (Grassegger and Krogerus, 2017).

Starbird et al. (2019) clarify that the purpose of an information operation can be to promote an individual, group, or idea, as well as mobilise people against it through polluting communication channels (e.g., chat rooms, discussion channels or hashtags). Woolley (2016) documented the use of political bots for both promotion and pollution (discussed in Section 2.2.1), and King et al. (2017) discussed the same techniques being employed by the "Fifty-Cent Army" in China. In contrast, disinformation campaigns are primarily aimed at eroding people's perception of the distinction between facts and non-facts – in doing so the population tends to believe those they trust, including their preferred politicians (p.10, Rid, 2020). Furthermore, Starbird et al. (2019) argue that modern SIOs and disinformation campaigns are now inherently "*participatory* in nature" (p.127:5, Starbird et al., 2019), meaning that, as alluded to above, suitably inclined members of the public are recruited to the campaign (whether they realise it or not) by appealing to them to engage and repost the message, such as occurred during a recent anti-White Helmet campaign[6] (Starbird and Wilson, 2020).

[6] The White Helmets are volunteer medics operating in Syria, providing aid to civilians harmed in the civil war.

In general, Starbird et al. (2019) determine that SIOs fall into three categories:

Orchestrated Highly *orchestrated* campaigns are directed from the top down with a traditional command and control structure (e.g., the Russian Internet Research Agency's (RU-IRA) activities during the 2016 US presidential election, Chen, 2015; Mueller, 2018).

Cultivated *Cultivated* campaigns supporting existing issue-motivated communities, seeking to exploit fissures in society, by disparaging opponents and disseminating preferred narratives[7] or alternatively hijacking and redirecting existing movements.[8]

Emergent Some campaigns *emerge* from the community gathering around particular false narratives and conspiracies, especially promoted by "alternative news" sites, in which discussion shifts from general theorising to specific political disinformation. An example of an emergent campaign was the "false flag" conspiracies regarding mass shootings in the US promoted by the InfoWars website[9] and then amplified by other "alternative" media and government-affiliated news sources in Russia (Benkler et al., 2018). Such patterns have also been observed in studies of the spread of COVID-19 and 5G conspiracies on social media (Bruns et al., 2021).

Elsewhere, Starbird (2019) also emphasises that disinformation is not simply false information; it is often a combination of true and false information in layers that prevent easy identification, such as the use of a false context. By appealing to the "sincere activists" and "unwitting agents", the content will also be embellished, as well as amplified, making it exceedingly hard to identify what content is genuine disinformation and what is simply misinformation. Other research has demonstrated that false news spreads much more widely than true news, potentially due to its novelty and the emotional reactions it generates, meaning that "false news spreads more than the truth because humans, not robots, are more likely to spread it" (p.1146, Vosoughi et al., 2018). Rather than classifying a single piece of information as true or false, Starbird (2019) argues it is more important to examine how it contributes to the broader campaign, and what the aims and methods of the campaign are.

This mixing of true and false content is clearly present in a propaganda strategy dubbed the "firehose of falsehoods" (Paul and Matthews, 2016). The strategy's four defining traits, observed since at least the 2014 annexation of Crimea by Russia, are:

[7] The campaign against the White Helmets in Syria, who rescue victims of the civil war, was conducted by a combination of social media activists and Syrian government accounts supported by Russian official news and alternative media (Starbird and Wilson, 2020).

[8] The `#StopAsianHate` campaign was reframed from being about anti-Asian racism inspired by COVID-19 misinformation to focus on those promulgating the 'laboratory-leak' conspiracy theory as a way to defend China and the Chinese Communist Party (Zhang, 2021). This theory states that the COVID-19 coronavirus originated, not in a wet market in Wuhan, but in a nearby virus research institute with military ties.

[9] https://www.msnbc.com/opinion/infowars-school-shooting-lies-cost-alex-jones-put-extremists-alert-n1280803. Posted 2021-10-06. Accessed 2021-12-09.

high volumes of messaging across many platforms; the content is "rapid, continuous and repetitive"; the messaging need not be true or even realistic; and the content need not even be consistent (p.2, Paul and Matthews, 2016). The fact that the content is repetitive means that detection methods that focus on amplification will remain useful while the strategy is employed, but the high volume of the messaging and its rapidity indicate that being able to respond quickly is vital. As Paul and Matthews (p.5, 2016) state, "first impressions are very resilient" and "Repetition leads to familiarity, and familiarity leads to acceptance", so countermeasures either need to identify and then stop the amplification before it spreads too far, or be very convincing or distracting to quickly change the narrative. Clarity and consistency in messaging is key for this to occur, perhaps even using the same amplification and multi-channel strategies, but not necessarily needing to be covert – Ronald Reagan's public call of "Mr Gorbachev, tear down this wall!" was public, but also clear, consistent and well-covered by the world media (p.133, Kent, 2020).

When influence campaigns are enabled with automation and access to big data resources (such as Cambridge Analytica's Facebook data), they have been referred to with the label *computational propaganda* (Shorey and Howard, 2016; Woolley and Guilbeault, 2018). More is said about computational propaganda in Section 2.2.4.

2.1.6 "Ampliganda"

Further complications have arisen with the increase in public awareness of such information campaigns, as grassroots movements start to use the same techniques, a concept which has been dubbed *ampliganda* (DiResta, 2021). In mid-2020, TikTok users registered interest in a political rally only to not attend and encouraged friends to do the same, resulting in a majority of empty seats for the venue and embarrassment for the politician.[10] Also in 2020, in response to the George Floyd riots, white supremacist users promoted `#WhiteLivesMatter`, only to have the hashtag polluted with pictures of Korean pop (K-pop) band members by fans.[11] The term ampliganda is designed to emphasise that it is simply an opinion that is being amplified, rather than something portrayed as a fact, and could be regarded as an agenda-driven meme (Dawkins, 1989). This technique can can also be detrimental when promoting misinformation (e.g., `#Ivermectin`, a non-effective COVID-19 treatment, and `#SaveTheChildren`, promoted by QAnon, The Soufan Center, 2021b). DiResta (2021) explained that a further danger of ampliganda occurs when the instigators are not careful with messaging (e.g., hashtag phrasing) and lose control of the conversation, overtaken by those with other agendas. Because of examples like these, studying coordinated amplification and the evolution of campaigns based on such techniques remain an important topic of research.

[10] https://www.nytimes.com/2020/06/21/style/tiktok-trump-rally-tulsa.html. Posted 2020-07-21. Accessed 2021-11-29.

[11] https://www.bbc.com/news/technology-52922035. Posted 2020-06-04. Accessed 2021-11-23.

Despite these difficulties, there are continuing efforts to measure the effects of information operations (e.g., Nimmo, 2020; Zannettou et al., 2017; Zannettou et al., 2019). For our part, we contribute to the characterisation of polarised online groups (discussed in Part II) and the detection of groups behind coordinated amplification and other related behaviours (discussed in Part III). Kumar and Shah (2018) provides a recent and detailed survey of information disorders and recent software-based techniques designed to address them.

2.1.7 Behaviour, not content

The complexity described in this section suggests that detection methods that do not rely on content analysis may have an advantage. Even analyses that rely on keywords, such as hashtags, to identify campaigns make assumptions about whether each post is promoting or attacking the keyword's concept. As a result, we favour the use of SNA and network methods, basing the majority of our behavioural analyses on the timestamped interactions between social media accounts, only looking to their content as evidence for confirmation or characterisation.

2.2 Online Influence and Inauthentic Behaviour

Since before the first documented use of social media to artificially influence an election in the 2010 special election in Massachussetts, America (Metaxas and Mustafaraj, 2012), people have been exploring how to exploit the features of the internet and social media that otherwise bring us benefits. The activities documented in that election are an example of *astroturfing*, the practice of generating fake grassroots movements, creating the impression of popular support for an idea or person through coordinated deception (Ratkiewicz et al., 2011; Cho et al., 2011). Exploitable features of the two-edge sword of social media include the following:

Specificity and reachability The ability to direct marketing to specific audiences that connects businesses with the most receptive customers also enables highly targeted non-transparent political advertising (Angwin et al., 2017; Woolley and Guilbeault, 2018) and the ability to expose people to propaganda and recruit them to extremist organisations (Berger, 2014; Berger and Morgan, 2015; Badawy and Ferrara, 2018). This was mentioned in Subsection 2.1.5.

Anonymity The anonymity that supports the voiceless in society to express themselves also enables trolls to attack others without repercussions (Hine et al., 2017; Burgess and Matamoros-Fernández, 2016; Bot Sentinel, 2021).

Automation The automation that underpins benign services from news aggregators to art projects also facilitates social and political bots that seek to manipulate public opinion (Ferrara et al., 2016; Woolley, 2016; Bessi and Ferrara, 2016; Cresci, 2020).

In summary, targeted marketing and automation coupled with anonymity provide the tools required for potentially significant influence in the online sphere, perhaps enough to swing an election.[12]

In this section, we introduce a number of the elements of inauthentic behaviour that exploit features of OSNs to influence others. The phrase "inauthentic behaviour" was coined by Facebook in the context of "coordinated inauthentic behaviour" (CIB, Gleicher, 2018), but a clear actionable definition continues to elude the major OSNs. Inauthentic behaviour is defined as "the use of Facebook or Instagram assets (accounts, pages, groups or events), to mislead people or Facebook" regarding identities and true purposes, popularity of said assets, the origins of content or in order to evade its Community Standards.[13] "Coordinated" inauthentic behaviour is defined as the use of "multiple Facebook or Instagram assets, working in concert to engage" in inauthentic behaviour.[14] Despite that, recent revelations indicate that Facebook in particular has overlooked some of their own terms and conditions for prominent users.[15] Douek (2020), a legal researcher, explained how the fuzzy definition of CIB and its various interpretations by platforms mean that they give themselves the flexibility to choose to enforce rules based on business pressures rather than "seriousness". This comment arose from the revelation by a whistleblower that Facebook has been slow to react to reports of CIB in low-priority environments and countries, meaning that some governments are able to persist with CIB to support themselves for many months and even years after CIB is reported.[16]

Two features of OSNs, in particular, are used to engage in and maximise the effectiveness of inauthentic behaviour: automation and anonymity, the second to enhance the first.

2.2.1 Automation

Automating social media activity is a relatively simple programming task, especially when connecting to OSN APIs, which provide direct access to their features and capabilities. The term *bot* refers to software that can carry out repetitive tasks that a human would otherwise have to do, such as posting or retrieving information (Ferrara et al., 2016). Crawlers and spiders which populate search indices or archive websites

[12] Inauthentic influence has been previously observed in Australia elections (Waugh et al., 2013), and an Australian Senate select committee investigating foreign interference threats recently warned of potential for interference in the 2022 Australian federal election. Source: https://www.theguardian.com/australia-news/2021/dec/20/morrison-warned-foreign-interference-campaign-on-social-media-is-a-serious-risk-to-australias-election. Posted 2021-12-20. Accessed 2022-01-05.

[13] https://transparency.fb.com/en-gb/policies/community-standards/inauthentic-behavior/. Accessed 2021-11-24.

[14] *ibid.*

[15] https://www.theguardian.com/technology/2021/sep/13/facebook-some-high-profile-users-allowed-to-break-platforms-rules. Accessed 2021-11-24.

[16] https://www.theguardian.com/technology/2021/apr/12/facebook-fake-engagement-whistleblower-sophie-zhang. Posted 2021-04-12. Accessed 2021-12-07.

for posterity are good examples of such automation. Other bots conduct "social listening" for trends or intelligence gathering, and some scan for copyright violations (Woolley, 2016). Some bots are clearly benign, such as joke bots[17] and art projects.[18] Oentaryo et al. (2016) would class these as *consumer* or *broadcaster* bots in their taxonomy. In contrast, the related term *daemon* refers more to services or persistently running processes that engage in system administration.[19] Automatons, in this sense, have existed for decades, and have certainly been useful since the inception of the internet, though they have long also been associated with internet blights, such as spam (Aiello et al., 2012), which is Oentaryo et al. (2016)'s third category.

2.2.1.1 Social bots

For more than a decade now, another breed of bot has been active on social media: *social bots* (Ferrara et al., 2016; Cresci, 2020). Hwang et al. (p.40, 2012) define them as designed for "creating substantive relationships among human users . . . and shaping the aggregate social behavior and patterns of relationships". Their definition nicely encapsulates the notion that social bots are meant to look human, act like humans, and interact with humans, to shape discussions and influence humans. Furthermore, they are an effective tool of influence, having been implicated in amplifying misinformation (Shao et al., 2018a). They may be fully or partially automated. Grimme et al. (2017) break down social bot behaviour as consisting of 1) building up a network of followers, 2) behaving realistically, i.e., exhibiting plausibly human-like patterns of life, and 3) generating content to interact with other users. Partially automated accounts are also referred to as *cyborgs* (Chu et al., 2012). These hybrid systems can disseminate content generated by humans via automation to followers that have been acquired also via automated strategies. Until recently, the content for bot tweets needed to be crafted by humans to be plausible, but recent advances in natural language generation suggest that automated language could soon be very hard to identify by eye – a bot using the GPT-3 model (Brown et al., 2020) remained active on Reddit for a week before it was identified and its account shut down (Heaven, 2020).

2.2.1.2 Overt bots

A bot's purpose may be benign or malicious, and vary in degree. Social bots may be *overt*, clearly automated, when they are designed for interaction, such as the chatbots that can be seen on commercial sites that provide the first layer of support for customers. When these kinds of bots pretend to be human, it is clearly not a malicious attempt to deceive. Simple spambots are designed for marketing purposes, and though some could be thought to be malicious, they are often better described as irritating. Then there are mobile device assistants, such as Apple's Siri or the Google

[17] E.g., https://www.abc.net.au/news/2021-08-04/stand-up-comedy-being-written-by-robots/100342712. Accessed 2021-11-24.

[18] E.g., https://inspirobot.me/. Accessed 2021-11-24.

[19] https://en.wikipedia.org/wiki/Daemon_(computing). Accessed 2021-11-24

Assistant, which make a mobile device's functions available through spoken dialogue. These are clearly benign bots.

2.2.1.3 Covert bots

In contrast, some bots are *covert* and hide their true identity deliberately, playing on humans' poor ability to judge increasingly sophisticated automated behaviour (Edwards et al., 2014; Guilbeault, 2016; Cresci et al., 2017b), especially when attention is so limited and vulnerable to manipulation (boyd, 2017; Ciampaglia et al., 2018; Lou et al., 2019). Malicious bots include 'fembots', which run on dating sites (Newitz, 2015), stock manipulating bots that spread fake news to scare investors (Ferrara et al., 2016), bots involved in 'pump and dump' and other financial schemes (Cresci et al., 2019; Pacheco et al., 2021), and political bots used to interfere with elections and political discussions (Woolley, 2016; Bessi and Ferrara, 2016; Grimme et al., 2017; Rizoiu et al., 2018). These financial and political examples are further instances of astroturfing.

2.2.1.4 Detection surveys

In their widely cited survey of social bot detection methods, Ferrara et al. (2016) divided them into three categories: systems that make use of the account's social network, systems that rely on crowd intelligence, and systems that use machine learning to distinguish bots from typical users by discovering highly discriminatory features. The primary issues considered include not just the question of whether an account is human-driven or automated, but whether the detection can be conducted at scale.

Network methods Early network-based methods focused on follower relations and assumed that bots would mostly form cliques of *sybils* to build credible friend and follower counts. Research found that bots could easily infiltrate communities of genuine users, who often accepted random friend requests (on Facebook, Tumblr and Twitter). Experiments with automated strategies to build follower networks have demonstrated that simply being active and retweeting or reposting is sufficient to avoid most detection and to gather followers (Freitas et al., 2015; Grimme et al., 2017; Fazil and Abulaish, 2020). These early detection methods were not adopted by OSNs due to high false positive rates, which would cause the OSNs to frequently flag genuine users and bad press would result if they were suspended.

Crowdsourcing Human investigators produce lower false positive rates but require time and training, and so the issue of scale is a challenge. Established platforms already have too many users to rely on humans, and as artificial intelligence techniques such as DeepFakes (Hwang, 2020) and natural language generation (Brown et al., 2020) mature, telling agenda-driven humans from bots will only become more difficult. This distinction is further obscured by humans deliberately employing deception (e.g., "users of the `#NotABot` hashtag were no more likely to be human than other users", p.203, Bellutta et al., 2021).

Machine learning methods These methods rely on extracting features from accounts based on their behaviour and metadata, and then distinguishing between genuine human users and automated accounts using classification or clustering. APIs provide a wealth of metadata, from which features can be extracted or calculated. Botometer (formerly BotOrNot, Davis et al., 2016) is an ensemble classifier relying on six sub-classifiers, each focused on a different category of features. The categories include (i) *network* features, drawn from the account's follower and friend connections, (ii) *user* profile features, (iii) *friend* profile features, (iv) *timing* features, (v) *content* features, and (vi) *sentiment* features. Updates to Botometer mean it is no longer English-centric and it now includes a Complete Automation Probability (CAP) measure, a sophisticated Bayesian-based calculation of the likelihood that a given account uses automation (Yang et al., 2019). Other ML-based detection systems include `tweetbotornot2`,[20] RTBust (Mazza et al., 2019), BotSlayer (Hui et al., 2019), and Birdspotter (Ram et al., 2021), but all rely on access to labelled datasets and need constant re-training, a shortcoming highlighted by Alizadeh et al. (2020). Alizadeh et al. (2020) built classifiers trained to detect evolving troll campaigns, addressing the question of retraining by evaluating how well classifiers trained on a month's data are at identifying trolls in the next month. The feature sets, APIs and policies of the platforms themselves are also always in flux, affecting the availability and accessibility of data for classification training and detection.

That said, a number of labelled datasets of Twitter data relating to bots and genuine users have been published, including recently by Feng et al. (2021), but it is unclear what conditions they have been released under[21] and how they will age. Feng et al. themselves note that Botometer's diminished performance on the new TwiBot-20 dataset indicates "the real-world Twittersphere has shifted and Twitter bots have evolved to evade previous detection methods" (p.4491, Feng et al., 2021), and thus it should be expected that TwiBot-20 will also soon become outdated.

In his follow-up survey, Cresci (2020) highlights the necessarily adversarial nature of bot detection, in that bots evolve (or their designers revise their functionality) to avoid the state-of-the-art detection methods. He also notes that the future inauthentic behaviour detection systems will need to focus on how malicious accounts (automated or otherwise) coordinate or engage in "orchestrated activities" (Grimme et al., 2018).

[20] https://github.com/mkearney/tweetbotornot2. Accessed 2021-11-24.

[21] These benchmarks are provided in structured formats such as CSV or JSON, but are not in the raw JSON form provided directly from Twitter's APIs, perhaps avoiding issues we raise in Part I, the Information Environment.

We explore the literature on coordinated behaviour in Section 2.6 and provide our own contribution to its detection in Part III.

Latah (2020)'s thorough survey of the state-of-the-art in bot detection methods provides an extensive taxonomy of social bot and botnet strategies and detection methods, as well as countermeasures. One point in particular that Latah makes is the importance of having consistent but diverse benchmark datasets, which is a topic we address directy in Section I.1.

2.2.2 Malicious actors

Several types of accounts engage in different malicious online behaviour, including trolls, sockpuppets, vandals and fake reviewers. Here we summarise their key features.

2.2.2.1 Trolls

Trolls are users that actively attempt to antagonise and harass, or at least sow division (Wardle and Derakhshan, 2017), and they can have a variety of motivations, such as:

Financial (i.e., paid) Examples include employees of the RU-IRA (Chen, 2015; Mueller, 2018; Dawson and Innes, 2019) or per-post bounty-based members of the Chinese "Fifty-Cent Army" (King et al., 2017).

Ideological The ideology can vary and may be, e.g., political, social or racial. For example, Milo Yiannopoulos was banned from Twitter for coordinating racist and sexist harassment of the cast of the all-female reboot of the film Ghostbusters, particularly the Black comedienne Leslie Jones (Romano, 2016), having previously championed the anti-feminist `#GamerGate` movement (Burgess and Matamoros-Fernández, 2016; Massanari, 2016).

Entertainment Such activities are typically conducted at the expense of a target individual or group, e.g., the `#BikiniBridge` challenge initiated on the 4chan forum aimed to convince young women to lose dangerous amounts of weight to achieve the latest fad 'fitness goal' (Drenten and Gurrieri, 2018).

The motivation is not always clear, however, such as of the persistent coordinated attacks on the Duchess of Sussex, Meghan Markle (Bot Sentinel, 2021).

Communities have also been observed engaging in *brigading*, coordinated trolling of individuals and other communities, by amplifying abuse or suppressing through downvoting (Massanari, 2016). Kumar et al. (2018) studied particular attack patterns and their effectiveness on Reddit, while Datta and Adar (2019) identified raids by automatically detecting accounts' 'home' communities and then analysing the aggressiveness of content they posted to other communities that triggered community sanctions, also on Reddit. Mariconti et al. (2019) used temporal analysis to detect preparations on 4chan for attacks on particular YouTube video comment sections. In terms of contributions to objectionable content online, 4chan has provided proportionally more than most online communities (Hine et al., 2017). Some attacks have a distinct political aspect, however, such as the 2020-21 aggression directed towards the state government of Victoria and its premier, Daniel Andrews, during the COVID-19 pandemic (Graham et al., 2020b), as well as the "tidal waves of abuse" directed at UK government COVID advisors.[22]

It should be noted that, under the right conditions, typical users can engage in troll-like behaviour (Cheng et al., 2017). This requires a confluence of the user's mood and surrounding discussion, and may go towards explaining some of the aggressive behaviour witnessed in the context of recent bushfire-related discussions, which we examine in Chapter 5.

2.2.2.2 Sockpuppets

Sockpuppets are accounts created with false personas that are used to promote a particular narrative and are often used in multiples, run by an individual (Chen, 2015; Kumar et al., 2017b). Carefully curated fake individual personas can have significant influence on the broader discussion, such as the RU-IRA's `@TEN_GOP` account during the 2016 US presidential campaign (Nimmo, 2017). At scale, using their Dawn of Glad Tidings mobile app the terrorist group ISIS co-opted Twitter accounts of followers to boost their army's appearance as they invaded Mosul in 2014 (Berger, 2014; Berger and Morgan, 2015). Further, sockpuppets are more likely to be used strategically. Kumar et al. (2017b) observed that sockpuppets started few discussions, interacted with each other more, forming tighter egonets, but also posted in the same discussions more often when run by the same person. Dawson and Innes (2019) identified several strategies that the RU-IRA used to build their audiences that highlight the effort employed to curate these accounts. In particular, the practice of *narrative switching* requires significant planning and execution to do effectively. Using this strategy, an account is prepared with a particular persona (as defined by the profile information, such as screen handle or name, profile pictures, background pictures, and account description) and promotes a particular narrative for a period. At some point the account goes dormant and all its posts are deleted. After a further period, the account's persona is changed and then it resumes posting, promoting a different narrative. Dawson and Innes (2019) also identified the use of the same *follower fishing* strategy experimented with by Grimme et al. (2017) and Fazil and Abulaish (2020) to build a follower base and infiltrate the broader community, enabling them to avoid a number of the network-based detection methods mentioned above.[23]

[22] https://www.theguardian.com/world/2021/dec/31/uk-governments-covid-advisers-enduring-tidal-waves-of-abuse. Posted 2021-12-31. Accessed 2022-01-06.

[23] Follower fishing is a strategy that involves following random accounts, which many people reciprocate out of politeness. If after a short period of time (e.g., one day) they do not reciprocate, then they are unfollowed. This is technique is also used to artificially inflate an account's reputation score, which is based on an account's friend and follower scores.

Automation can value-add to these strategies, leading to a commercial industry in hosting botnets of fake followers (Aggarwal and Kumaraguru, 2015; Woolley, 2016; Confessore et al., 2018). Such botnets can become very large — in 2017, a network of 350,000 bots swapping Star Wars quotes was discovered (Echeverria and Zhou, 2017). Its purpose was not clear but its potential for large-scale influence was (e.g., through pollution or amplification). Additionally, botnets posting human-generated content can survive in the wild for considerable periods of time (Grimme et al., 2018; Fazil and Abulaish, 2020).

2.2.2.3 Vandals

Vandals modify and damage online content, including through the use of automation. A particular challenge for crowdsourced efforts, such as Wikipedia and Wikidata, is conflict between contributors on contentious pages (Giles, 2005; Sarabadani et al., 2017). Content-editing bots have been observed to delete and replace each others' edits for years (Tsvetkova et al., 2017).

2.2.2.4 Fake reviews

A particular consternation for commercial entities is fake reviews, organised campaigns of which are referred to as *crowdturfing* (Wang et al., 2012). Typically, these involve commercial interests, such as companies paying for good reviews of their products or bad reviews of their competitors' products. The restaurant and travel industry struggles with fake reviews, with up to a sixth of them being fake (Kumar et al., 2017a), and user-initiated reporting systems are also not always reliable (Freeman, 2017). Even streaming sites need to contend with fake activity, to avoid stream view counts being manipulated (Shah, 2017). Synchronicity is used as a detection tool in this field also (Li et al., 2017a).

Some attacks on commercial ventures appear to be ideological brigading or astroturfing, which grow as sympathetic users are recruited, such as boycotts. Examples of these include anti-diversity attacks on movie reviews for the films "Mad Max: Fury Road" (2015)[24] and "The Last Jedi" (2017),[25] which received one star reviews from men's rights activist and alt-right communities for having prominent female and non-White characters. Racially-motivated attacks on the Black Panther (2018) film were foiled before they could launch.[26]

[24] https://www.gq.com/story/mra-calls-for-mad-max-boycott. Posted 2015-05-15. Accessed 2021-12-06.

[25] https://www.gq.com/story/last-jedi-spam-rotten-tomatoes. Posted 2017-12-20. Accessed 2021-12-06.

[26] https://www.polygon.com/2018/2/2/16963988/rotten-tomatoes-black-panther-review-bombing-alt-right. Posted 2018-02-02. Accessed 2021-11-24.

2.2.3 Further inauthentic activities

As referred to above, inauthentic behaviour is most effective when used en masse in a coordinated manner to mimic genuine community activity. Some coordinated inauthentic activities are not easy to identify, due to limitations on the data provided by OSNs (see Section I.1). For example, the practice of name switching to avoid detection has been observed in the literature (Mariconti et al., 2017; Ferrara, 2017). Because the identifier of a Twitter account is separate from the account's handle, the handle can be modified, and thus accounts can agree to swap them. This particular strategy is used by triples of accounts to avoid being identified and reported by genuine users.

Other areas of inauthentic online behaviour that have received scholastic attention include the trade in fake follower accounts mentioned earlier and, more broadly, the study of rumours.

A common measure of online popularity is *reputation*, defined as the ratio between an account's number of followers and their friends and followers (i.e., $\frac{|followers|}{|followers|+|friends|}$). The closer to 1 this expression is, the more an account is followed (implying desired, admired, or respected) than follows others. This also helps obscure inauthentic accounts, presenting them as typical users, if their reputation is around 0.5, implying they have equal numbers of friends and followers. As a result, there are commercial opportunities in providing accounts to act as followers (Aggarwal and Kumaraguru, 2015; Aggarwal et al., 2018; Confessore et al., 2018), and using automation is the easiest way to establish and manage these accounts at scale.

The notion of rumours is closely related to misinformation and disinformation, inasmuch as it is information that may be used in a benign or malicious manner, but its distinction is that its veracity is not confirmed. Kumar et al. (2017a) associate rumours with hoaxes, which are arguably a type of disinformation (as they are for personal gain or entertainment), and their dissemination has been long studied. Notable contributions have been Vosoughi et al. (2018)'s comparison of the flow of false and true news online, and the Hoaxy hoax-tracking system (Shao et al., 2016). Given the emergence of *citizen journalism* (Gillmor, 2006), through which social media enables any individual to broadcast and report on events around them to the world (which draws concerns of bias and balance), researchers have investigated ways to test the credibility of rumours based on social media activity (Mitra et al., 2017).

2.2.4 Computational propaganda

We introduced computational propaganda as the combination OSNs, automation and big data resources in ways designed to influence public opinion in Section 2.1.5. Here, we explore how these elements relate to the goal.

Many online influence techniques have been inspired by incidental discoveries by curious and motivated technical users. boyd (2017) traces the history of those who found they could game OSN trending algorithms, and then started to exploit this ability to 'hack the *attention economy*' for their own (or their client's) benefit. In this context, the term "attention economy" refers to using economic principles to the scarce commodity of people's attention, which is necessarily limited due to the deluge of information they receive through the online and offline media (Simon, 1971). Similarly, Gorwa and Guilbeault (2017) discuss how young political activists in the UK found they could influence voters by first attracting potential voters through flirtatious bots on the Tinder dating platform. In a participatory study of the major political campaigns in the 2016 US election, Woolley and Guilbeault (2018) found that the campaigns in America were also very experimental in their approaches: "We will throw anything against the wall and see what sticks" stated one Republican National Committee employee (p.195, Woolley and Guilbeault, 2018).

The first documented use of social media to artificially influence voters was in the 2010 special election in Massachussetts, America (Metaxas and Mustafaraj, 2012), and people have been exploring how to exploit the features of the internet and social media to manipulate public opinion in a more organised and repeatable fashion. These features that enable this include

- the ability to target marketing to specific audiences that connects businesses with the most receptive customers also enables highly targeted non-transparent political advertising (Chessen, 2017; Angwin et al., 2017; Woolley and Guilbeault, 2018), which also facilitates the ability to expose people to propaganda and even recruit them to extremist organisations (Berger and Morgan, 2015; Badawy and Ferrara, 2018; Singer and Brooking, 2019; Waldek et al., 2020);
- the anonymity that supports the voiceless in society to express themselves also enables trolls to attack others without repercussions (Hine et al., 2017; Burgess and Matamoros-Fernández, 2016; Bot Sentinel, 2021); and
- the automation that enables news aggregators also facilitates social and political bots (Ferrara et al., 2016; Woolley, 2016; Ferrara, 2017; Cresci, 2020).

In summary, targeted marketing and automation coupled with anonymity provide the tools required for potentially significant influence in the online sphere, perhaps enough to swing an election. Recent surveys have found deliberate attempts to manipulate public opinion are widespread, having been observed in at least 81 countries, well-resourced (nearly US$10m has been spent on political advertising), and often directed internally (i.e., used domestically or against domestic targets, and not used for foreign interference, Bradshaw et al., 2021).

It is this "assemblage of social media platforms, autonomous agents, and big data tasked with the manipulation of public opinion" (p.185 Woolley and Guilbeault, 2018) that defines 'computational propaganda'. The practice can take many forms: Twitter botnets, sockpuppets on Facebook, YouTube and Instagram, or chatbots on Tinder, Snapchat and Reddit (Howard, 2018). It is clear that the automation of bots and botnets is only one element of the propaganda system, which is why devoting attention to how accounts coordinate their behaviour rather than to whether or not they are automated will be a better use of resources in the long term. Part III of this thesis focuses on detecting and characterising this coordination.

A second significant element to computational propaganda is the use of highly targeted advertising mechanisms offered by the OSNs, which are not publically announced, even if the advertising is political – these are known as 'dark posts'.[27] By choosing from the thousands of niche categories available, it is even possible to target a single individual (González-Cabañas et al., 2021). Not only are the categories highly specific, but they have also raised ethical questions. In response to a ProPublica investigation (Angwin et al., 2017) in the wake of the 2016 US presidential election, Facebook removed 5,000 discriminatory categories (Howard, 2018). ProPublica had discovered, by using Facebook's self-service ad-buying facility, it could target thousands of user who had expressed anti-Semitic sentiments. These categories were available because they were derived algorithmically from user behaviour, and had not all been manually reviewed. Recent reporting indicates this is an ongoing issue, four years later.[28] The influence of platform algorithms on what kind of user behaviour they encourage has been of concern for some time (e.g., Pariser, 2012).

To target a particular audience, advertisers will look to personality traits and influence techniques, because different techniques will work better for different personality traits (Cialdini, 2007). Research leading up to 2016 found that quantifiable personality traits could be predicted from Facebook activities (Kosinski et al., 2013; Youyou et al., 2015). Investigations by Grassegger and Krogerus (2017) revealed that, using this research and a significant amount of fraudulently obtained Facebook data, the data science firm Cambridge Analytica guided the advertising of the Brexit LEAVE campaign[29] and the Trump election campaign,[30] contributing to their successful conclusions (though how effectively is a matter of some debate). Although later revelations caused Cambridge Analytica to be banned from Facebook,[31] thinktanks such as the Atlantic Council believe these kinds of activities will continue to occur. Some predict that the combination of big data analytics, psychometric profiling and machine learning will be used to develop personalised propaganda based on personality traits, political, religious, sexual and gender preferences, and demographic information, all of which will improve over time as more data is collected (Chessen, 2017).

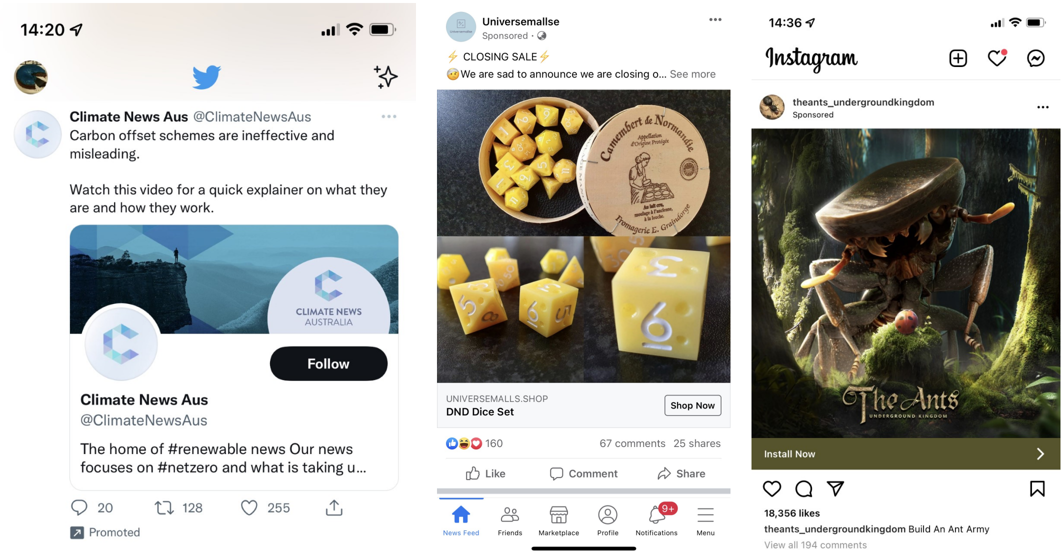
[27] https://insense.pro/blog/dark-posting-on-facebook-what-is-a-facebook-dark-post. Posted 2021-10-21. Accessed 2021-11-25.

[28] https://edition.cnn.com/2021/12/02/tech/facebook-vaccine-holocaust-misinformation/index.html. Posted 2021-12-03. Accessed 2021-12-03.

[29] https://www.theguardian.com/technology/2017/may/07/the-great-british-brexit-robbery-hijacked-democracy. Accessed 2021-11-25.

[30] https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election. Accessed 2021-11-25.

[31] https://www.theguardian.com/uk-news/2018/mar/20/cambridge-analytica-execs-boast-of-role-in-getting-trump-elected. Accessed 2021-11-25.


14:20

Climate News Aus @ClimateNewsAus
Carbon offset schemes are ineffective and misleading.

Watch this video for a quick explainer on what they are and how they work.

Climate News Aus
@ClimateNewsAus

The home of #renewable news Our news focuses on #netzero and what is taking u...

20 128 255

Promoted

(a) Twitter.

Universemallse
Sponsored ·

⚡ CLOSING SALE ⚡
😲We are sad to announce we are closing o... See more

UNIVERSEMALLS.SHOP
DND Dice Set

160 67 comments 25 shares

(b) Facebook.

14:36

theants_undergroundkingdom
Sponsored

18,356 likes
theants_undergroundkingdom Build An Ant Army
View all 194 comments

(c) Instagram.

FIGURE 2.3. Promoted posts observed on the author's Twitter, Facebook, and Instagram activity feeds. Note the ability to like (favourite, react to or like, respectively), comment on and share the posts, and the engagement they have already generated by the time the screenshots were taken.

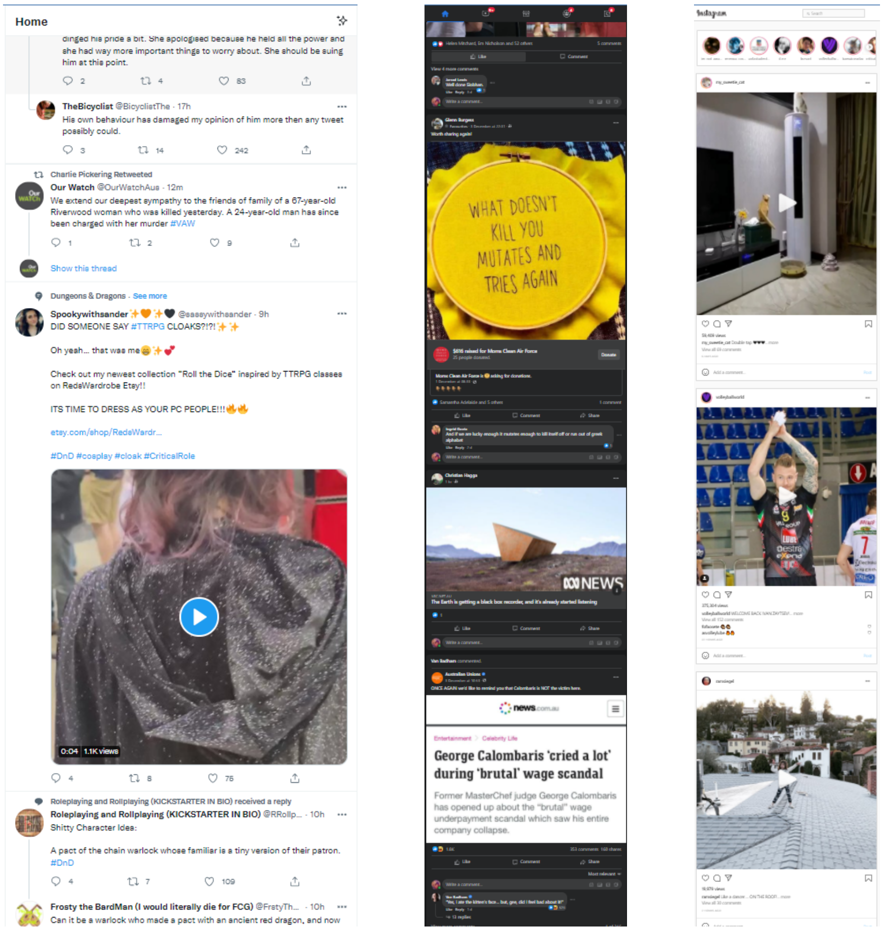
Increasingly, advertisements and promoted posts on OSNs, such as Twitter, Facebook and Tumblr, are specifically designed so they can be interacted with through liking, commenting and sharing (e.g., Figure 2.3), so much future computational propaganda may not need bots and trolls to spread it, but simply the OSNs advertising distribution systems.

2.3 Online Social Networks and Social Media Analytics

The primary vehicle for much of the online behaviour discussed so far is the OSNs that enable us to connect with others, establishing and maintaining relationships and forming communities around shared values. Major OSNs include Facebook, Twitter, Reddit, Instagram and WhatsApp (three of which are owned by Meta, formerly Facebook[32]). Many of them present the user with an infinite activity feed, filled with posts by accounts the user follows interspersed with their own. The selection of posts is decided by opaque recommendation algorithms and ordered either according to posts' temporal information or personalised by other algorithms, all finely tuned to maintain engagement (e.g., Figure 2.4). Additionally, many OSN features have analogies, so for most people OSNs differ not so much by what the platform will let them *do* but more by which of their friends is already present to connect with. Table 2.1 presents examples of common features of some of the major OSNs.

Due to these commonalities, as a researcher it is convenient to initially target one platform, knowing that, with careful design, analytics developed for that platform

[32] https://www.bbc.com/news/technology-50838013. Posted 2019-12-18. Accessed 2021-11-29.



Home

dinged his pride a bit. She apologised because he held all the power and she had way more important things to worry about. She should be suing him at this point.

2 4 83

TheBicyclist @BicyclistThe · 17h
His own behaviour has damaged my opinion of him more then any tweet possibly could.

3 14 242

Charlie Pickering Retweeted
Our Watch @OurWatchAus · 12m
We extend our deepest sympathy to the friends of family of a 67-year-old Riverwood woman who was killed yesterday. A 24-year-old man has since been charged with her murder #VAW

1 2 9

Show this thread

Dungeons & Dragons · See more
Spookywithsander✨🧡✨🖤 @sassywithsander · 9h
DID SOMEONE SAY #TTRPG CLOAKS?!?!✨✨

Oh yeah... that was me😁✨💕

Check out my newest collection "Roll the Dice" inspired by TTRPG classes on RedsWardrobe Etsy!!

ITS TIME TO DRESS AS YOUR PC PEOPLE!!!🔥🔥

etsy.com/shop/RedsWardr...

#DnD #cosplay #cloak #CriticalRole

0:04 1.1K views

4 8 75

Roleplaying and Rollplaying (KICKSTARTER IN BIO) received a reply
Roleplaying and Rollplaying (KICKSTARTER IN BIO) @RRollp... · 10h
Shitty Character Idea:

A pact of the chain warlock whose familiar is a tiny version of their patron. #DnD

4 7 109

Frosty the BardMan (I would literally die for FCG) @FrstyTh... · 10h
Can it be a warlock who made a pact with an ancient red dragon, and now

Like Comment

Well done Siobhan.

Write a comment...

Worth sharing again!

WHAT DOESN'T KILL YOU MUTATES AND TRIES AGAIN

Donate

Like Comment Share

Write a comment...

The Earth is getting a black box recorder, and it's already started listening

Like Comment Share

Write a comment...

ONCE AGAIN we'd like to remind you that Calombaris is NOT the victim here.

news.com.au

Entertainment > Celebrity Life

George Calombaris 'cried a lot' during 'brutal' wage scandal

Former MasterChef judge George Calombaris has opened up about the "brutal" wage underpayment scandal which saw his entire company collapse.

Like Comment Share

Write a comment...

(a) Twitter. (b) Facebook. (c) Instagram.

FIGURE 2.4. Activity feed examples

TABLE 2.1. Equivalent social media interaction primitives.

OSN	POST	REPOST	REPLY	MENTION	TAG	LIKE
Twitter	tweet	retweet	reply tweet	@mention	#hashtag	favourite
Facebook	post	share	comment	mention	#hashtag	reactions
Tumblr	post	repost	comment	@mention	#tag	heart
Reddit	post	crosspost	comment	/u/mention	/subreddit	up/down vote
Parler	parley	echo	reply parley	tag	#hashtag	up/down vote
Gab	gab	repost	comment	@mention	#hashtag	up/down vote
Instagram	insta	regram	comment	@mention	#hashtag	like

will be transferable to others. Although all OSNs have APIs, over the past decade many have changed the constraints under which access to those APIs is permitted, responding to societal and commercial pressures. After the Cambridge Analytica scandal (discussed in Section 2.1.5), Facebook withdrew many elements of its API and focused more on trusted relationships with researchers rather than providing open access to all (Bruns, 2019a). Investigations revealed Facebook provided more access to personal data than it should have,[33] so this is an understandable response, but it

[33] https://www.ftc.gov/news-events/press-releases/2019/07/ftc-imposes-5-billion-penalty-sweeping-new-privacy-restrictions. Posted 2019-07-24. Accessed 2021-12-08.

hampers research of misinformation on the largest OSNs in the world (Facebook alone has almost 3 billion monthly active users, as of late 2021[34]). Furthermore, questions have been raised about the validity of the data it *has* released as part of attempts to rectify the situation, such as its Social Science One initiative, launched in 2018 (Hegelich, 2020).

Since late 2017, Twitter has broadened its offerings to researchers[35] and has continued to provide free access to some of its data. Although Twitter's user base is smaller than Facebook's, it is still widely used in political contexts, and journalists often refer to content on it as indicative of broader community sentiment, acknowledging that a majority of people are not Twitter users. Twitter's microblog model is replicated by Parler (Aliapoulios et al., 2021) and Gab (Fair and Wesslen, 2019), both of which are popular in right-wing circles (Aliapoulios et al., 2021), many users having shifted to them after being banned on Twitter for violating usage terms (or in solidarity with those banned).[36] Twitter data is also very widely used in the literature to study influence and information disorders (as discussed in Sections 2.1 and 2.2), which improves our opportunities for comparability.

2.3.1 Social network analysis

SNA facilitates exploration of social behaviours and processes by providing concepts and tools to model social relationships among actors (Borgatti et al., 2009). It is based on the premise that an actor's position in the network impacts their ability to access opportunities and resources and therefore allows us to understand social behaviours and processes in network terms (Borgatti et al., 2013). An individual's importance and role in a network can be examined with centrality analysis, and communities can be characterised with concepts such as homophily, echo chambers, and filter bubbles, including how they relate to polarisation between communities. These are discussed in more detail in the sections ahead, and their mechanics are explained, along with the graph theory that underpins them, in Section 3.2.

OSNs are often considered convenient proxies for offline social networks, because they seem to offer a wide range of data on a broad spectrum of individuals, their expressed opinions and inter-relationships. It is assumed that examining the social networks present on OSNs can inform the study of information dissemination and opinion formation, contributing to an understanding of offline community attitudes. Though such claims are prevalent in the social media literature, there are serious questions about their validity due to an absence of comprehensive social science theory and SNA techniques focused on online behaviour, the mapping between online and offline

[34] https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/. Posted 2021-11-01. Accessed 2021-11-24.

[35] https://developer.twitter.com/en/products/twitter-api/academic-research. Accessed 2021-11-24.

[36] https://www.usatoday.com/story/tech/2020/11/11/parler-mewe-gab-social-media-trump-election-facebook-twitter/6232351002/. Posted 2020-11-11. Accessed 2021-11-29.

phenomena, and the repeatability of such studies. In particular, the issue of reliable data collection is fundamental. Collection of OSN data is often prone to inaccurate boundary specifications due to sampling issues, collection methodology choices, as well as platform constraints. Chapter 4 presents a specific methodology for examining the effects of such collection issues on SNA.

The establishment of datasets in which the research community can have confidence, as well as the ability for the replication of studies, including through common benchmarks, is vital for the validation of research findings (Assenmacher et al., 2021, and Section I.1).

2.3.2 Social networks from social media data

The content for this subsection is drawn from Publication ***V****.*

Using SNA to explore social behaviours and processes from OSN data presents many challenges. Most easily accessible OSN data consists of timestamped interactions, rather than details of long-standing relationships, which form the basis of SNA theory. Additionally, although interactions on different OSNs are superficially similar, how they are implemented may subtly alter their interpretation. They offer a window into online behaviour only, and any implications for offline relations and behaviour are unclear. Beyond modelling and reasoning with the data is the question of collection—accessing the right data to construct meaningful social networks is challenging. OSNs provide a limited subset of their data through a variety of mechanisms, balancing privacy and competitive advantage with openness and transparency.

2.3.2.1 Interactions and relationships online

Given the availability, nature and structure of much OSN data, the use of network-based techniques is a natural choice for the analysis of online social behaviour, with the obvious candidate for nodes in these networks being accounts.

There is, however, an important distinction between the relatively stable, long-term relationships that are typically studied in SNA and the social connections among online actors (Wasserman and Faust, 1994; Nasim, 2016; Borgatti et al., 2009). On social media, accounts can easily fulfil the role of actors, but precisely what constitutes a relationship is unclear. An obvious candidate is the *friend* or *follower* relationship common to most OSNs, but, due to how OSNs present their specific features to users, each online community develops its own social relation culture. Therefore, such connections do not necessarily easily translate between OSNs. Is a Facebook *friend*ship really the same as a *follow* on Twitter, even if it is reciprocated? And how do each relate to offline friendships?

OSNs offer ways to establish and maintain relations with others. This is done primarily through interactions, many of which are common between OSNs, such as replying to the posts of others, mentioning others (causing the mentioning post to appear in the mentioned user's activity feed), using hashtags to reach broader communities, or sharing or reposting another's post to one's followers or friends. A sample of interactions with equivalents on different OSNs is offered in Table 2.1. (NB, we distinguish interactions from *follow*ing or *friend*ing actions, which define information flows (i.e., they tell the OSN where to send posts), which are persistent once created.) Specific interactions may be visible to different accounts, intentionally or incidentally (*cf.*, replying to one post versus using a hashtag). Exploration of these differences may lead to an understanding of the author's intent and the identity of the intended audience. Is replying to a politician's Facebook post a way to connect directly with the politician, or is it a way to engage with the rest of the community replying to the post, either by specifically engaging with dialogue or merely signalling one's presence with a comment of support or dismay? A reply could be all of these things but, in particular, it is evidence of engagement at a particular time and indicates information flow between individuals (Bagrow et al., 2019). Since most online interactions are directed towards a particular individual or group, they offer an opportunity to study the flow of information and influence. On the other hand, although friend and follower connections may indicate community membership, they obscure the currency of that connection. Through their dynamic interactions, a user who *liked* a Star Wars page ten years ago can be distinguished from one who not only *liked* it, but posted original content to it on a monthly basis. Therefore, we specifically focus on interactions rather than friend and follower relations in this study.

2.3.2.2 Social network analysis theory

Relationships between individuals in a social network may last for extended periods of time, vary in strength, and be based upon a variety of factors, not all of which are easily measurable. Because of the richness of the concept of social relationships, data collection for SNA is often a qualitative activity involving directly surveying community members for their perceptions of their direct relations and then perhaps augmenting that data with observational data such as recorded interactions (e.g., meeting attendance, emails, phone calls). Just like it is tempting to believe that delving into *Big Data* will bring quick rewards, only to discover that extracting semantic information can be remarkably challenging (Emani et al., 2015), it is tempting to believe that the richness of social relationships should be discoverable in the vast amount of interaction data provided by OSNs. This relates to contemporary work comparing 'thick' versus 'thin' data (Janetzko, 2017). Thick data is fundamental information about an individual, such as values and goals, while thin data is the digital traces of their behaviour and ideas obtained from social media. It is challenging to derive thick data from thin data. Issues to consider include:

1. Links between social media accounts may vary in type and across OSNs—it is unclear how they contribute to any particular relationship;

2. What is observed online is only a partial record of interactions in a relationship, where interactions may occur via other OSNs or online media, or entirely offline; and
3. Collection strategies and OSN constraints may also hamper the ability to obtain a complete dataset.

Although many interactions seem common across OSNs (e.g., a retweet on Twitter resembles a repost on Tumblr and a share on Facebook), nuances in how they are implemented and how data retrieved about them is modelled (beyond questions of semantics) may confound direct comparison. For example, a Twitter retweet refers directly to the original tweet, obscuring any chain of accounts through which it has passed to the retweeter (Ruths and Pfeffer, 2014). There are efforts to probabilistically regenerate such chains (Rizoiu et al., 2018; Gray et al., 2020), but, in any case, is one account sharing the post further evidence of a relationship? What if it is reciprocated once, or three times? What if the reciprocation occurs only over some interval of time? These questions require careful consideration before SNA can be applied to OSN data.

2.3.2.3 Challenges obtaining OSN data

Social media data is typically accessed via an OSN's APIs, which place constraints on how true a picture researchers can form of any relationship. Via its API an OSN can control: *how much data* is available, through rate limiting, biased or at least non-transparent sampling, and temporal constraints; *what types of data* are available, through its data model; and *how precisely data can be specified*, through its query syntax. Many OSNs offer commercial access, which provides more extensive access for a price, though use of such services in research raises questions of repeatability (Ruths and Pfeffer, 2014; Assenmacher et al., 2021). This is done to protect users' privacy but also to maintain competitive advantage. Researchers must often rely on the cost-free APIs, which present further issues. Twitter's 1% Sample API has been found to provide highly similar samples to different clients, and it is therefore unclear whether these are truly representative of Twitter traffic (Joseph et al., 2014; Paik and Lin, 2015). If the samples were truly random, then they ought to be quite distinct, with only minimal overlap. Studying social media data therefore raises questions about the "the coverage and representativeness" (p.17 González-Bailón et al., 2014) of the sample obtained and how it therefore "affects the networks of communication that can be reconstructed from the messages sampled" (p.17 González-Bailón et al., 2014).

Empirical studies have compared the inconsistencies between collecting data from search and streaming APIs using the same or different lists of hashtags. Differences have been discovered between the free streaming API and the full (commercial) "fire-hose" API (Morstatter et al., 2013). There is general agreement in the literature that the consistency of networks inferred from two streaming samples is greater when there is a high volume of tweets even when the list of hashtags is different (González-Bailón et al., 2014). More concerning is the ability to tamper with Twitter's Sample API to insert messages (Pfeffer et al., 2018), introducing unknown biases at this early stage of data collection (Tromble et al., 2017; Olteanu et al., 2019).

Assuming that *Big Data* will provide easy success without deep understanding of the data can also lead to inappropriate generalisations and conclusions (Lazer et al., 2014; Tufekci, 2014; Emani et al., 2015). This is well illustrated, for example, by the range of motivations behind retweeting behaviour including affirmation, sarcasm, disgust and disagreement (Tufekci, 2014). Similarly, in the study of collective action, there are important social interactions that occur offline (Venturini et al., 2019). Furthermore, relying solely on observable online behaviours risks overlooking passive consumers, resulting in underestimating the true extent to which social media can influence people (Falzon et al., 2017).

Big Data and its precursors in databases and data warehouses have had to address issues of data quality since the late 1960's (Scannapieco et al., 2005), both in terms of the cleanliness of the data (e.g., missing or incorrect values, poorly designed schemas, difficulties in the enforcement of consistency or other validation practices) as well as techniques to manage the distribution of values within the data. Machine learning (ML) algorithms (discussed in Section 3.4) have long benefited from techniques to manage class imbalance for classifiers (Sun et al., 2009), and careful human input is very much needed to guide ML system design. For example, Roccetti et al. (2020) describe their experiences studying faulty water meters in Italy, finding the contribution of subject matter experts invaluable in defining 'clean' data to train their ML classifiers. Others have begun to systematise how to study the effect of data quality on the performance of ML algorithms (Foidl and Felderer, 2019; Breck et al., 2019), though the phenomenon has long been known (Sessions and Valtorta, 2006).

In the case of OSN data, the quality of the data is high (as it has already been processed by the OSN platforms) and thus the further challenges are at least twofold:

- To determine the completeness of a given dataset; and
- To extract meaningful network information (i.e., semantic information) from datasets using OSN-specific schemas, which are provided by OSN-specific APIs, many of which have unique and idiomatic characteristics.

For the first challenge, it is unclear when a dataset obtained via an OSN's API is complete, because only the OSN knows the extent of its holdings and whether all query results have been provided. Repeatability requires that a query returns the same results (ignoring other effects, such as the introduction or removal of data, i.e., adding new posts or losing them when rate limits are reached); however, it is not necessary for complete results to be returned, only the same results. The primary requirement for repeatability comes from benchmarking, and recent efforts have begun to examine how to ensure repeatability for benchmarking without a requirement for complete results (Assenmacher et al., 2021). The second challenge requires careful design of networks from the data available, including an awareness of what information can be extracted from particular OSNs' data models and, therefore, how transferable methods applied to the data of one OSN are to the data of another.

OSN APIs provide data by streaming it live or through retrieval services, both of which make use of OSN-specific query syntaxes. Conceptually, therefore, there are two primary collection approaches to consider: 1) focusing on a user or users as seeds (e.g., Gruzd, 2011; Morstatter et al., 2018; Keller et al., 2017) using a snowball strategy to discover the accounts that surround them (Goodman, 1961); and 2) using keywords or filter terms, defining the community as the accounts that use those terms (e.g., Ratkiewicz et al., 2011; Ferrara, 2017; Morstatter et al., 2018; Woolley and Guilbeault, 2018; Bessi and Ferrara, 2016; Nasim et al., 2018). Focusing on seeds can reveal the flow of information within the communities around the seeds, while a keyword-based collection provides the ebb and flow of conversation related to a topic. These approaches can be combined, as exemplified by Morstatter et al. (2018) in their study of the 2017 German election: an initial keyword-based collection was conducted for eleven days to identify the most active accounts, the usernames of which were then used as keywords in a subsequent six-week collection.

Once a reasonable dataset is obtained, there may be benefit in stripping what Foidl and Felderer (2019) call 'context-dependent Data Smells'. This includes junk content introduced by automated accounts such as bots (Ferrara et al., 2016; Davis et al., 2016). The question, however, of whether to remove content from social bots, which actively pretend to be human, depends on the research question at hand; because humans are easily fooled by social bots (Cresci et al., 2017b; Nasim et al., 2018; Cresci, 2020), their contribution to discussions may still be valid (unlike, e.g., that of a sport score announcement bot). Several studies have examined how humans and bots interact, especially within political discussions (Bessi and Ferrara, 2016; Rizoiu et al., 2018; Woolley and Guilbeault, 2018), and we provide our own contribution in Chapter 7.

The question of how to reliably obtain and model social media data is relevant to all the technical chapters in this thesis (i.e., Chapters 4–7), but is explored in detail in Chapter 4.

2.4 Social Media during Crises

*The content for this subsection is based on Publications **I** and **VIII**.*

Polarisation, i.e., differences in opinion which may form the basis of argument, can occur over any two-sided issue about which people feel strongly. Some instances are relatively benign, such as between dog-loving and cat-loving communities or the 2015 dress controversy,[37] but some can result in considerable conflict, especially when it relates to politics, religion and ideology. The importance of the ongoing environmental crisis, and the role of climate change in it, is a particularly common point of argument (Williams et al., 2015). Unsurprisingly, conflict and emerging polarisation are prevalent in online discussions of related issues, as our investigation in Chapter 5 shows. The structured, linked nature of OSN data also sometimes allows researchers to observe it with relative ease, compared with identifying the boundaries of offline polarised communities, depending on the topic of the discussion and the relations between participants being examined. The danger to society of ongoing online polarisation is the focus of Part II of this thesis.

The study of the use of OSNs (including Twitter especially) during crises and times of political significance is well established (Bruns and Liang, 2012; Bruns and Burgess, 2012; Flew et al., 2014; Marozzo and Bessi, 2017; Graham et al., 2020c), and has provided recommendations to governments and social media platforms alike regarding its exploitation for timely community outreach (Saleem and Mehrotra, 2021). The social media response of the Australian Queensland State Government was praised for its use of social media to manage communication during devastating floods (Bruns and Burgess, 2012), and analyses of coordinated behaviour have revealed significant organised anti-lockdown behaviour during the COVID pandemic (Graham et al., 2020c; Magelinski and Carley, 2020; Loucaides et al., 2021) and in the lead up to the January 6 Capitol Riots in America (Scott, 2021; Ng et al., 2021). The continual presence of trolling and bot behaviour (see Sections 2.2.1 and 2.2.2) diverts attention and can confuse the public at times of political significance, whether it is to generate artificial support for policies and their proponents (Keller et al., 2017; Rizoiu et al., 2018; Woolley and Guilbeault, 2018), harass opponents (Keller et al., 2017; CREST, 2017) or just pollute existing communication channels via ‘content injection’ (Conover et al., 2011; Woolley, 2016; Nasim et al., 2018; Kušen and Strembeck, 2020). Malign actors can also foster online community-based conflict (Kumar et al., 2018; Datta and Adar, 2019; Mariconti et al., 2019), as well as polarisation (Conover et al., 2011; Garimella et al., 2018b; Morstatter et al., 2018; Villa et al., 2021).

Misinformation on social media has also been extensively studied, as we discussed in Section 2.1, with growing attention to its overall effect on society (Starbird, 2019; Carley, 2020). Many relevant current events, however, are yet to be explored in the peer-reviewed literature. Instead, researchers have turned to other methods to quickly warn of the dangers of misinformation via other channels; examples include Graham and Keller's interview with the technology magazine ZDNet (Stilgherrian, 2020) and their follow-up article on The Conversation (Graham and Keller, 2020), a publisher of “research-based news and analysis”,[38] while commissioned reports provide an opportunity to present more comprehensive yet still not peer-reviewed analyses (e.g.,

[37] https://en.wikipedia.org/wiki/The_dress. Posted 2015-02-27. Accessed 2022-01-24.
[38] https://theconversation.com/au/who-we-are

Wardle and Derakhshan, 2017; Graham et al., 2020c; Smith et al., 2020). Because social media has become such a mainstay of modern communication, misinformation on social media is often amplified on the mainstream media (MSM), or by prominent individuals, often when it aligns with their ideological outlook, which then feeds back into social media as people discuss it further.[39] This is exacerbated by persistent low levels of media literacy in the population, resulting in a limited ability to identify misinformation (Notley et al., 2021). This cycle is shown above in Figure 2.2. These cycles are useful tools in influence and information operations, enabling the exploitation of fissures in society (Benkler et al., 2018; Phillips, 2018; Starbird and Wilson, 2020; Badham, 2021).

Specifically considering fire-related misinformation, patterns of online discussion messaging and activity documented over more than the last decade (including our own investigation in Chapter 5), resurfaced in the US during Californian wildfires in mid-2020, causing armed vigilante gangs to form to counter non-existent Antifa activists who had been blamed for the fires on social media.[40] Arson has again been blamed for the 2021 fires around the Mediterranean, throughout southern Europe and in northern Africa,[41] even as the United Nations' Intergovernmental Panel on Climate Change released its sixth Assessment Report stating that humans' effect on climate is now "unequivocal" (IPCC, 2021). Furthermore, when the misinformation relates to conspiracy theories involving public health measures during a global pandemic, the risk is that adherents will turn away from other evidence-based policies, as we see with vaccine hesitancy (Ball and Maxmen, 2020), adoption of flat earth beliefs (Brazil, 2020), and other conspiratorial anti-government sentiments (The Soufan Center, 2021b).

These conditions foster an environment in which conflict is easy to spark and then entrench as people's polarised opinions are strengthened through the social bonds they form with those who agree with them. This contributes to an us-versus-them mentality, which exacerbates the polarisation. Studying the characteristics of online polarisation by examining the use of social media during times of crisis can be extremely informative. This forms the basis for much of the research presented in Part II of this thesis.

One particular social process that contributes significantly to worsening polarisation is the formation of echo chambers around opposite poles of an issue.

2.5 Echo Chambers and Polarisation

The content for this subsection is extended from Publication ***IX****.*

[39] https://www.abc.net.au/triplej/programs/hack/spread-of-arson-disinformation-us-wildfires-similar-to-australia/12666336. Posted 2020-09-15. Accessed 2022-01-07.

[40] https://www.theguardian.com/us-news/2020/sep/16/oregon-fires-armed-civilian-roadblocks-police. Posted 2020-09-16. Accessed 2022-01-07.

[41] https://edition.cnn.com/2021/08/11/world/wildfires-climate-change-arson-explainer-intl/index.html. Posted 2021-08-11. Accessed 2022-01-07.

In this section, we explore the notion of polarisation further, and also the contributing social phenomenon of echo chambers. In the previous section, we defined polarisation as differences in opinion over any two-sided issue, which people find important and that can therefore form the basis of argument. Precise definitions are often lacking in the literature, relying most often on the lay English meaning, primarily because the term is only truly meaningful in the context in which it is applied. A recent simple definition of *political* polarisation, for example, is "a divide existing between groups on either side of the political orientation spectrum" (p.186, Weber et al., 2021b), but others clarify that there are nuances in how it is manifested. In his review of research on the US liberal-conservative divide, Lelkes (2016) suggests that much of the arguments between those claiming polarisation is increasing or not is caused by how they 'operationalise' the idea of polarisation. Polarisation may be *ideological*, and can refer to a) how tightly the opinions of followers of or parties proclaiming an ideology 'align' with that of their ideology (implied to be a set of opinions on a number of issues), or b) the 'divergence' or distance between different ideologies (in terms of Likert scale-based survey responses). Perceived ideological polarisation refers to the extent people believe the community to be polarised, and is often compared with measures of actual polarisation. Finally, affective polarisation refers to people's sentiments regarding people in their own camp and those in the opposite camp. Data obtained to address these nuances is often drawn from surveys, and is analysed by examining standard deviations and bimodality of Likert scale responses. Moving beyond political polarisation to groups, "polarization is said to occur when an initial tendency of individual group members toward a given direction is enhanced following group discussion" (p.1141, Isenberg, 1986), and again, these studies often rely on survey data.

In this research, our focus is on divisions in social and political opinions, primarily studied through the lens of the interactions and behaviour of the opinion-holders, and thus a dictionary definition for 'polarisation' will suffice, though care must be taken when considering the nature of actor connections when searching for evidence of polarisation. If connections are negative (e.g., representing attacks or arguments), a highly polarised discussion may result in a tight network of interactions with only a single cluster, which would not be expected if the connections are positive (implying support or similarity) and resulting in a cluster for each like-minded community.

The Cambridge Dictionary defines 'polarisation' to be "the act of dividing something, especially something that contains different people or opinions, into two completely opposing groups."[42]

An 'echo chamber' is an environment in which the participants only share "ideologically congenial" content, which reinforces their opinions (p.1531, Barberá et al., 2015). If the ideology is somehow political in nature, this can lead to group polarisation, mentioned above, which is "when members of a deliberating group move toward a

[42] https://dictionary.cambridge.org/dictionary/english/polarization. Accessed 2021-12-08.

more extreme point in whatever direction is indicated by the members' predeliberation tendencies" (p.176, Sunstein, 2002). In this way, the presence of an echo chamber is not evidence of polarisation, as the opinions or ideology shared requires a counter-community to oppose. If there are many different opinions on an issue (e.g., favourite television shows), it is difficult for communities to oppose each other in a polarising fashion. From a network perspective, assuming positive edges, an echo chamber is a community dominated by internal connections, though some may still exist with the external community (p.22, Bruns, 2019b).

A 'filter bubble' is a specialisation of an echo chamber, referring to internet users only being exposed to information they prefer as a result of the personalisation algorithms that decide their individual results from using search and other information provision services (Pariser, 2012). These bubbles can apply to individuals or communities in the online sphere. From a network perspective, a filter bubble is an echo chamber that has cut most if not all of its external connections (p.22, Bruns, 2019b). A related concept is 'epistemic bubbles', in which a community lacks knowledge on an issue *because* it is cut off from suitable information sources (Nguyen, 2018). Nguyen (2018) argues that echo chambers can be more dangerous than epistemic bubbles, because an epistemic bubble may be broken through the introduction of new information, while members of an echo chamber do not *trust* those outside the community and thus refuse to accept new information. In fact, new information can actually strengthen already-held opinions (e.g., Bail et al., 2018).

It is important to note that, especially in social science literature, the dangers of echo chambers and filter bubbles have been argued against, as people within such communities still have access to many other sources of information (e.g., Bruns, 2019b). Although it is important to consider the audience when using such terms as 'echo chambers' and 'filter bubbles', and do so with care (Bruns, 2019c), they are still useful concepts, especially when analysing networks based on social media behaviour. The datasets that underpin those networks are defined by their collection criteria, so they represent only a partial view of all online behaviour, and (very real) polarisation observed within them is therefore still meaningful, just constrained by those criteria.

Due to the breadth of related work in this field, it is necessary to structure our review. We first elaborate on dangers caused by allowing polarisation to flourish online, then consider the difficulties of opinion formation in real-world environments where contentious issues and opinions on them abound. We finally touch on related polarisation research.

2.5.1 The broken promise of social media

As mentioned, OSNs allow people to easily form communities with shared ideas, ideals and beliefs. This notion of people connecting based on similarities is known as *homophily* in network science and sociology (Rogers and Bhowmik, 1970; McPherson et al., 2001). The open nature of the internet and social media was expected to facilitate broader engagement in society, allowing ordinary folk to communicate directly with elites (Woolley and Guilbeault, 2018), leading to what Habermas referred to as *deliberative democracy* (Habermas, 1996), where people could more easily come to a consensus on issues of interest, or gain an understanding of opposing views (as discussed by Graham and Ackland, 2017). Instead, social media users have found a plethora of ways to use the features of OSNs beyond their intended functions. The downsides of social media include the following:

- The formation of echo chambers and subsequent filter bubbles (Pariser, 2012; Barberá et al., 2015) leads to opportunities for anti-social groups to form (Massanari, 2016), incite and radicalise their members, and conduct organised raids on other online communities (Datta and Adar, 2019; Burgess and Matamoros-Fernández, 2016; Mariconti et al., 2019). This extremism can move offline also, ultimately resulting in terrorist attacks and other ideologically-motivated violence (Brooking and Singer, 2016; CREST, 2017; Waldek et al., 2020; Scott, 2021).

- Automation, big data holdings, and organised inauthentic effort can be used to influence both domestic and foreign politics (Woolley, 2016; Shorey and Howard, 2016; Ferrara, 2017; King et al., 2017; Dawson and Innes, 2019), and has been observed as far back as 2010 (Metaxas and Mustafaraj, 2012). (See Section 2.2.4.)

- Misinformation can be monetised online (even without the need for malicious intent on the part of the scammers). For example, Macedonian teenagers found they could generate revenue from GoogleAds when they created highly conservative but entirely fictional news articles in the lead up the US 2016 presidential election (Subramanian, 2017). This is an example of boyd (2017)'s concept of "[h]acking the attention economy".

- Public support can be artificially manufactured with the use of paid workers and motivated volunteers (King et al., 2017; Woolley and Guilbeault, 2018; Jamieson, 2020) or, failing that, simply faked with automated follower accounts (Aggarwal and Kumaraguru, 2015; Confessore et al., 2018).

- Finally, with relative ease and anonymity, coordinated malicious campaigns can be conducted against prominent individuals (e.g., Bot Sentinel, 2021), groups (Pacheco et al., 2020; Starbird et al., 2019; Graham et al., 2020b) or countries (Graham et al., 2020c; Strick, 2021) for ideological reasons or simply for the lulz[43] (Drenten and Gurrieri, 2018; Hine et al., 2017).

[43] 'Lulz' (also 'lolz') is defined as "laughs at someone else's or one's own expense". Collins Dictionary. Source: https://www.collinsdictionary.com/dictionary/english/lulz. Accessed 2022-01-24.

Some of these dangers have been foreseen, however, as recent revelations from whistleblowers have revealed that Facebook knew an increase in inflammatory content would be a likely consequence of its policy to weight Reactions five times more than Likes (Merrill and Oremus, 2021). In some cases the OSNs have even facilitated the spread of misinformation through favourable treatment of prominent accounts (Timberg et al., 2021).

These are all factors contributing to the increased aggression and polarisation observed not only in the online space (Garimella and Weber, 2017), but also offline as a direct result of those online events. These have real-world effects, such as vaccine hesitancy (Broniatowski et al., 2018) and coordinated anti-lockdown movements (Loucaides et al., 2021) during a global pandemic, and extremism facilitated by conspiratorial thinking (The Soufan Center, 2021b; Brazil, 2020). They are exacerbated by bias in the media (e.g., Benkler et al., 2018; Barry, 2020) and biased online amplification (Huszár et al., 2021), which contributes to radicalisation (Berger and Morgan, 2015; Badawy and Ferrara, 2018; Waldek et al., 2020) and associated violence (Samuels, 2020; Scott, 2021; Mackintosh, 2021).

2.5.2 Opinion formation in a complex opinion space

Classical opinion modelling theory tells us that, assuming people have an opinion on any matter, increased interaction will shift the population towards consensus as people find more reasons that they are similar than different (DeGroot, 1974; Baronchelli, 2018). Despite the increased opportunity for interaction provided by the internet and social media, what we observe is the contrary: an increase in polarisation on certain issues (Garimella and Weber, 2017), which then spills to sets of related issues and from the online world to the offline world. Although it might be reasonable to assume that accounts highly polarised on certain issues are unlikely to change their stance on those issues over time, they still may be receptive to alternative viewpoints on a variety of other topics. The online opinion space is complex, however, as it consists of many competing diverse, often incompatible, often orthogonal, sets of opinions. For example, in the recent COVID-19 coronavirus pandemic, healthcare and economics experts' opinions on lock-downs were consistently contested and undermined by conflicting information on social media and in the news media (Ali and Kurasawa, 2020; Tasnim et al., 2020). Studies have shown that online polarisation can even persist across conceptually unrelated issues (Häussler, 2018), and that stances on some issues seem to align (Baumann et al., 2021), which may contribute to online friendships, consolidating the social groups and reinforcing the polarisation between them. These findings highlight the need to better understand how humans respond to multiple and sometimes conflicting opinions, particularly in the online context, especially given low media literacy (Notley et al., 2021) and the effect on passive consumers who do not interact with content (Falzon et al., 2017).

2.5.3 Polarisation research

Polarisation is a broad and well studied topic (Garimella et al., 2018b; Kligler-Vilenchik et al., 2020), particularly in the context of politics and from a variety of analytical perspectives and disciplines, including:

- computer science (Conover et al., 2011; Morstatter et al., 2018; Bail et al., 2018);
- network science (DeGroot, 1974; Krackhardt and Stern, 1988; Newman, 2003; Häussler, 2018);
- sociology (Rogers and Bhowmik, 1970; McPherson et al., 2001);
- political science (Jost et al., 2003; Jost, 2017; Weber et al., 2021b); and
- general linguistic analysis (especially in the US context) (Sylwester and Purver, 2015; Li et al., 2017b; Demszky et al., 2019), as well as
- specific linguistic analysis of the use of moral terms and verbs in political arguments (Graham et al., 2013; Wang and Inbar, 2020).

Contentious issues such as immigration (Albada et al., 2021) and climate change (Williams et al., 2015) have also long provided opportunities for case studies, including our own (see Chapter 5). As social media has increasingly been used for political communications, election-related discussions have become rich sources of study of polarisation (Bessi and Ferrara, 2016; Woolley and Guilbeault, 2018; Morstatter et al., 2018; Garimella et al., 2018a). The primary fear associated with online polarisation and the echo chambers and filter bubbles that contribute to it is that, by their very nature, they may restrict the opportunities to exchange views, let alone establish common ground. They reinforce existing opinions, risking extremism and radicalisation (Barberá et al., 2015; Bail et al., 2018; Baumann et al., 2021; Jiang et al., 2021). The concern around echo chambers creating filter bubbles is not shared by all, however: Bruns (2019b) argues that although a group may form an echo chamber in an online space, the individuals have many opportunities to obtain information via other communication channels, both online and offline. Nevertheless, online polarisation appears to be increasing (Garimella and Weber, 2017) and can be costly to those attempting bipartisanship (Garimella et al., 2018a). This is in part due to the dynamics between social media, and traditional and alternative media, exacerbated by political pressures (Benkler et al., 2018; Jamieson, 2020; Badham, 2021), leaving us particularly vulnerable to the influence of misinformation and disinformation (Wardle, 2019a; Carley, 2020; Notley et al., 2021).

2.6 Detecting Coordinated Behaviour

The detection of coordinated online behaviour has emerged gradually and from a variety of application domains, but has gained momentum since 2017 as researchers began to investigate the online activity surrounding the unexpected successes of the

Brexit LEAVE campaign and the Trump presidential campaign. Prior to the advent of social media, spam campaigns were (and still are[44]) prevalent in email, and social media have simply provided more avenues for dissemination. Lee et al. (2013) used graph- and content-based methods to detect free text campaigns and machine learning to classify them as spam, promotional material, template messages, news headlines or appeals to celebrities. Cao et al. (2015) examined URL sharing behaviour to identify active campaigns, a concept extended by Giglietto et al. as Coordinated Link Sharing Behaviour (CLSB) in their studies of Italian politics (Giglietto et al., 2019; Giglietto et al., 2020a; Giglietto et al., 2020b). Campaigns are also identifiable by the hashtags they use (Conover et al., 2011; Howard and Kollanyi, 2016; Varol et al., 2017b; McKew, 2018; Graham et al., 2020b), and the networks of co-occurring hashtags used to study these campaigns are sometimes referred to as *semantic networks* (Radicioni et al., 2021). More recently, attention has turned to explicitly searching for coordination in online behaviour as evidence of campaigns beyond the presence of duplicated and highly similar content (e.g., text, URLs, hashtags, imagery). At least three categories of approaches have been employed:

Machine learning ML approaches (e.g., Davis et al., 2016; Varol et al., 2017b; Alizadeh et al., 2020), some of which rely on clustering to detect temporal synchronicity (e.g., Cao et al., 2014; Chavoshi et al., 2017; Dawson and Innes, 2019; Mazza et al., 2019);

Mathematical models These include point processes (e.g., Rizoiu et al., 2017; Sharma et al., 2021) and Bayesian methods (e.g., Yu et al., 2015); and

Networks Network-based co-activity analyses (e.g., Keller et al., 2017; Giglietto et al., 2019; Nizzoli et al., 2021; Pacheco et al., 2021).

This has provided the opportunity for researchers to focus on identifying not just individuals engaging in systematic inauthentic behaviour but also the groups behind such activities (Yu et al., 2015; Şen et al., 2016; Grimme et al., 2018; Graham et al., 2021). This is because it is the coordinated actions of influential groups, not just individuals, that have the most success with computational propaganda, whether its aim is to amplify a message, pollute a discussion channel with inauthentic or counter-narrative content or attack an opponent (Nasim et al., 2018; Mariconti et al., 2019; Giglietto et al., 2020b). Further, as discussed earlier, the societal concern with campaign detection has shifted from spam campaigns to misinformation campaigns, disinformation campaigns, influence operations and computational propaganda (Carley, 2020). This has put a greater emphasis on not just identifying the groups of accounts behind the campaigns, but also characterising their behaviour and speculating about their aims. For computer science, automatically distilling *intent* from actions and content is still

[44] Forbes claimed that, in May 2020, 320 billion spam emails were sent each day. Source: https://www.forbes.com/sites/daveywinder/2020/05/03/this-surprisingly-simple-email-trick-will-stop-spam-with-one-click/?sh=415d423b3791. Posted 2020-05-03. Accessed 2021-11-25.

TABLE 2.2. A timeline of research relating to inauthentic behaviour detection. Works associated with this thesis are denoted with an asterisk and in bold. Reference details can be found in the Bibliography.
Type key: P = Poster, C = Conference paper, J = Journal paper, T = Talk only, A = Pre-print (e.g., arXiv), R = Report, D = Demo/Data challenge

Year	Month	Author(s)	Venue	Type	Notes
2013	December	Lee et al. (2013)	TIST	J	Campaign classification (ML+NLP)
2014	November	Cao et al. (2014)	CCS	C	Temporal clustering (ML)
2015	October	Yu et al. (2015)	Trans. KDD	J	Group anomaly detection (point processes)
2016	April	Davis et al. (2016)	WWW	C	Bot detection (ML)
	July	Ferrara et al. (2016)	C.ACM	J	Survey of social bot research
	September	Cresci et al. (2016)	IEEE Int. Sys.	J	DNA-inspired bot detection
2017	April	Chavoshi et al. (2017)	WWW	C	Time-based bot detection (ML)
	April	Cresci et al. (2017b)	WWW	C	comparison of bot detectors
	May	Keller et al. (2017)	ICWSM	C	Co-activity (co-tweeting)
	May	Rizoiu et al. (2017)	ICWSM	C	Virality analysis (point processes)
	May	Varol et al. (2017a)	ICWSM	C	Social bot detection (ML)
	July	Vo et al. (2017)	ASONAM	C	Retweeter groups (network+temporal)
	July	Varol et al. (2017b)	EPJ DS	J	Hashtag campaign detection (ML)
	November	Carnein et al. (2017)	ER	C	Clustering of text streams (ML+NLP)
	November	Zannettou et al. (2017)	IMC	C	URL campaign detection (Hawkes processes)
	December	Grimme et al. (2017)	Big Data	J	Social bot/botnet design
2018	April	Nasim et al. (2018)	WWW	C	Polluter detection (Network+temporal)
	May	Rizoiu et al. (2018)	ICWSM	C	Retweet campaign analysis (point processes)
	July	Grimme et al. (2018)	SCSM/HCI	C	Social bot/botnet design
	July	Beskow and Carley (2018b)	SBP-Brims	D	Bot detection (ML)
	August	Beskow and Carley (2018a)	ASONAM	C	Bot detection (ML+Network)
2019	January	Gupta et al. (2019)	COMAD/CODS	C	Retweeter group detection (ML)
	January	Yang et al. (2019)	HB+ET	J	Bayesian automation prediction (CAP)
	May	Dawson and Innes (2019)	Pol. Quart.	J	Troll behaviour analysis
	May	Zannettou et al. (2019)	WWW	C	Troll behaviour analsyis (Point processes)
	June	Mazza et al. (2019)	WebSci	C	Retweeter groups (ML+DL+temporal)
	September	Giglietto et al. (2019)	socArxiv	A	URL campaign detection (network)
	October	Keller et al. (2019)	Pol. Comm.	J	Co-activity strategies (co-tweeting)
	November	**Weber (2019)***	**ASNAC**	**P**	**Groups by co-activity (network+temporal)**
2020	March	Giglietto et al. (2020b)	IC&S	J	URL campaign detection (network)
	March	Fazil and Abulaish (2020)	JIFS	J	Campaign strategy evaluation (network)
	May	Assenmacher et al. (2020)	FLAIRS	C	Tweet clustering campaign detection (ML+NLP)
	May	Pacheco et al. (2020)	WWW	C	Groups by behavioural traces (network)
	June	Giglietto et al. (2020a)	ICSM+S	C	URL campaign detection (network)
	June	Graham (2020)	QUT	T	Groups by co-activity (network)
	July	Alizadeh et al. (2020)	Science	J	Troll strategy detection (ML+temporal)
	September	Cresci (2020)	C.ACM	J	Survey of social bot research
	November	Graham et al. (2020a)	ASNAC	T	Groups by co-activity (network)
	November	Vargas et al. (2020)	SIGSAC	C	Coordination detection (network+ML)
	November	Magelinski and Carley (2020)	IDeaS	T	Groups by co-activity (network+temporal)
	December	**Weber et al. (2020b)***	**ASNAC**	**T**	**Groups by co-activity (network+temporal)**
	December	**Publication II***	**ASONAM**	**C**	**Groups by co-activity (network+temporal)**
2021	March	Ram et al. (2021)	WSDM	C	Bot detection (point processes)
	May	Broniatowski (2021)	IDDP Report	R	URL campaign detection (point processes)
	May	Magelinski et al. (2021)	MAISoN	T	Groups by co-activity (network+temporal)
	May	Schliebs et al. (2021)	OII, Oxford	R	Groups by co-text in RTs & replies (network)
	June	Nizzoli et al. (2021)	ICWSM	C	Account similarity networks
	June	Pacheco et al. (2021)	ICWSM	C	Groups by behavioural traces (network)
	July	Yu (2021)	ICICT	C	Media URL campaign detection (network)
	August	Sharma et al. (2021)	SIGKDD	C	Campaign detection (point processes)
	September	Graham et al. (2021)	ECREA	T	URL campaign detection (point processes)
	September	Ng et al. (2021)	SBR-Brims	T	Groups by URLs (network)
	October	**Publication VII***	**SNAM**	**J**	**Groups by co-activity (network+temporal)**
	October	Zhang et al. (2021)	NeurIPS	C	Coordination detection (point processes+Bayes)
	November	Cresci et al. (2021)	arXiv	A	Group detection optimised via GANs (ML)

very much an open research area, but the more tractable challenges of group discovery and strategy identification has received considerable scholastic attention in recent years. This is evident in the number of publications since 2017 shown in Table 2.2, which provides a historical timeline of recent works contributing to the discovery of inauthentic online behaviour and coordinated groups, and other relevant research, including our own arising from this PhD project.

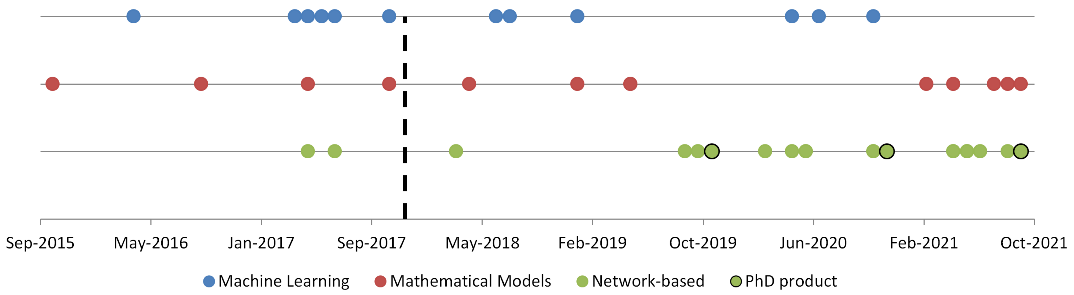


FIGURE 2.5. A timeline of selected relevant papers, categorised by the approach taken to analyse inauthentic or coordinated behaviour, from Yu et al. (2015) onwards. This PhD project started on 2017-12-06, marked by the dashed vertical line. PhD products are papers that were published as part of this PhD research.

Below, we offer a discussion of contributions in the literature using each of the three approaches mentioned above. First, however, we present a timeline of a selection of categorised papers as exemplars in Figure 2.5. We can see that machine learning approaches have been applied often but sporadically, while mathematical models have been consistently popular to the greater extent, but particularly popular in 2021. Network-based methods have gained significant popularity in recent years, and the approaches described in the papers reviewed share many similarities, with the primary differences being the method to extract meaningful communities from the networks and the validation methods to confirm findings.

2.6.1 Machine learning approaches

The primary benefit of ML approaches is that they are data-driven, i.e., they 'learn' from examples of the data, allowing the researcher to use as many 'features' of the data they have as they choose. The majority of uses of ML in the selected papers are supervised (i.e., trained with labelled data) though some use unsupervised clustering methods (e.g., based on textual similarity or synchronicity). The supervised methods are for classification, judging whether:

- accounts are bots or genuine humans (Davis et al., 2016; Chavoshi et al., 2017; Varol et al., 2017a; Beskow and Carley, 2018a; Beskow and Carley, 2018b; Yang et al., 2019);
- groups of accounts are coordinating, either as the main focus of the detector (Vargas et al., 2020) or as a confirmation measure (see Chapter 7 and Publication **VII**);

- accounts are members of malicious retweeter botnets, specifically (Gupta et al., 2019; Mazza et al., 2019);
- campaigns were of particular types (Lee et al., 2013; Varol et al., 2017b); or
- particular campaign strategies were used (Alizadeh et al., 2020).

Some classifiers are built on very few features, such as Beskow and Carley (2018b)'s Bot-hunter, which uses a selection of 6 profile, 3 network, 6 content, and 2 timing features. Vargas et al. (2020)'s group classifier relies on seven network features drawn from six different types of co-activity networks for a total of 42 features. In contrast, Botometer (Davis et al., 2016) and its revisions (Varol et al., 2017a; Yang et al., 2019) uses more than a thousand features spanning six similar categories (network, user, friend, temporal, content and sentiment features) to create an ensemble classifier. Many of the features Botometer uses relate to distributions of values, e.g., minimum, maximum, mean, median, mode and standard deviation of hashtag/mention/URL uses per account.

Temporal information is the focus of Chavoshi et al. (2017)'s DeBot classifier, which uses distributions of interarrival times (i.e., the times between posts) to identify regular posting behaviour as well as *burstiness*, a feature also used for fake review detection (Li et al., 2017a) and studying online virality (Deusser et al., 2018; Jansen, 2019). Mazza et al. (2019) also use temporal information to successfully identify retweeting botnets using unsupervised feature extraction and clustering, making use of deep learning (DL) rather than traditional ML algorithms such as SVM or Random Forest.

Other unsupervised ML techniques are also applied to clustering social media posts and accounts according to similarities in the text they use (Lee et al., 2013; Carnein et al., 2017; Assenmacher et al., 2020; Schliebs et al., 2021). A very recent development has been the use of adversarial ML to train fake news and social bot classifiers in the same way as it has been applied to image classification and voice recognition (Cresci et al., 2021).

2.6.2 Mathematical modelling approaches

The label chosen for this category of approaches is necessarily broad, but the majority of these approaches rely on analysing distributions of values drawn from the social media data, often with regard to interarrival times. 'Point processes' are often used to model these, which define the probability of the arrival of a new event in a stream of timestamped events. One commonly used for social media data is the Hawkes process (Hawkes, 1971), which models 'self-exciting' processes, processes where the arrival of an event affects the probability of another arriving within a given time period (Laub et al., 2015). An extension to model the expected volume of flow-on events rather than their timing, known as the Hawkes Intensity Process (HIP), was first used to predict the popularity of YouTube videos based on their views, shares and tweets

(Rizoiu et al., 2017). Since then, it has been used to examine retweet cascades during election debates (Rizoiu et al., 2018), identify social bots (Ram et al., 2021) and study disinformation campaigns (Graham et al., 2021). Broniatowski (2021) used a Poisson point process with the interarrival times of tweets mentioning the same URL to study "near simultaneous" (within 25 seconds) link sharing behaviour. He identified groups of coordinated actors by connecting those engaging in such link sharing, prioritising those who did it most frequently for further investigation in terms of their identity and their content. Dawson and Innes (2019) may have used point processes in their 'synchronicity analysis' of RU-IRA accounts engaging in "simultaneous patterns of messaging" (p.251, Dawson and Innes, 2019) to find sockpuppet accounts, but their methods were not described in sufficient detail. Sharma et al. (2021) use a sophisticated model that relies on account activity and hidden group behaviours, and incorporates temporal point processes to distinguish anomalous levels of within-group activity from the rest of the population. Although point processes have been used the longest in our selection, they have been used consistently since, indicating that they remain useful. Future approaches could be enhanced by combining point processes with other methods.

Other approaches use Bayesian probability, such as Yu et al. (2015)'s *Group Latent Anomaly Detection* (GLAD) model, which identifies anomalous group behaviour in networks built from social media data. Yu et al. (2015)'s dynamic version of GLAD detects anomalies in series of network snapshots, which is similar to the notion of 'graph streams' (McGregor, 2014) in which Eswaran et al. (2018)'s Spotlight can detect anomalies. Spotlight is designed to detect the appearance or disappearance of significant cliques between snapshots. Bayesian reasoning was also introduced to Botometer by Yang et al. (2019) to calculate a Complete Automation Probability (CAP) statistic for Twitter accounts.

Finally, Cresci et al. (2016) proposed a method for detecting 'spambots' that treated account behaviour as a DNA-like sequence, and then applying sequence mining techniques (ones that rely on unsupervised clustering of subsequences can regarded as machine learning). In one experiment, each of an account's tweets is encoded as a letter corresponding to tweet type (simple tweet, reply or retweet), and then, in a second experiment, to its content (URLs, hashtags, mentions, media, combination or plain). In each case, each account is then represented as a unique sequence of as many characters as tweets were collected, with accounts sharing long common subsequences assumed to be automated. In two labelled datasets of spambots, one political and one commercial, the technique successfully distinguished humans from bots by tweet type and in one case by tweet content, and also matched and at times outperformed other contemporary techniques. The approach was extended to identify botnets of spambots with a technique known as 'Social Fingerprinting' (Cresci et al., 2017a), but neither the original nor extended approach exploited temporal information, such as the interarrival times between tweets.

2.6.3 Network-based approaches

Where earlier research had assigned accounts to loose groups based on the URLs or text they shared (Lee et al., 2013; Cao et al., 2014; Giglietto et al., 2019), or tight groups based on their synchronised retweeting behaviour (Vo et al., 2017; Keller et al., 2017; Gupta et al., 2019; Mazza et al., 2019), a number of recent network-based efforts have attempted to generalise the 'reasons' for associating accounts into groups while also experimenting with the temporal aspects of those reasons: this is done by creating *coordination networks* of accounts engaging in *co-activities* in a constrained timeframe, e.g., sharing the same URL or retweeting the same tweet within 10 seconds. A variety of Twitter-based co-activities have been previously identified (Ratkiewicz et al., 2011; Keller et al., 2019; Vargas et al., 2020), such as

- *co-tweet*: tweeting the same or similar text;
- *co-retweet*: retweeting the same tweet;
- *co-url*: using the same URL;
- *co-mention*: mentioning the same account; and
- *co-hashtag*: using the same hashtag.

Some approaches (Fazil and Abulaish, 2020; Nizzoli et al., 2021; Pacheco et al., 2021) are even more flexible about the association 'reason', relying on the application domain to help define what makes a pair of accounts 'similar', rather than restricting it a co-activity, and thus create "user similarity networks" (p.444, Nizzoli et al., 2021). The edge weights on such networks represent the frequency of instances of coordination (or similarity) found in the data. There are several network-based efforts along these lines (Graham et al., 2020a; Fazil and Abulaish, 2020; Nizzoli et al., 2021; Magelinski et al., 2021; Pacheco et al., 2021) including our own (Weber, 2019, and Publications **II** and **VII**),while Vargas et al. (2020) introduced a ML classifier to identify coordinating groups modelled as networks. All of these approaches can constrain the time window, to ensure the co-activities are temporally similar, but this does assume that any 'coordination' will be contemporaneous, and thus misses the more operation-level, long-term coordination required for uncovering information operations such as Secondary Infektion (Nimmo et al., 2020), which was conducted over many years with a small number of highly curated personas. Such activities are highly unlikely to be detectable in any automated way in any case. The constrained time window is typically applied in a conceptual sliding time window, though the detail of how the window slides and the degree of overlap are not explored in detail in most of the work to date – only our work goes into any considerable detail about the mechanics of the sliding window approach (Publication **VII** and Chapter 7).

The general process used by all these approaches is to ingest social media posts, extract relevant timestamped elements, and then associate accounts according to common features in those elements. A temporal constraint can be applied to most of these, e.g., finding accounts that retweeted the same tweet within thirty seconds, with the use of a sliding window concept. Within each window (which may overlap or be adjacent), the timestamped elements are analysed to connect accounts which fit the association criteria, progressively adding to a 'coordination network', which can be mined for the most highly 'coordinating' accounts to filter out coincidental associations (e.g., people coincidentally use the same hashtag all the time – this is the basis for 'Currently trending' lists provided by many OSNs). The methods for this 'community extraction' can vary, depending on the context, and even the time constraint may not always be needed. A good counter-example is finding accounts that switch their names – accounts are associated when they shared a screen handle (publicly visible account name, independent of the underlying account ID), meaning they have swapped IDs at some point in the corpus. In this case, the community extraction method need only be to find connected components.

Bearing the general process in mind, we now summarise notable approaches.

- Graham et al. (2020a)'s approach finds a variety of co-activities, including co-retweeting, co-replying, co-mentioning, co-hashtag use, co-URL use, co-text (posting identical text) and co-simtext (posting similar text, which requires a tuning parameter). The sliding window is applied through a database join query, overlapping to whatever degree is necessary to uniquely identify all possible co-activities, which increases the computational load and also forces the analysis to be conducted post-collection. To remove coincidental associations, an edge weight threshold is applied, culling edges with a weight of only 1. Graham et al. (2020a) endorse the findings of Keller et al. (2017), emphasising that coordinating groups are often directed by one of their members, what is referred to as the *principal-agent* model. This model was recently observed by Graham et al. in a study, which successfully identified not just coordinated amplification strategies in use but also apparent promotion activity schedules, according to which articles from hyper-partisan information sources were disseminated at stipulated times (Graham et al., 2021). These strategies made use of Twitter accounts as 'agents' guided and encouraged by a principal actor (Keller et al., 2019) according to the dissemination schedule to ensure the content's longevity in the social memory and its virality.

- Nizzoli et al. (2021) uses the concept of a 'user similarity network' along with a multi-resolution community extraction algorithm to identify coordinating communities within, meaning that groups engaging in different degrees of coordination can be revealed progressively. They test their approach with seed-based networks, i.e., they begin their data collection with known accounts (e.g., *super-producer*[45] or *superspreader*[46] accounts), and then snowball outwards by finding 'similar' accounts linked to the seeds. In this sense, the temporal constraint

[45] Superproducer: an account that creates many new posts.

[46] Superspreader: an account that re-shares many posts (e.g., via retweets).

and sliding window is only applied if it is part of the semantics of the chosen 'similarity' measure.

- Fazil and Abulaish (2020)'s approach relies on vectors of statistically-derived similarity features (e.g., account age, posting rate, follow and friend rates) to associate accounts in another 'user similarity network', and then applies markov clustering to extract 'campaigns' (i.e., clusters of coordinating accounts). Their approach is validated through the use of a social bot network they designed and ran, as a way to explore the effectiveness of their social bots infiltration strategies. Temporal information is considered with regard to account age and rate-based features, and thus this approach neglects to employ the sliding window concept.

- Magelinski et al. (2021) highlight the importance of distinguishing between genuine grassroots activities which appear coordinated, like activism and fandoms, and inauthentic ones, such as cyberhate campaigns (Bot Sentinel, 2021) or organised raids (Mariconti et al., 2019). They also discuss ethical issues associated with publishing detection techniques, thereby making them available to authoritarian regimes or police states. They create multi-view coordination networks, in which each layer corresponds to a particular type of co-activity, and then rely on a density-based clustering technique to identify coordinating communities. Magelinski et al. (2021) use centrality measures as a validation method, particularly focusing on degree, which they claim can reveal the principal actor of a coordinating group. Searching for co-activity types together, they suggest, may help focus the search for inauthentic behaviour (by searching for posts with, e.g., specific hashtag and URL pairings). Using this approach, they identified several 'template campaigns' (using similar text to promote their content) that use the same mention/URL and hashtag/URL combinations to pressure the Mexican government over environmental policies, and could distinguish between those focused around individuals (i.e., organised) and those not (i.e., grassroots movements). They also showed the value in multi-view networks built from both co-mentions and co-hashtags (i.e., accounts are linked if they are paired according to either co-activity), examining the activities of a 'Reopen America' advocate account, and contrasted the account's network to a coordinated official news media dissemination group.

- Pacheco et al. (2021)'s approach elegantly uses the concept of a digital 'behavioural trace' as the reason to associate two accounts. Initially, an account/trace bipartite network is formed, where edges only link accounts to traces (e.g., indicating the use of a URL or hashtag). The traces are converted to 'features', and in this step the edge weights can be scaled to represent the strength of the 'evidence' that the trace represented. An account/account network is projected from this second bipartite network, combining the edge weights to create the account/account edge weight. This account/account network is conceptually

> much the same as the coordination networks above, and clusters of accounts can be detected in it using a variety of methods, the choice of which is left as an exercise for the reader. The authors demonstrate with several case studies that, depending on the trace chosen, simply identifying connected components is an effective clustering method. The case studies provide evidence of the flexibility of the trace approach: the traces used include name (i.e., Twitter screen handle) switching, image co-sharing (requiring image analysis for comparison, not just URL comparison), sharing the same *sequences* of hashtags, retweeting the same tweet, and short interarrival times as evidence of synchronised activities. In several cases, *tf-idf* (term frequency–inverse document frequency) is used to scale the bipartite network edges to help reduce the effect of highly popular traces (e.g., original tweets in the co-retweet case study). An early variation on the co-retweet study was used in a study of amplification of anti-White Helmet tweets (Pacheco et al., 2020), in which case the temporal constraint was part of the bipartite edge: the trace nodes were original tweets retweeted within ten seconds of their creation, rather than constraining pairs of co-retweets within the desired time window. This approach is clearly aimed at detecting automated amplification rather than more organic, human-driven efforts, which may operate over longer timeframes.

None of these approaches, as described in their publications, is immediately suitable to a near real-time processing pipeline execution model. None provide significant detail of how their sliding window or compare different community extraction methods. Because of this, and because of the arbitrary overlap in sliding windows, they are necessarily bound to process their datasets once the collection activity is complete. None include the ability to associate accounts active in the same conversation, such as via Twitter's 'conversation ID' API feature,[47] which is valuable to the analysis of online conversation patterns (e.g., Gonzalez-Bailon et al., 2010; Tamine et al., 2016; Beskow and Carley, 2018a; Bagavathi et al., 2019; Ackland, 2020). Only Magelinski et al. (2021) discusses the benefits of explicitly combining co-activities, though it could be assumed possible with the other approaches. A positive point for these approaches is that they are very flexible. Nizzoli et al. (2021) and Pacheco et al. (2021)'s concept of account 'similarity' provides significant scope to associate accounts on a basis other than co-activities or highly similar content. For example, taking inspiration from Pacheco et al. (2021), Yu (2021) recently exploited this flexibility to associate accounts not on the basis of their URLs but by comparing the media content (i.e., imagery, video) to which the URLs refer.

2.6.4 Validation

Validation of campaign and coordination detection tools in the papers surveyed is typically done via manual examination of results. An outlier in this regard is Fazil

[47] https://developer.twitter.com/en/docs/twitter-api/conversation-id. Accessed 2021-12-10.

and Abulaish (2020)'s campaign detection system, which relies on the social bots they designed and ran to develop a ground truth dataset. Manual inspection best suits those in possession of interesting datasets and relevant domain knowledge of contemporaneous offline events, a situation that favours the more established research groups, given the difficulties in publishing social media datasets (which we discuss in Section I.1 and Chapter 4). Furthermore, campaigns are amorphous entities and defining which activity is part of a campaign or not is not always clear, therefore traditional evaluation measures for classifiers are not always suitable (e.g., confusion matrices). Traditional classification measures are used extensively for the evaluation and comparison of bot detection approaches (e.g., Feng et al., 2021), as well as for classification of *types* of campaigns (e.g., Lee et al., 2013), but these still require something to act as ground truth, whether it is manual inspection and labelling (e.g., Lee et al., 2013) or external knowledge (e.g., court documents identifying social media actors, Keller et al., 2017; Keller et al., 2019).

When ethics protocols restrict the publication of individual identifiers, however, evaluation that highlight specific groups of named accounts and associated messaging are not available. One must therefore focus on other methods to build confidence in results. This can be achieved by developing a variety of statistically-generated methods to validate proposed approaches. These can also form the basis for comparison between approaches. This is the approach we have taken in Chapter 7, and the basis for many of our validation methods are discussed in Chapter 3.

One primary advantage of manual inspection of the results of campaign and coordination detection is that is easier for a human observer to distinguish between genuine (incidentally) coordinated activities, such as those of fandoms and activists, and inauthentic coordination. At least two approaches have focused on this distinction:

1. The multi-resolution coordinating community extraction method used by Nizzoli et al. (2021) enables the researcher to vary how deeply into the coordination network they delve. The stronger the threshold for coordinated behaviour, the more likely it is to be deliberate, often involving automation.

2. Content analysis, in the form of the classification of propaganda linguistic techniques, has been recently combined with coordination detection to specifically identify malicious propaganda campaigns on social media as distinct from activist efforts during an election (Hristakieva et al., 2021).

Analysing a corpus of social media posts using a variety of co-activity types as coordination criteria can produce a similar effect. Manual inspection, however, is still beneficial to interpret and confirm the results, as demonstrated by Magelinski et al. (2021) and we incorporate this approach in Chapter 7. A broader study of several established information campaigns could be used to identify which co-activities and combinations thereof are used in different campaign styles, similar to Alizadeh et al. (2020)'s identification of distinct campaign strategies used by Russian, Chinese and

Venezuelan troll teams. This could lead to opportunities for more automated detection, because it could aid in prioritising which co-activity combinations are most frequently used, by whom and for what purposes.

2.6.5 Active research communities

A last note is that a co-author network[48] of the selected works in Table 2.2 provides insight into the groups active in this field. In Figure 2.6 we present the largest groups of co-authors based on the non-survey papers in Table 2.2. We can see that there is a very large and active community from Indiana University, home to the Observatory on Social Media (OSoMe) project, centred around Flammini, Ferrara and Menczer, which also has links through other publications to the community focused around Cresci of the Institute of Informatics and Telematics in the Italian National Resarch Council (IIT-CNR) – Ferrara and Cresci co-edited a recent special issue of the Journal of Computational Social Science regarding malicious online behaviour during the COVID-19 pandemic (Ferrara et al., 2020). Another community is centred around Grimme of the University of Münster, the work of which focuses on the design of hybrid social bots and botnets and the effectiveness of detection systems in finding them, while Carley's community in the Center for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University focuses on inauthentic behaviour detection, particularly with counter-terrorism and cyber-security applications. Zannettou's mostly Europe-based community focuses on information disorders, online conspiracies and malicious online behaviour. Finally, the community to which this work belongs is bridged to the Rizoiu/Graham community by Mitchell and represents a primarily Australian contingent. The cluster around myself, Weber, has focused on online polarisation and coordinated online behaviour (as documented in this thesis), while the work of Graham and Rizoiu has looked mostly at bot detection, information dissemination and online influence.

2.7 An Observation

There is a parallel to be drawn between the limited selection of co-authors in Figure 2.6, representative of the limited set of works referred to in this section (listed in Table 2.2, and the other works cited throughout this chapter, and the study of social dynamics through the lens of social media. Human users of social media use other means (i.e., outside of social media) to communicate and interact, and their activities on social media (based on the data we can obtain about them from the OSN APIs) are only indicative of but a subset of their overall engagement in society (online and offline), not to mention the lack of data we have for passive consumers who leave no (easily accessible) digital trace. Thus, there is only so much we can tell about the effects, say, of a disinformation campaign from social media data, but valuable insights can still be gleaned. Similarly, despite the length and breadth of this review chapter,

[48] A co-author network is itself a form of coordination network.

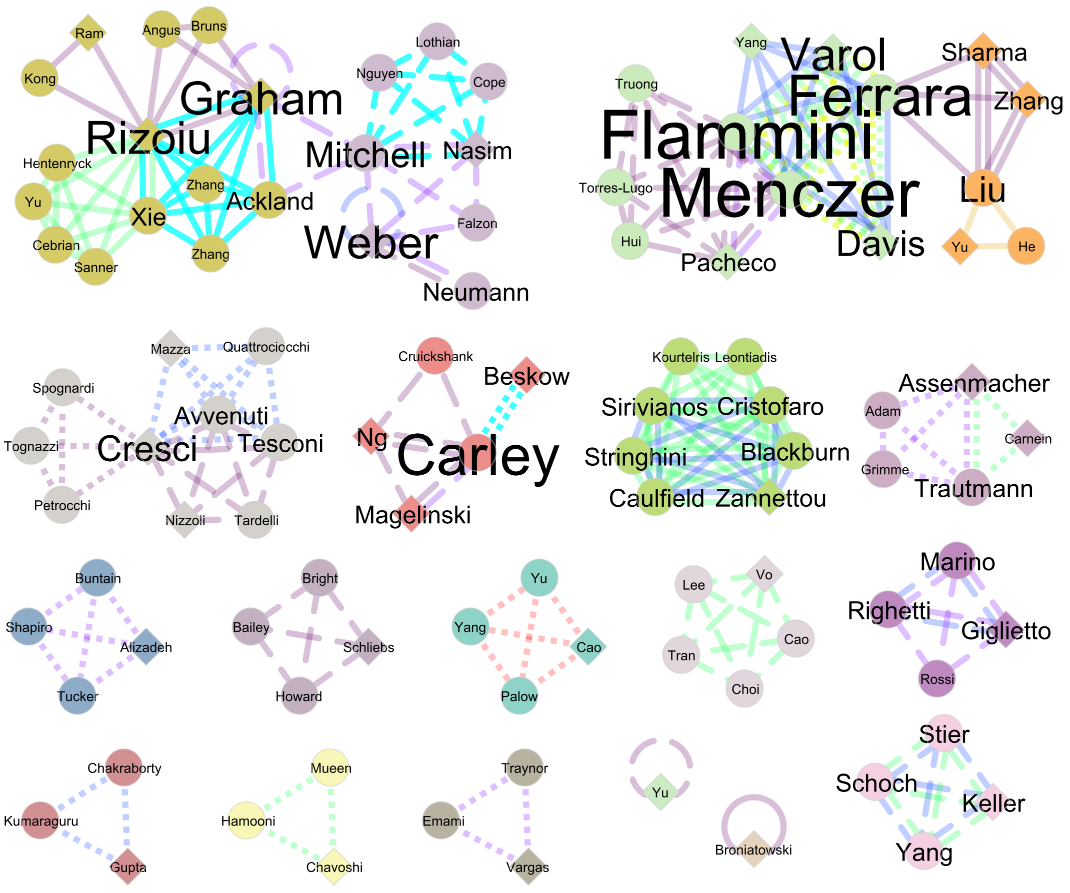


FIGURE 2.6. Communities of co-authors researching inauthentic online behaviour. Nodes represent authors, with first authors drawn as diamonds. Nodes labels are authors' surnames, which are sized according to the number of papers in the sample they have co-authored. Nodes are coloured according to Louvain cluster (Blondel et al., 2008). Each edge indicates the joined authors worked on a paper, and the edge style indicates the approach taken: ML are dotted, network-based methods are dashed, while mathematical model-based approaches are complete lines. The line colour represents the year the paper is published, ranging from 2014 in red to 2017 in green to 2021 in deep purple.

there are many more relevant works for the reader to find, but the reader will now have a strong grounding in some of the complexities involved in this research, not just in the sense of technical details, but also in the degree to which this research spans many different research disciplines and application domains.

Chapter 3

Methods

We adopt a network-based approach to the majority of analysis in this PhD project, but also use a variety of other analytic methods depending on the demands of the research question at hand. In several cases, we adopt a non-network approach as a method of validation, especially for Chapter 7. First, however, we provide a description of the data model to be used, as although our aim is to generalise our analytic efforts by relying on features common to multiple OSNs, in practice we begin with representative data obtained from Twitter. We then introduce network science, specifically for its specialisation for studying social networks, and methods for comparing and characterising a variety of types of data and group content and behaviour. We then conclude by briefly touching on machine learning techniques and bot analysis.

3.1 Dataset Statistics

The datasets that we examine in this thesis are based on data from Twitter, which shares many features with other major OSNs. We introduce the relevant elements of Twitter's data model in this subsection. Before considering more sophisticated analysis, we can gain significant insights into such a dataset with descriptive statistics, especially when we can group the statistics by types of user (e.g., through membership of a community or group). These are used particularly when considering the reliability of social media data (Chapter 4) and when characterising the differences between polarised groups (Part II).

3.1.1 Twitter data model

As discussed in Section 2.3.2, Twitter data is available via its APIs, which include both streaming access (i.e., filtering of live tweets as they are posted) and search or look-up of existing data holdings. The access level dictates the rate limits and search periods available to a requestor. The base level provides up to 500k tweets per month and a search facility that extends back seven days from the time of the query.[1] Though the primary data type is the tweets themselves, it is possible to obtain other data, such

[1] https://developer.twitter.com/en/docs/twitter-api/getting-started/about-twitter-api#v2-access-level. Accessed 2021-12-13.

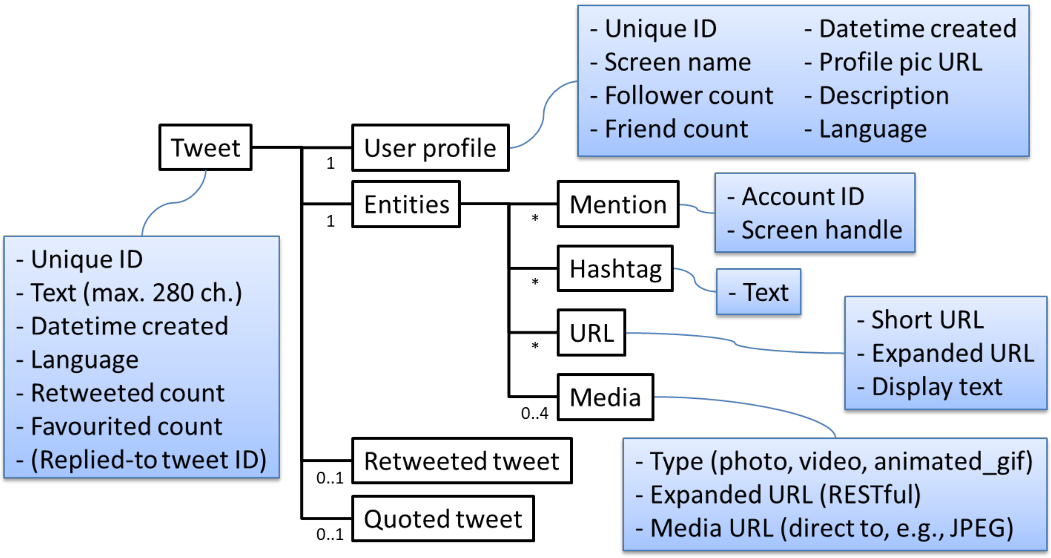


FIGURE 3.1. A conceptual data model of a tweet, in which the black and white boxes are key-value maps, and the blue boxes list map keys. Values with a cardinality of '0..1' or with a name in parentheses are optional.

as user profiles, including in bulk; the retrieval service permits bulk requests for such data, provided the entity IDs are known.

Each tweet is a microblog post of up to 280 characters, not including a trailing URL (i.e., a URL at the end of the tweet text). Its structure is a nested key-value map, where values may be arrays, maps, or primitives (i.e., strings, numbers, or booleans). A selection of the notable keys and values is shown in a conceptual representation of the data model in Figure 3.1. Each tweet has a unique ID and incorporates a snapshot of the profile of the user that posted it at the time they posted it, even if it is a retweet. The user profile has its own unique ID. Twitter populates a number of meta-data fields based on the tweet text: information regarding hashtags, URLs, and mentions of other accounts by screen name are included in the `entities` substructure. Twitter also populates a language field (`lang`) with a calculated prediction of the tweet's language. When tweets include only a URL, a mention, or a hashtag (which is not infrequent), this field is often left as "und", meaning 'undefined'.

In contrast, the `lang` field in the user profile is manually set by the user. The user profile also contains a user-specified free text description and location, and people are often whimsical in how they populate these fields, confounding reliable geolocation analysis. To study inauthentic follower behaviour, Dawson and Innes (2019) examined an account's profile snapshots over a period of time, attending to follower count changes, identifying both the sudden appearance of highly likely fake followers as well as *follower fishing* behaviour.

If a tweet is a retweet or a quote tweet, it will retain a copy of the *original* retweeted or quoted tweet in the correspondingly named field (`retweeted_status` and `quoted_status`, respectively). If both the `quoted_status` and the `retweeted_status` are populated, then the tweet is a retweet of a quoted tweet (rather than a quote of a retweet). The tweet's `retweeted_count` and `favorite_count` fields are updated whenever it is retweeted or favourited, respectively, so if a tweet is obtained via the streaming API, its `retweeted_count` and `favorite_count` will always be zero.

3.1.2 Example statistics

Based on the model described above, useful descriptive statistics for interpreting a dataset of tweets include frequencies and maximums, though examination of distributions could also be of value, especially for very large (multi-million strong) datasets. Such values are often used in machine learning applications (e.g., Davis et al., 2016; Alizadeh et al., 2020). The frequency statistics relate to the absolute count of the following features:

- *Tweets*: The number of tweets in the corpus.
- *Accounts*: The number of unique accounts that posted tweets in the corpus (i.e., does not include those that were only mentioned or whose tweets were retweeted).
- *Retweets*: The number of tweets which were native retweets, i.e., created by clicking the 'retweet' button on the Twitter user interface, rather than manually typing in "RT @original_author: *original text*", which is another valid, though time consuming, way to retweet. Both include an implicit mention of the account being retweeted, but the second will not populate the `retweeted_status` field.
- *Quotes*: The number of tweets which were quote tweets (non-native retweets, or retweets with comments).
- *Replies*: The number of tweets which were replies, including replies to tweets outside of the corpus.
- *URLs*: The number of tweets using URLs, the number of unique URLs used and the number of URL uses.
- *Hashtags*: The number of tweets using hashtags, the number of unique hashtags used and the number of hashtag uses.
- *Mentions*: The number of tweets containing mentions of other accounts, the number of unique mentioned accounts, and the number of mentions overall.

The remainder relate to the highest values of the following features:

- *Tweeting account*: The most prolific account and the number of tweets they posted.
- *Mentioned account*: The most mentioned account and the number of times they were mentioned.
- *Retweeted tweet*: The most retweeted tweet and how often it was retweeted.

- *Replied-to tweet*: The tweet with the most direct replies, and the number of those replies.
- *Used hashtags*: The first and second most used hashtags, and the number of times they were used.
- *URLs*: The most used URL, and the number of times it was used.

If groupings of accounts are known at this point, their statistics can be grouped and compared, providing insight into how the groups differ in behaviour and appearance (e.g., through examining number of tweets posts and distribution of follower counts, respectively).

3.2 Graph Theory and Social Network Analysis

Graph theory provides us with the tools to study the interactions and relations between entities by representing the entities as nodes in a graph and the connections as the edges between them (Brandes and Erlebach, 2005; Newman, 2010). Though widely used in a variety of disciplines, from physics and chemistry to economics, it is particularly well suited to social network analysis (SNA), facilitating the study of social relational structures and processes by modelling individuals as nodes and their relationships as edges (Borgatti et al., 2009). Graph theory provides a range of analytics that can inform us about

- the importance of individuals and the roles they play, e.g., based on different centrality scores;
- the similarities and position-based associations between individuals (e.g., same age or gender, and peripheral or central);
- the flow of information and influence throughout the network; and
- the presence and nature of any communities or clusters (sets of nodes that are more closely associated than the rest of the network).

In this thesis, we employ graph theory to model both social and content networks to study the communication patterns and the relationships between discussion themes (respectively) that are found in social media datasets.

Next, we consider the theory underpinning graphs and SNA, followed by a discussion of how network science is put into practice.

3.2.1 Theory and concepts

Though the majority of this subsection will focus on graph theory and network science, it is appropriate to introduce some key social science concepts to help keep in mind how the theory will be applied in practice. Borgatti et al. (2009) neatly summarises the main concepts of social science modelled by SNA:

- An *actor* in an individual in the community under study.
- *Ties* represent the information that links a pair (a *dyad*) of actors. Ties may be directional and can be broken down into four types: similarities (e.g., geographic, membership), social relations (e.g., kinship, cognitive), interactions (e.g., emailed), and flow (e.g., beliefs, resources, Borgatti et al., 2009).
- The *relation* defines the type of link between a pair of actors, i.e., the nature of their relationship for a given period of time. The relation also defines the direction of the link, or if it is bidirectional.
- Both actors and ties can have *attributes*, e.g., actors may have a gender attribute while ties could have a starting timestamp and a duration.

3.2.1.1 Graphs

Turning now to graph theory, we refer to Brandes and Erlebach (ch. 2, 2005)'s excellent introduction. First, we use the term 'graph' to refer to the mathematical concept of the combination of a set of *nodes* or *vertices*, V, and a set of *edges*, E, each of which represents a connection or *relation* between two nodes $v_i, v_j \in V$ (the *endvertices* of the edge). A *simple* graph is an unweighted, undirected graph, with no loops or multiple edges between nodes. Most of the graphs we discuss in this section will not be simple. We use the term 'network' as the informal term applied to an arrangement of entities and their pairwise inter-relationships that can be modelled as a graph. Simply put, we construct a graph to model a particular domain-specific problem, e.g., the internet is a network of computers and the communication cables between them, which we can model and reason about as a graph. In practice, the terms are often used interchangeably, despite their specific definitions.

For a given graph, $G{=}(V, E)$, V is the graph's nodeset, i.e., $V = \{v_1, v_2, \ldots v_n\}$ and E is its edgeset, i.e., $E = \{e_1, e_2, \ldots e_m\}$. The number, or *cardinality*, of nodes in G is $n = |V|$ and the number of edges is $m = |E|$. For such a graph, an undirected edge e between $v_i, v_j \in V$ is denoted by $\{v_i, v_j\}$, while a directed one from v_i (the *tail* or *origin*) to v_j (the *head* or *destination, respectively*) is denoted by (v_i, v_j).[2] An edge between two nodes implies they are *neighbours* and *adjacent* to each other. The set of nodes adjacent to a given node, v_i, form its *neighbourhood*, which is denoted by $N(v_i)$. Some graphs permit *loops*, a type of edge that links a node to itself.

The strength of a connection (representing, e.g., the frequency of interaction, capacity, physical distance or the similarity of its adjacent nodes) is often modelled by a weight on an edge; such edges belong to *weighted graphs*. Edge weights are represented by the function $\omega : E \to \mathbb{R}$, where $\mathbb{R}$ is the set of all real numbers; this function defines the weight on each edge $e \in E$ as $\omega(e)$ (similarly $\omega(e) = \omega(v_i, v_j)$ if $e = (v_i, v_j) \mid v_i, v_j \in V$). An unweighted graph can be thought of as a weighted graph in which each edge

[2] By introducing direction to edges, the graph is no longer simple.

has a weight of 1. Both nodes and edges may have attributes, in addition to a ‘weight’ value on edges.

The *degree* of a node v_i is the cardinality of its neighbourhood (i.e., its adjacent edge count), so $deg(v_i) = |N(v_i)|$. The set of these edges, for which v_i is an endvertex, is $\Gamma(v_i)$. If the edges are weighted, the *weighted degree* is the sum of the weights on the edges. If the network is directed, each node will have an *indegree*, $deg^-(v_i)$, and an *outdegree*, $deg^+(v_i)$, i.e., a count of all incoming and outgoing edges, respectively. The sets of incoming and outgoing edges for a node v_i are denoted by $\Gamma^-(v_i)$ and $\Gamma^+(v_i)$, respectively.

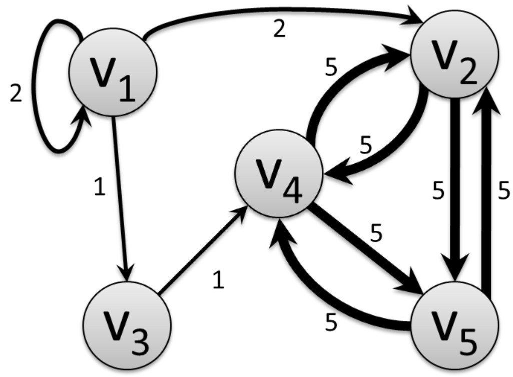


FIGURE 3.2. An example of a directed weighted graph with labelled nodes.

TABLE 3.1. The weighted adjacency matrix for the directed graph in Figure 3.2.

		Destination				
		v_1	v_2	v_3	v_4	v_5
Origin	v_1	2	2	1	0	0
	v_2	0	0	0	5	5
	v_3	0	0	0	1	0
	v_4	0	5	0	0	5
	v_5	0	5	0	5	0

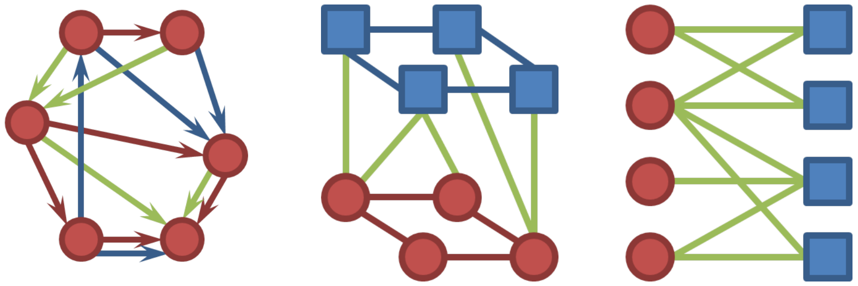
An example of a basic graph is provided in Figure 3.2. This could represent a network created from emails between office workers, or packets sent by computers to each other, or the number of marbles exchanged between school students in lunchtime games. Here, $V = \{v_1, v_2, v_3, v_4, v_5\}$ and we can represent the edgeset, E, as the matrix in Table 3.1. We define the *adjacency matrix* A_{ij}, where $1 <= i, j <= |V|$ as

$$A_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{otherwise.} \end{cases}$$

A *weighted* adjacency matrix simply uses $\omega(e)$ in place of 1 for each edge, as in the example given. The adjacency matrix representation can make some calculations easier. By adding the values in v_2’s and v_4’s columns, we can determine that v_2’s indegree is 12, while v_4’s is 11. Adding the values in v_1’s row, we can see its outdegree is 5. The degree of a node is the sum of values in its row and column, counting the diagonal value only once, e.g., the degree of v_5 is 20. The neighbourhood of v_3 is $\{v_1, v_4\}$. We can see that $\{v_2, v_4, v_5\}$ form a *clique* together, a fully connected set of nodes, and that v_1 has a self-loop edge of weight 2.

Paraphrasing Brandes and Erlebach (p.9, 2005), a *walk* between nodes v_0 and v_k describes the alternating sequence $v_0, e_1, v_1, e_2, v_2, \ldots, e_k, v_k$ of nodes and edges, where $e_i = (v_{i-1}, v_i)$ for directed graphs and $e_i = \{v_{i-1}, v_i\}$ for undirected graphs. If no edge is duplicated (i.e., $e_i \neq e_j, \forall\, i \neq j$), then the walk is a *path*. A *cycle* is a path in which $v_0 = v_k$. A chordlesscycle over k nodes is denoted by C_k. Triangles are a special case, and are sometimes referred to as K_3 cycles. The *length* of a walk, path or cycle is the number of edges it contains. A *connected component* is a set of nodes in a graph, such that each pair of nodes has a path between them. Furthermore, the *distance* between two arbitrary nodes, v_i and v_j, is the length of the shortest path between them (assuming edge weights of 1) and defined as $d(v_i, v_j) = \min\{|P| \mid P$ is a path from v_i to $v_j\}$ (p.295, Brandes and Erlebach, 2005).

3.2.1.2 Complex graphs



(a) A multiplex network. (b) A 2-layer network. (c) A bipartite network.

FIGURE 3.3. Complex graph examples. Colours indicate different node and edge types. In the 2-layer network in Figure 3.3b, the blue squares could represent web pages that link to each other while the red circles are users who email each other, and the green edges indicate that a user has used a URL to a website in one of their emails. In the bipartite graph in Figure 3.3c, the red circles could represent tweets and the blue squares could represent hashtags, and the green edges indicate which hashtags are used in which tweets. The nodes in the multiplex network in Figure 3.3a could represent accounts, while the coloured edges indicate different directed interactions, such as retweet, mention and reply.

Some graphs, known as *multigraphs*, permit multiple edges (of the same direction, if it is a directed graph) between the same pair of nodes. Examples are offered in Figure 3.3. If the edges have different types, the network may be known as a *multiplex* or a *multirelational* graph (Kivela et al., 2014), an example of which is shown in Figure 3.3a. Nodes can also come in different varieties, beyond simply having different labels, and their arrangement can form different types of networks for different types of reasoning (Kivela et al., 2014). If a graph has two sets of nodes, V_1 and V_2, each of a different type, then they can form:

- a type of *multi-layer* graph, in which case each node set forms a layer, and there can be intralayer edges (i.e., connections between the nodes within each layer) and interlayer edges (i.e., connections between nodes in different layers), all of which have different semantics, such as shown in Figure 3.3b; and
- a specialisation, which only permits interlayer edges known as a *bipartite* graph, shown in Figure 3.3c.

Such variations in network structure can introduce complications with some network statistics and algorithms, and addressing these remains an area of active research (Kivela et al., 2014).

3.2.1.3 Graph-level statistics

The following network statistics are used to characterise graphs as a whole: number of nodes, edges, mean degree, mean edge weight, density, number of connected components and the size and diameter of the largest, Louvain (Blondel et al., 2008) cluster count and the size of the largest, reciprocity, transitivity, and largest k-core. Although most of these are self-explanatory, a few warrant explanation.

The *diameter* of a graph $G = (V, E)$, $diam(G)$, is the greatest distance between two arbitrary nodes in the graph (p.296, Brandes and Erlebach, 2005), i.e.,

$$diam(G) = \max\{d(v_i, v_j) \mid v_i, v_j \in V\}. \tag{3.1}$$

Reciprocity r is the proportion of edges in a directed graph that connect two nodes to each other in both directions (sec. 7.10, Newman, 2010). It can be calculated, given $m = |E|$ and A is G's (unweighted) adjacency matrix, with

$$r(G) = \frac{1}{m} \sum_{ij} A_{ij} A_{ji}. \tag{3.2}$$

Density is a ratio of a graph's number of edges to the potential number of edges it could have (i.e., when fully connected) and provides an indication of how sparsely connected (or not) the graph is (p.131, Brandes and Erlebach, 2005). Given $n = |V|$, it is calculated with

$$density(G) = \frac{2m}{n(n-1)}. \tag{3.3}$$

Graphs can exhibit *transitivity* in a similar way to mathematics. In mathematics, a binary operator $\circ$ is *transitive* if, for $x \circ y$ and $y \circ z$ then $x \circ z$, such as is the case for '$=$', '$<$', and '$>$'. This concept is useful to apply to social networks, especially, because if Kelly and Sam are friends and Sam and Alex are friends, then it is more likely than random that Kelly and Alex will also be friends. A common way to consider transitivity in a graph is in its relations: if (v_i, v_j) and (v_j, v_k) are in the graph, the relations are transitive if (v_i, v_k) also always exists (sec. 7.9, Newman, 2010). Fully transitive graphs representing real-world networks are cliques and are rare. *Near transitive* graphs, such as that used to model the social network of Kelly, Sam, and Alex above, are more common. The *global clustering coefficient* is a measure of the level of transitivity in a graph G, which is described by the *transitivity index*, $T(G)$.

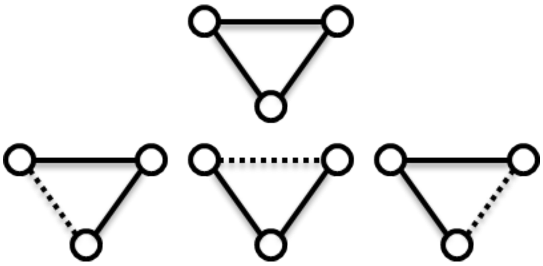


FIGURE 3.4. A triangle (above) consists of three separate paths of length 2 or *connected triples* (below).

To formalise this, we need to define some contributing concepts (p.302, Brandes and Erlebach, 2005). For a path of length 2 in a graph, (v_i, v_j, v_k), if edge (v_i, v_k) also exists, then the path is *closed*. The transitivity index of a graph G is (informally) then calculated with

$$T(G) = \frac{\text{number of closed paths of length 2}}{\text{number of all paths of length 2}}.$$

A *triple* is a set of three nodes, while a *connected triple* is a triple with only two edges, equivalent to a path of length 2. For practical purposes moving forward, when we refer to a triple, we mean a connected triple. Each triangle $\Delta = \{V_\Delta, E_\Delta\}$ of three nodes and the edges between them therefore has three triples, as per Figure 3.4. The number of triangles a node v_i helps form is given by $\lambda(v_i) = |\{\Delta \mid v_i \in V_\Delta\}|$, and the number of triangles in a graph G is defined as $\lambda(G)$. As such, the number of triangles in a network is a third of the number of triangles that its nodes are part of, i.e., $\lambda(G) = \frac{1}{3}\sum_{v_i \in V} \lambda(v_i)$.

A *triple at node* v_i is a triple where v_i is the middle node (i.e., it is adjacent to both other nodes in the triple). Knowing the degree of v_i, $deg(v_i)$, the number of triples v_i is in, $\tau(v_i)$, is given by

$$\tau(v_i) = \binom{deg(v_i)}{2} = \frac{deg(v_i)(deg(v_i)-1)}{2} = \frac{deg(v_i)^2 - deg(v_i)}{2},$$

and so the number of all the triples in a graph is $\tau(G) = \sum_{v_i \in V} \tau(v_i)$. In this sense, $\tau(G)$ is the number of closed and unclosed (i.e., all) triples in the graph. The number of triangles in the graph, $\lambda(G)$, is three times the number of closed paths of length 2, and so we can formulate $T(G)$ with

$$T(G) = \frac{3\lambda(G)}{\tau(G)}. \tag{3.4}$$

A related concept is based on the *local clustering coefficient* (p.303, Brandes and Erlebach, 2005), which is the ratio of a node's neighbours that are also connected. Socially, this is a useful concept for describing what are known as *structural holes* (sec. 7.9.1, Newman, 2010), where neighbours of a node are not connected, giving the node some degree of control over information flowing between the two neighbours (as they lack their own connection). This is related to betweenness centrality, which is much more expensive to calculate (given it relies on knowing all shortest paths in a graph), but is more locally focused. Formally, the clustering coefficient of a node v_i with $\tau(v_i) \neq 0$ (i.e., v_i is at the centre of at least one triple) is

$$c(v_i) = \frac{\lambda(v_i)}{\tau(v_i)}. \tag{3.5}$$

Extending this, the clustering coefficient for the whole graph G, $C(G)$ is the average of the local clustering coefficients. Considering only nodes at triples, $V' = \{v_i \in V \mid deg(v_i) >= 2\}$, we calculate

$$C(G) = \frac{1}{|V'|} \sum_{v_i \in V'} c(v_i). \tag{3.6}$$

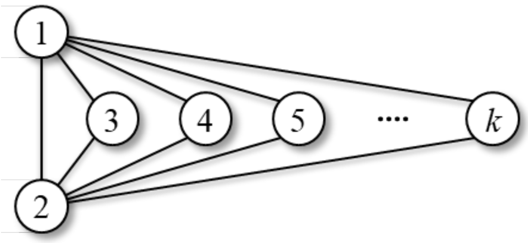


FIGURE 3.5. A graph with k nodes, which has a higher clustering coefficient and lower transitivity, as k increases. This figure is a reproduction of Figure 11.2 in (p.303, Brandes and Erlebach, 2005).

Though both $T(G)$ and $C(G)$ lie in $[0, 1]$ and describe the proportion of complete triangles in a graph, they work differently and give very different results for some graphs (p.304, Brandes and Erlebach, 2005). Newman (sec. 7.9.1, 2010) notes that low-degree nodes tend to dominate $C(G)$. Brandes and Erlebach (p.304, 2005) provide an example of a graph with k nodes, where $k >= 3$ and two of the nodes connect to all the other nodes (recreated in Figure 3.5). As more nodes are added, $k \mapsto \infty$; of these nodes, only two have $c(v_i) \neq 1$ so $C(G) \mapsto 1$, and more unclosed triples are added, so $T(G) \mapsto 0$ as its denominator rises. Though both approaches are valid for specific types of research endeavour, the local clustering coefficient is better suited to localised considerations, while transitivity may better suit all-of-graph calculations, such as how tightly communities are connected (*cf.*, near transitivity).

From a practical perspective, these measures relate to density but give a better sense of the how clustered the nodes are, which in turn relates to the distribution of node degrees. If every node in the network has the same degree, then it will not have many clusters, but if the degrees vary greatly, more clustering will be present. Transitivity and the clustering coefficient reveal this, without requiring analysis of the degree distribution.

Finally, a k-core is a maximally connected subset of nodes in which each node is connected to at least k of the other nodes in the set (sec. 7.8.1, Newman, 2010). A graph with a large k-core (i.e., meaning k is large) has a highly cohesive core or cores (a graph may have more than one). The ratio between the maximum k-core and the overall graph size (in nodes) also provides a sense of how cohesive the graph is overall.

These measures provide us with an understanding of the 'shape' of the networks in terms of how broad and dense they are and the strength of the connections within.

3.2.1.4 Centrality

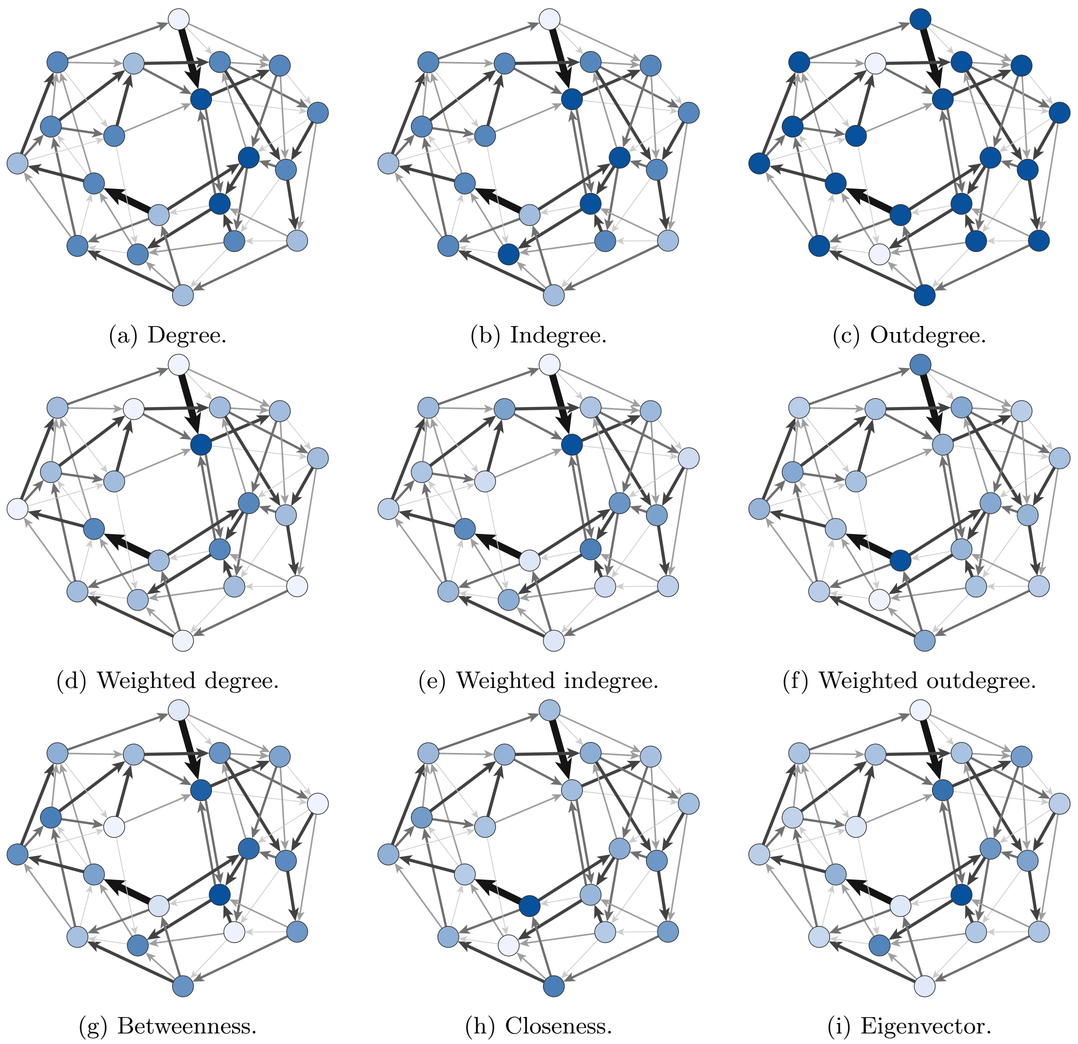
Centrality measures offer a way to consider the importance of individual nodes within a graph (Brandes and Erlebach, 2005; Newman, 2010). The centrality measures for a node v_i that we consider here include the following.

The *degree* centrality $c_D(v_i)$ indicates how many other nodes the node v_i is directly linked to, i.e., $c_D(v_i) = deg(v_i)$ (p.20, Brandes and Erlebach, 2005). The *indegree* centrality $c_{iD}(v_i)$ indicates how many incoming edges v_i has in a directed graph (i.e., how many edges for which it is the head or destination), thus $c_{iD}(v_i) = deg^{-}(v_i)$. Correspondingly, the *outdegree* centrality $c_{oD}(v_i)$ indicates how many outgoing edges v_i has in a directed graph (i.e., how many edges for which it is the tail or origin), and so $c_{oD}(v_i) = deg^{+}(v_i)$.

The *weighted degree* centrality $c_{wD}(v_i)$ is the sum of the weights of v_i's adjacent edges, i.e., $c_{wD}(v_i) = \sum_{v_j \in \Gamma} \omega(v_i, v_j)$. Correspondingly, the *weighted indegree* and *weighted outdegree* centralities $c_{wiD}(v_i)$ and $c_{woD}(v_i)$ are the sums of the weights on v_i's incoming and outgoing edges, respectively, and are calculated with $c_{wiD}(v_i) = \sum_{v_j \in \Gamma^-} \omega(v_j, v_i)$ and $c_{woD}(v_i) = \sum_{v_j \in \Gamma^+} \omega(v_i, v_j)$.

Betweenness centrality refers to the number of shortest paths between all pairs of nodes in the graph that a node is on and thus to what degree the node is able to control information flowing between other nodes (p.29, Brandes and Erlebach, 2005). Given σ_{st} denotes the number of shortest paths between arbitrary nodes $s, t \in V$, and $\sigma_{st}(v_i)$ is the number of them that contain v_i, we need to define the proportion of shortest paths in the graph that contain v_i:

$$\delta_{st}(v_i) = \frac{\sigma_{st}(v_i)}{\sigma_{st}}.$$



(a) Degree.

(b) Indegree.

(c) Outdegree.

(d) Weighted degree.

(e) Weighted indegree.

(f) Weighted outdegree.

(g) Betweenness.

(h) Closeness.

(i) Eigenvector.

FIGURE 3.6. Centrality examples demonstrated with a 20-node small world weighted directed graph. Darker nodes have higher centrality values, and darker, wider edges have greater weight. Centrality values for the second and third rows were calculated using the edge weights. The closeness centrality values in Figure 3.6h were calculated using the inverse of the weight attributes on each edge as a proxy for distance. Figures 3.6a and 3.6d are calculated ignoring edge direction, and reciprocal edges are aggregated.

The betweenness centrality c_B for node v_i is then

$$c_B(v_i) = \sum_{s \neq v_i \in V} \sum_{t \neq v_i \in V} \delta_{st}(v_i). \tag{3.7}$$

Closeness centrality $c_C(v_i)$ provides a sense of how topologically close a node v_i is to the other nodes in the graph, and thus is maximised as the reciprocal of the sum of the v_i's distances to all other nodes in the graph (p.22, Brandes and Erlebach, 2005):

$$c_C(v_i) = \frac{1}{\sum_{v_j \in V} d(v_i, v_j)}. \tag{3.8}$$

The last centrality measure we consider, *eigenvector* centrality $c_E(v_i)$, measures how important a node v_i is based on the importance of nodes to which it is connected (sec. 7.2, Newman, 2010). The eigenvector centrality values are calculated iteratively, as initially the importance of nodes is unknown, but it is calculated with

$$c_E(v_i) = \kappa_1^{-1} \sum_{j \in V} A_{ij} c_E(v_j). \tag{3.9}$$

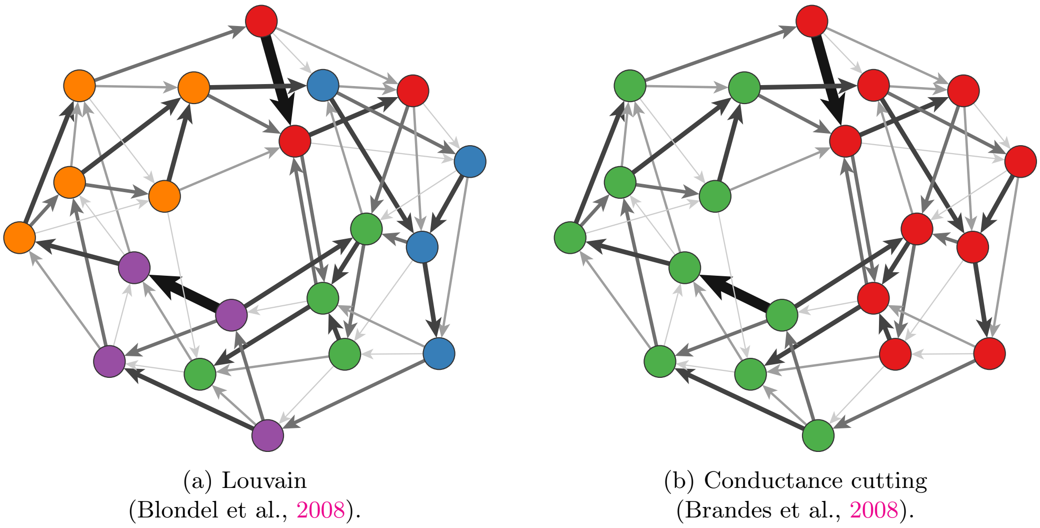
where κ_1 is the largest eigenvalue of A, and thus the eigenvector centrality of node v_i is proportional to that of its neighbours. Eigenvector centrality is often compared with Google's PageRank algorithm (Brin and Page, 1998), which gives a measure of the importance of nodes (e.g., websites) based on the importance and count of other nodes that reference them.

An example is shown in Figure 3.6 in which subfigures show how the centrality scores vary in the same graph. Beyond the works of Brandes and Erlebach (2005) and Newman (2010), further details can be found in the works of Robins (2015) and Wasserman and Faust (1994).

Incidentally, considering the k-core concept mentioned in the previous section, a node may belong to many k-cores with different values of k, so its highest k-core value indicates how deeply embedded the node is within the graph. Whereas a node's centrality can indicate its importance within a graph, its maximum k-core score more clearly indicates its location within the graph. After all, a node may have a very high eigenvector centrality score even if it is on the periphery of a graph, as long as it is connected to enough other nodes with high eigenvector centrality scores.

3.2.1.5 Groups

Communities in graphs can be revealed in a number of ways. They can be manually labelled, based on knowledge of the nodes (e.g., they could represent people from particular organisations or countries). Alternatively, they can be calculated computationally based on the structure of the network and the information in its nodes and



(a) Louvain (Blondel et al., 2008).

(b) Conductance cutting (Brandes et al., 2008).

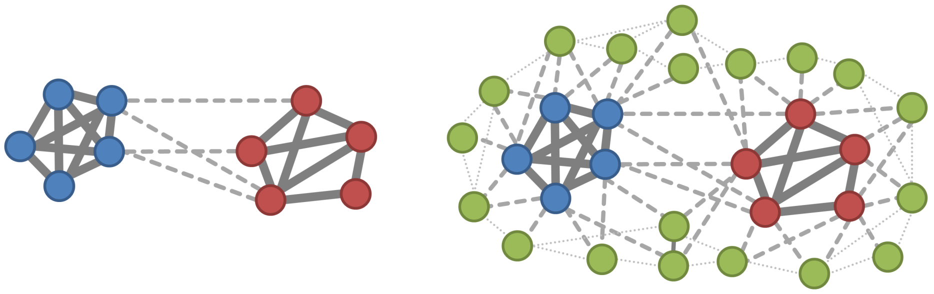
FIGURE 3.7. Examples of the 20-node small world graph in Figure 3.6, in which clusters detected with different algorithms are highlighted.

edges with clustering algorithms, such as k-nearest neighbour (kNN), Focal Structures Analysis (FSA, Şen et al., 2016), the Louvain algorithm (Blondel et al., 2008) and conductance cutting (Brandes et al., 2008). The Louvain method works well with large and small networks (Yang et al., 2016) and is well known in social media analysis research (e.g., Morstatter et al., 2018; Nasim et al., 2018; Nizzoli et al., 2021). kNN is used by Cao et al. (2015) in their study of URL-sharing campaigns. Other methods used in the literature include Markov method (Fazil and Abulaish, 2020) and multi-view modularity clustering (Magelinski et al., 2021). Examples of Louvain and conductance cutting are shown in Figure 3.7. Louvain identifies small, tightly connected sets of nodes, while conductance cutting forms communities by dividing the graph where the connections are weakest. Both attend to edge weight, but not all methods do.

The difference between cluster detection and community extraction algorithms is that cluster detection will typically assign every node in a graph to a cluster, whereas community extraction may only identify a subset of nodes as community members. A good example of a community extraction method is to search for non-overlapping cliques, fully connected subsets of nodes which share no nodes (i.e., the intersection between the two subsets is empty, p.114, Brandes and Erlebach, 2005). Cliques are somewhat rare in practice, so a variety of other near-cliques have been defined, such as the k-cores mentioned above, or distance-based measures like N-cliques, N-clubs, and N-clans (p.115, Brandes and Erlebach, 2005). FSA also specifically seeks out influential sets of nodes in a graph that may only be near-cliques (Şen et al., 2016).

3.2.1.6 Homophily

The online communities formed by shared ideas, ideals and beliefs provide examples of the sociological concept of *homophily*, which is the tendency for individuals to prefer to



(a) Only polarised groups.

(b) The broader network.

Figure 3.8. Two groups polarised over a contentious issue with their internal and external edges, representing positive interactions or similarities, highlighted (dark solid edges are strongest and are internal to each group, lighter dashed edges are moderately strong and connect group members to the broader network, and the light dotted edges are not relevant to the groups). The subfigure on the right includes the groups' edges into broader network. Here, the red and blue actors are highly homophilic with respect to one another, but they all interact with many green nodes.

connect to or interact with other individuals who are similar in some way (Rogers and Bhowmik, 1970; McPherson et al., 2001). For social networks where edges represent positive interactions regarding contentious issues (e.g., agreement or support), then homophily may correlate with polarisation within the network, particularly with regard to some communities within the network. When characterising communities and groups in social networks, beyond simple frequency metrics of numbers of accounts, interactions, and ratios like internal to external connection counts (i.e., how many connections are between members inside a group versus connections between members of different groups), we primarily rely on two measures when considering homophily: the assortativity coefficient (Newman, 2003) and a variation on the Krackhardt E-I Index (Krackhardt and Stern, 1988). Assortativity is a calculation of the degree to which nodes connect to similar nodes, based on a specified value of 'similar', agnostic of network semantics. In this thesis, it is typically defined by the node's label attribute, which refers to the 'name' of the node's community. This measure makes no use of edge weights. The Krackhardt E-I Index is a simple ratio of edges internal to a community (i.e., between community members) and edges external to that community (i.e., edges which have only one endpoint within the community). In a graph, $G = (V, E)$, for a community c consisting of a nodeset $V' \subseteq V$ and for which the corresponding edgeset is $E' = \{(v_i, v_j) \mid v_i \in V'\}$, its internal edges are $E'_{int}(c) = \{(v_i, v_j) \mid v_i, v_j \in V'\}$ while its external edges are $E'_{ext}(c) = \{(v_i, v_j) \mid v_i \in V', v_j \in V - V'\}$.[3] Its E-I Index $EIidx$ is given by

$$\text{EIidx}(c) = \frac{|E'_{ext}(c)| - |E'_{int}(c)|}{|E'_{ext}(c)| + |E'_{int}(c)|} \tag{3.10}$$

Extending this, the E-I Index of a graph with k known non-overlapping communities,

[3] NB, incoming directed edges are not included in E'_{ext}. If the edges are directed, we count only interactions 'reaching outwards' from inside the community, and if the edges are undirected (e.g., when using the same hashtag), then incoming edges are not distinct from outgoing ones.

$C = \{c_1, c_2, \ldots, c_k\}$, can be calculated only considering the nodes in the communities together $c_{all} = \bigcup_{i=1}^{k} c_i$ and the edges between them $\{(v_i, v_j) \mid v_i, v_j \in c_{all}\}$ (as in Figure 3.8a), in which case it measures the homophily of the communities with respect to one another. Alternatively, edges to the remainder of the network can also be considered using the edgeset $\{(v_i, v_j) \mid v_i \in c_{all}, v_j \in V\}$ (as in Figure 3.8b), in which case it measures the homophily of the communities but also accounts for connections with the broader network. The E-I Index for the set of communities C is given by

$$\text{EIidx}(C) = \frac{\sum_i |E'_{ext}(c_i)| - \sum_i |E'_{int}(c_i)|}{\sum_i |E'_{ext}(c_i)| + \sum_i |E'_{int}(c_i)|} \tag{3.11}$$

Our variation takes into account the weights of edges, because the weights represent the frequencies of individual interactions.[4] This ensures that the strength of connections between nodes is considered, rather than simply the size of the neighbourhood. Both measures lie within $[-1, 1]$, but their meaning is reversed: an assortativity score close to 1 implies high polarisation, with the majority of edges connecting nodes with the same label, whereas an E-I Index of 1 implies that all edges reach outside the group and no edge joins members of the same group. A value of 0 for both metrics implies a balance between internal and external edges.

Using the graphs in Figure 3.8, we can illustrate the two ways to use the homophily metrics mentioned above: one considers only the polarised groups (Figure 3.8a), while the other also considers them in the context of the rest of the network (Figure 3.8b). Figure 3.8a shows two categories of nodes (red and blue), with the internal edges coloured the same as the nodes they join and the external edges coloured purple. For the combined set of members of known polarised communities c_{all}, $|E'_{int}(c_{all})| = 18$ while $|E'_{ext}(c_{all})| = 4$, resulting in a high E-I Index of -0.64 and the conclusion that the red and blue category nodes are highly homophilic. When we consider the broader network in Figure 3.8b, including the green category of accounts, $|E'_{ext}(c_{all})|$ rises to 38, which shifts the E-I Index to 0.36, suggesting that the red and blue category nodes are actually only moderately heterophilic.[5] Using both of these variations allows us to see whether the polarised groups indeed form a filter bubble (in which case both E-I Index values would be negative) or whether they are just minimally connected to one another but are still strongly connected to the broader community, forming echo chambers, as is the case in our example in Figure 3.8b.

Binomial tests are used to test the statistical significance of the homophily measures. We consider p-value thresholds of 0.05, 0.01, 0.001, and 0.0001 to express the confidence in the significance.

[4] Edge weights are ignored in the implementation of the E-I index in the version of NetworkX (Hagberg et al., 2008) that we used, version 2.5, which is why we implemented our own.

[5] NB, the orange edges between green nodes are included in the figure for completeness but are not used in the calculation of the E-I Index.

It is important to note the importance of the choice of edge in network construction to determine if homophily measures can provide insights into community polarisation. It may be that homophily measures, as a quantification of clustering, can help direct the researcher to strong communities which may hold very similar and potentially strong views within networks based on contentious issues. If the network is based on tweet replies and there is significant argumentation, then edges will predominate between members of opposite camps, in which case measures such as assortativity and the E-I Index will negatively correlate with polarisation.

For these reasons, when we discuss polarisation in networks in this thesis, it is with some knowledge of the dominant opinions, narrative or content produced by two communities on either side of a divisive issue or position (e.g., whether abortion should be permitted as a choice of the child bearer, or whether immigration should be increased) or set of opinions that together form an ideology (e.g., left-wing versus right-wing politics).

3.2.1.7 Structural analysis and visualisation

Social theories of friendship indicate that not all ties are equal, and we have options to define the strength of ties in our networks. For networks based on interactions and content, it is possible to use frequencies as edge weights, but agnostic of the edge semantics (i.e., the reason for the presence of the edge), we can use the quadrilateral Simmelean backbone to identify the strongest ties in a given social network (Nick et al., 2013; Nocaj et al., 2014). This approach gives high weight to edges embedded in cycles of length 4. The intuition behind this approach is that dyads that share more common neighbours (meaning they are part of a triangle, K_3, or cycle, C_4, Nastos and Gao, 2013) are more strongly tied – this weight is therefore referred to as the *backbone strength* of the edge. This can be used in the rendering of edges, but also the layout of network nodes.

Visualising networks using force-based layouts can also provide insight into their macro-, meso- and micro-structures (Tollis et al., 1999; Brandes and Erlebach, 2005). When visualising these networks,[6] nodes can be laid out using the union of all maximum spanning trees as a sparsifier (referred to herein as the *backbone layout*, Serrano et al., 2009; Nocaj et al., 2014) and edge colours can indicate how strongly the ties are embedded. Naturally, the ties in the core of cohesive subgroups are strongly embedded compared with those on the periphery or those between subgroups. Alternatively, a spring embedder layout can be used to decide node placement (Tollis et al., 1999). Nodes representing accounts are often most meaningfully coloured according to the group to which they belong and may be sized by indegree, outdegree or activity (i.e., how many tweets they posted), depending on the semantics of the network edges.

[6] The network visualisations in this thesis were created with *visone* (https://visone.info) and Gephi (https://gephi.org), or designed using Microsoft PowerPoint.

Alternatively, the nodes may be sized according to weighted degree centrality (Brandes and Erlebach, 2005), whether the weight is the activity frequency or backbone strength values mentioned above.

3.2.2 Practice

In the context of social media, we can use the following to guide our modelling decisions.

- For social networks, the actor would be typically the (human) user of an account, but the only accessible data is the social media data itself, so our only option is the *representation* of the user: the account. Furthermore, a user (a real person) may control multiple accounts. For this reason we typically focus on modelling the account. For non-social networks, other social media elements can be represented as nodes, such as hashtags and URLs or even social media posts. The way themes relate to one another in an online discussion can be observed in how hashtags are used together in *semantic networks* (Radicioni et al., 2021). Significant insight can be gained by analysing 2-layer networks of accounts and URLs or hashtags to see how people use them, and social behaviours can be observed in the conversation structures of reply chain-based trees (e.g., Gonzalez-Bailon et al., 2010; Ackland, 2020).

- Ties may be based on static connections, such as friend or follower links, or membership to a particular WhatsApp group, or they may be more tightly related to a particular period or specific time, such as a video call (which has a start time and duration), or a comment on a Tumblr post (which has an occurrence time only). Where retweeting is a clear example of a directed tie, an association between two accounts because they retweeted the same tweet is non-directional. We regard this last as an example of a *co-activity*. Ties between nodes of different types can often indicate a ‘use’ or ‘includes’ relation, such as between an account and a URL, or a tweet and a hashtag, respectively.

  Ties are representative of more than just intermittent connections between accounts based on short-lived interactions, they inform us of flows of information, beliefs, ideas and influence (Borgatti et al., 2009), and the direction of the tie may need to change to represent each of these. If account B retweets a tweet by A, then a directed edge between A and B could represent many aspects, depending on the research question at hand. It may indicate just that B has retweeted A (B’s retweet includes a reference back to A’s tweet), so the A is the edge’s destination, or it could mean that B is part of A’s audience and the arrow refers to A’s effective reach or the flow of information or influence from A to B, so B may be the edge’s destination. This flexibility in representation underscores the importance of using clearly designed research questions.

- The relation defines the type of the link. In social networks, there are four categories of such links: similarities (e.g., attributes or memberships), social relations (e.g., kinship or role in a group), interactions (e.g., retweets), and flows (e.g., information or influence, Borgatti et al., 2009). They may have direction or be bidirectional. For example, retweets, comments on a forum post and follow links (such as those available on Twitter, Facebook, and Tumblr) are all directional, while a Facebook 'friend' relation is bidirectional, as both parties engage in the relationship. The initial 'friend request' to establish the relationship is, however, directional, as it requires one party to sent the request to the other.
- Node attributes often include an ID, a name or label (useful for human-readable representations), and often a category of some kind, for distinguishing groups of nodes. Edges can have weights, as discussed, but may also include a category or label.

The majority of network analysis in this thesis relies on three simple networks: interaction networks, semantic (or hashtag co-occurrence) networks, and coordination networks. Both interaction and coordination networks are social networks inasmuch as their nodes represent accounts, while the nodes of semantic networks are hashtags.

3.2.2.1 Interaction networks

A variety of information is available to build social networks of accounts from OSN data. In traditional SNA, relations are evidence of long-standing relationships between actors, such as familial or friend relations, or organisational structures, such as supervisory or collaborative relations, but online connections differ (Wasserman and Faust, 1994; Nasim, 2016; Borgatti et al., 2009). Even 'friend' links on Facebook, which possibly provide the best online analogy for an offline social relationship, are relatively easy to create, but then forget, especially with activity feed algorithms that prioritise 'best' friends and those most interacted with. As a result, these links can quickly become stale and meaningless, a fact that is not always apparent when obtaining such data from OSN APIs. The situation is even worse on microblogs such as Twitter, because a follower link typically does not require the attention of the followed account, so a user might see a tweet that appeals to them, decide to follow its author in case they ever post similar content, and then never see another of the author's tweets, either because they stopped posting them, or because they were lost in the user's activity feed, and so the follower link remains, but it is hard to say any relationship exists. Beyond follower relations, most OSNs provide no other data on long-standing relations between accounts. Instead, direct interactions between accounts, such as mentions, comments and shares or retweets, can provide evidence of the currency of connectivity, the degree of interaction activity and its direction, and thus we focus on these interactions to study online communities.

Four social networks built from interaction types common to many OSNs are 'mention' networks, 'quote' networks, 'reply' networks, and 'retweet' networks (retweets are analogous to Facebook shares or Tumblr reposts, and replies are analogous to comments on posts on Reddit, as shown in Table 2.1). We define a social network $G{=}(V, E)$ of accounts $u \in V$ linked by directed, weighted edges $(u_i, u_j) \in E$ based on the criteria below. For polarisation studies, it is useful for nodes to have a 'category' attribute to hold the ID or label of the group to which they belong (or are assigned).

Mention networks Twitter users can *mention* one or more other users in a tweet. In a mention network, an edge (u_i, u_j) exists if and only if u_i mentions u_j in a tweet, and the weight corresponds to the number of times u_i has mentioned u_j.

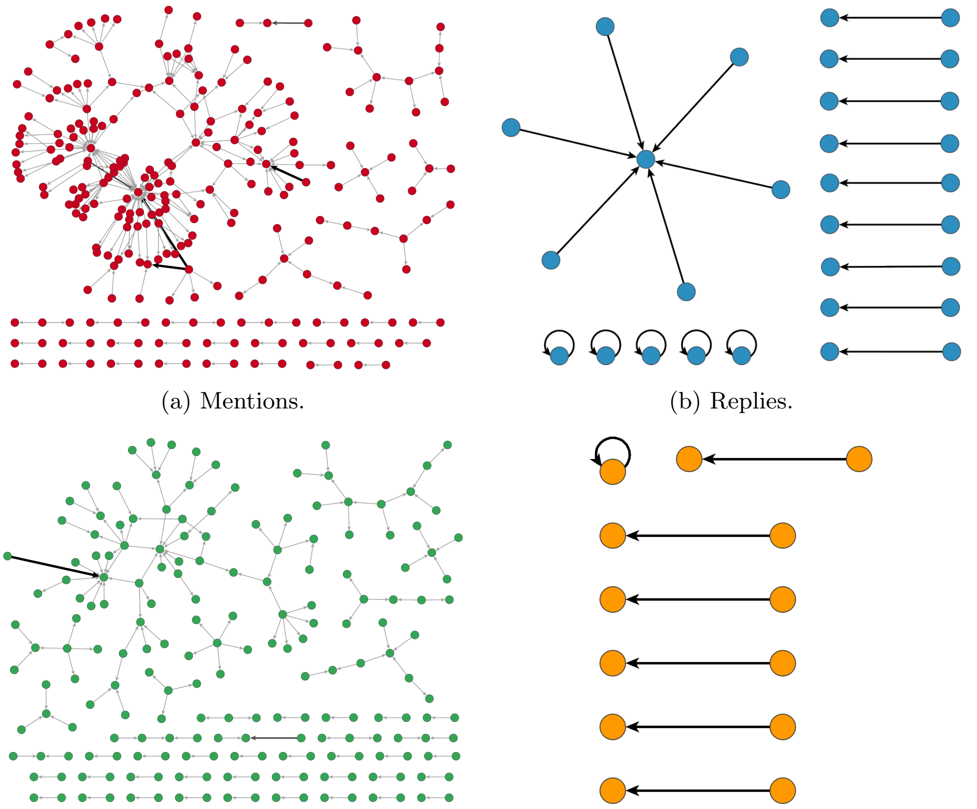
Reply networks A tweet can be a reply to one other tweet. In a reply network, an edge (u_i, u_j) exists if and only if u_i replies to a tweet by u_j, and the weight corresponds to the number of replies u_i has made to u_j's tweets.

Retweet networks A user can repost or 'retweet' another's tweet on their own timeline, which is then visible to their own followers. Though retweets are not necessarily direct interactions (Ruths and Pfeffer, 2014), which we elaborate on below, they can be used to determine an account's reach, and are widely used in the literature (e.g., Vo et al., 2017; Rizoiu et al., 2018; Woolley and Guilbeault, 2018; Morstatter et al., 2018; Gupta et al., 2019; Mazza et al., 2019). In a retweet network, an edge (u_i, u_j) exists if and only if u_i retweets a tweet by u_j, and its weight corresponds to the number of u_j's tweets u_i has retweeted.

Quote networks Similar to retweeting, 'quoting' is equivalent to adding a comment while sharing a post on Facebook or Tumblr, and the semantics are the same as retweets. In a quote network, an edge (u_i, u_j) exists if and only if u_i quotes a tweet by u_j, and its weight corresponds to the number of u_j's tweets u_i has quoted.

To illustrate the differences between the four types, we present networks built from them from the same dataset in Figure 3.9. Of the four interaction types, quotes (Figure 3.9d) are the least common, while replies (Figure 3.9b) are the next least frequently used. All but quotes are dominated by a single large component. Mention networks (Figure 3.9a) exhibit relatively high cohesiveness. The similarity between retweets (Figure 3.9c) and mentions is because the data model of a retweet includes a mention of the retweeted account, and thus the retweet edges form a subset of the mention edges. Removing these implicit mention links, if they are unwanted, would be part of data preparation, after collection but prior to network construction.

There is no clear comparability between interaction types (after all, absent of any further context, is a mention worth the same as a quote?), so we do not combine interaction networks with different types of edges without very good reason. Quotes and retweets *could* be merged without significant concerns regarding the semantics of the interactions (they are both further disseminating a tweet), but whereas a retweet is



(a) Mentions.

(b) Replies.

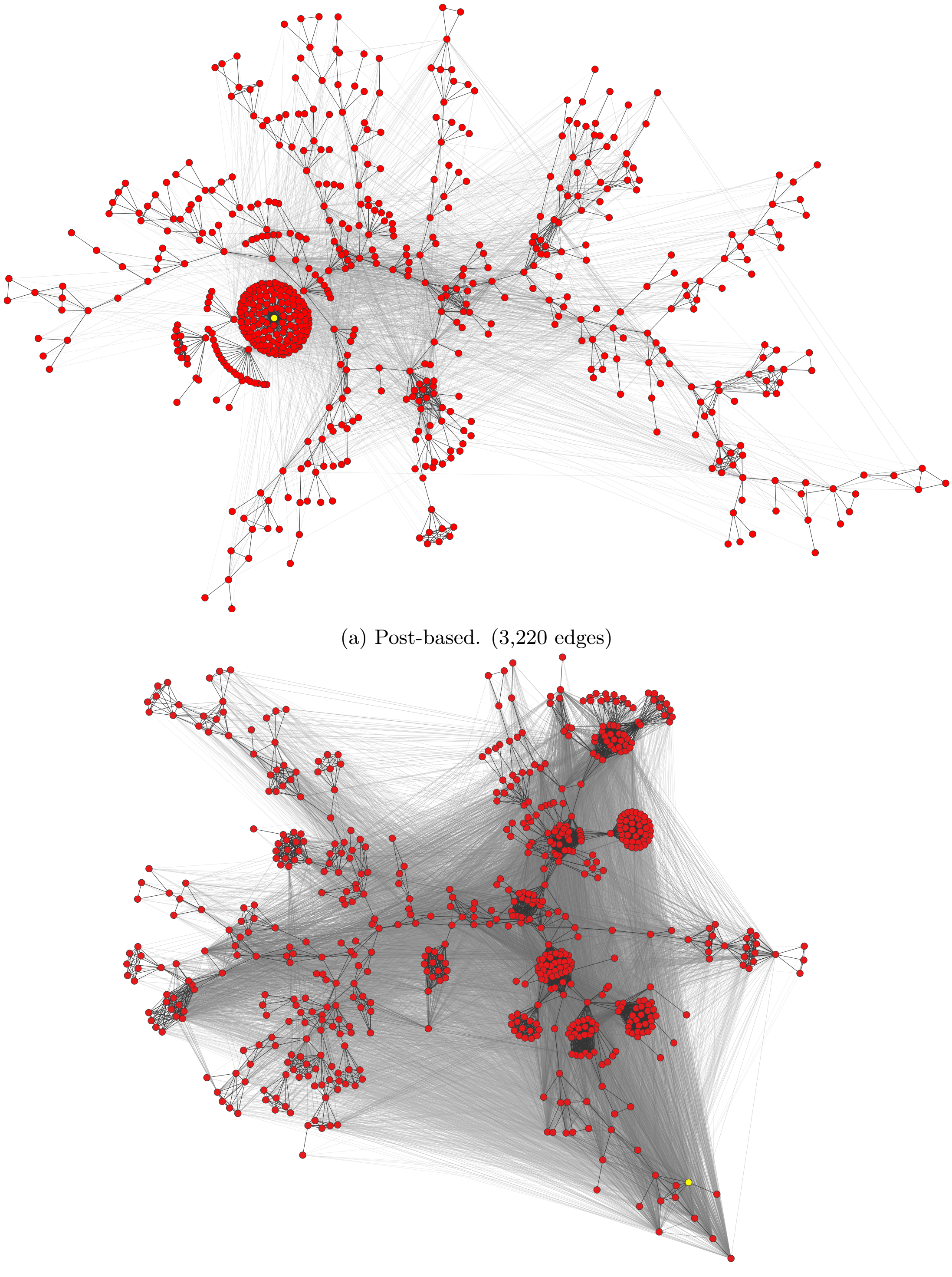
(c) Retweets.

(d) Quotes.

Figure 3.9. Sample networks of accounts built from 5 minutes of Twitter data. Nodes may appear in one or more networks, depending on their behaviour during the sampled period.

often seen as an endorsement (Metaxas et al., 2015), a quote may introduce a negative interpretation of the quoted tweet. Merging interactions through the use of multi-layer modelling and interaction-specific edge types, as exploited by others (Magelinski et al., 2021) examining `URL+hashtag` combinations, is a possibility for specific problems, but introduces complexities that are bound to the research question.

As an aside, Twitter's retweets introduce a specific semantic complication. In the metadata of a retweet, the details of the original retweeted tweet are provided, but, if the retweet in question is a retweet of another retweet of the original tweet, that path information is not. For this reason, edges in our retweet networks always refer back to the original tweeter, and are thus limited when studying information flow to a certain degree (Ruths and Pfeffer, 2014). Research is underway to reconstruct possible retweet paths via probability distributions (Rizoiu et al., 2018). Similar approaches are used in the study of contagions (e.g., Gray et al., 2020). These efforts will ultimately underpin methods based on mathematical models, such as point processes, better than discrete network approaches. In contrast, the metadata in replies, mentions, and quotes describe true direct interactions between accounts.



(a) Post-based. (3,220 edges)

(b) Account-based. (21,707 edges)

Figure 3.10. Semantic networks (based on hashtag use) of the Supporter community using the hashtag #ArsonEmergency in early 2020, as discussed in Chapter 5. The 645 nodes are hashtags, linked when used in the same tweet (Figure 3.10a) or by the same account (Figure 3.10b). All hashtag nodes are red, except #ArsonEmergency, which is highlighted in yellow. Edge width indicates frequency of co-use (i.e., the number of tweets a pair of hashtags appeared in and the number of accounts that used a pair of hashtags), while darkness is determined by *backbone strength*, provided by the *backbone* layout in *visone*. The layout uses the quadrilateral Simmelian backbone to calculate the importance of edges and guide the layout of nodes to cluster those most embedded in the network (Nick et al., 2013; Nocaj et al., 2014).

3.2.2.2 Semantic networks

Hashtags can be regarded as proxies for content, so to characterise the nature of a discussion in terms of the topics and themes that arise and how they inter-relate, we can construct a network of hashtag nodes, linked when they are used by the same accounts or in the same social media posts. Such hashtag networks are sometimes referred to as *semantic networks* (Radicioni et al., 2021; Ackland, 2020), but can also be more prosaically labelled *hashtag networks*. The post-based and account-based networks shown in Figure 3.10 clearly have different structures: each of the branch formations in the post-based network in Figure 3.10a give a clear indication of the hashtags that are used together, which offers an indication of their narrative or line of argument, whereas the cluster structures of the account-based network in Figure 3.10b tell us more about the topics being discussed by groups of individuals over time (i.e., not in the same posts). The tweets that these networks were based on were collected using a specific term, 'ArsonEmergency', the vast bulk of which appeared as the hashtag `#ArsonEmergency` (99.7%), which is the focus of discussion in Chapter 5, and thus the hashtag is included in almost every tweet (and is used by almost every account). This results in its node being centrally located in the tweet-based network, highlighted as the yellow node in Figure 3.10a, but the discussion topics clearly vary significantly given its location in the account-based network (highlighted again in yellow in Figure 3.10b). Using a clustering method such as the Louvain method (Blondel et al., 2008) and colouring hashtags by their clusters can provide a further statistical measure of hashtag relations.

3.2.2.3 Coordination networks

Coordination networks consist of accounts linked by evidence of coordination, for a given value of 'coordination'. What is regarded as coordination is domain-specific and can vary accordingly, but the focus of much of the literature discussed in Section 2.6 has been on similar or related activities, often conducted in brief timeframes, often aimed at amplifying content or a particular message or narrative. Examples include: co-activities, such as retweeting the same tweet, using the same URL or hashtag, or mentioning the same account (e.g., Ratkiewicz et al., 2011; Keller et al., 2017); temporal proximity (e.g., Chavoshi et al., 2017; Dawson and Innes, 2019; Pacheco et al., 2020; Broniatowski, 2021); using the same or similar media (Pacheco et al., 2021; Yu, 2021); or sharing screen names (Ferrara, 2017; Mariconti et al., 2017).

In essence, coordination networks are weighted undirected networks of accounts. As discussed in Section 2.6, there is no settled terminology in the literature, with the same concept referred to variously as "user similarity networks" (Nizzoli et al., 2021), "synchronous action networks" (Magelinski et al., 2021), "account networks" (Pacheco et al., 2021), and "latent coordination networks" (Chapter 7). The names used speak to the nature of the evidence discussed in the works, with both Nizzoli et al. (2021) and Pacheco et al. (2021) using a wide range of methods to determine 'similarity',

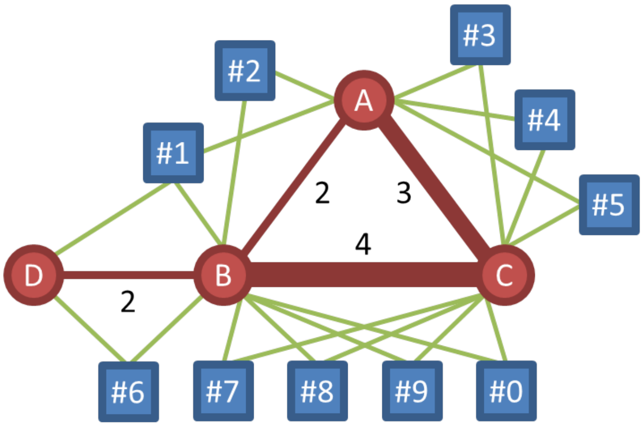
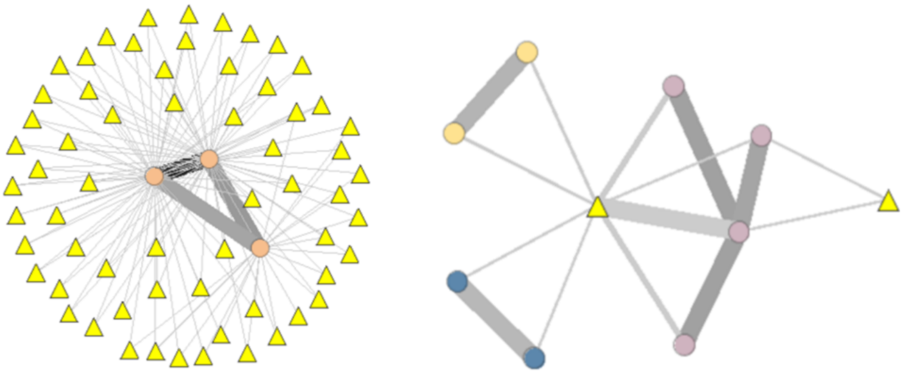


FIGURE 3.11. An example of a coordination network consisting of accounts (red circles) using URLs (blue squares). Each green edge indicates a single use of an URL by an account. The red weighted edges indicate the degree of 'coordination' between each pair of accounts (i.e., the number of URLs they both used within a specified timeframe). The red edges are labelled with their weights. NB, A and D are not connected, as they did not share the same URL within the time window, though A and B did, and B and D did, i.e., the time windows of A and B's coordination overlapped with B and D's.

while Magelinski et al. (2021) and we have focused on specific online actions or co-activities (particularly interactions). As we discuss in Chapter 7, we link two accounts when they engage in the same co-activity within a constrained timeframe. Because of the likelihood of incidental 'coordination' in this sense (e.g., using the same hashtag), care must be taken in data selection and network construction. Using focused datasets and specifically designed criteria for evidence of coordination will minimise dataset size and improve the likelihood of finding genuine coordination (as well as reducing computation costs), but this process will be aided by the choice of community extraction, once the network is constructed.

It is important to note that coordination networks created as described above are, in fact, aggregations of pairwise associations. Examining the example provided in Figure 3.11, we can see accounts A, B, C and D (red circles) are each linked according to how many URLs they use within a specified constrained timeframe (blue squares). Each green edge indicates a single use. While the fact that D and B both use URL #1 in the same time window can imply a potential association with A, as they all use the same URL (just in different, but overlapping time windows), the fact that D and B both use URL #6 does not. The result is that by only considering the account network ABCD we may come to the conclusion that D is more tightly associated with ABC than it actually is. The DB association may be the result of coincidence or it may be genuine coordination at a low level of intensity, and will likely require a domain expert's opinion to tell the difference, however it is in these situations that the choice of community extraction is important. A simple edge weight filter (e.g., remove all edges of weight 2 or less) may remove important edges (leaving nodes as isolates), thus obscuring important relationships.

3.2.2.4 Account/reason networks



(a) Accounts coordinating by sharing many URLs.

(b) Accounts coordinating by focusing on two URLs.

FIGURE 3.12. Genuine account/URL networks appearing in coordinating groups detected during the Republican National Convention in 2020 sharing URLs within ten-second windows (discussed in Section 7.4.4). Accounts are represented as circles, coloured according to Louvain cluster (Blondel et al., 2008), and the URLs they shared are represented as yellow triangles. The width and darkness of edges indicates the strength of coordination detected. The network in Figure 3.12a is a set of accounts sharing the same many URLs in short succession, which is a pattern often observed in automated news media and news aggregator accounts. In contrast, the small groups identified by node colour in Figure 3.12b, are clearly focused on sharing a single particular URL, raising the question of whether they are, in fact, all in one coordinating group.

As alluded to above in Section 3.2.1.2 on complex graphs, further insights can be gained by combining nodes of different types in the one network, as can be seen in Figure 3.11. By introducing nodes to represent the evidence of coordination recorded between accounts (i.e., the *reasons* why they are thought to be coordinating), we can form 2-layer account/reason networks, where accounts link to each other as well as the reasons why they are connected. In the example given, we can see that each pair of accounts only uses a URL once each, indicating a high degree of variation in the content they are sharing. If, instead, the network showed two accounts strongly linked (i.e., with an edge with a heavy weight), but then they only connected to a single shared URL, then it would tell us the two accounts are frequently posting that same URL, so not only are they trying to boost a URL, but they are both attempting to boost the *same* URL and thus may be working together. These considerations can guide our analysis of the examples provided in Figure 3.12.

3.2.2.5 Co-hashtag *account* networks

The *co-hashtag account network* is a coordination account network with a content focus, as a fundamental element in the data pre-processing stage is the selection of the hashtags. Typically, a collection has so many hashtags that the resulting co-hashtag network would be so large and dense as to be inaccessible to meaningful interpretation. Starting with specific seed hashtags used by target communities (e.g., partisan hashtags), a co-hashtag network created based on only the seed hashtags,

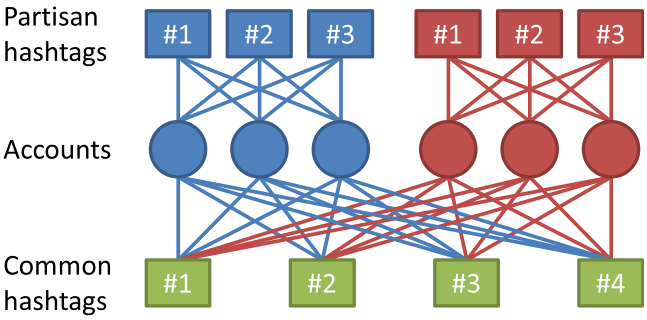


Figure 3.13. A bipartite network of two communities of accounts (blue and red circles) and hashtags (rectangles) linked when accounts use a hashtag, which demonstrates how the common non-partisan hashtags (in green) dominate the uses, but the less frequently used partisan hashtags (in red and blue) clearly delineate the red and blue communities.

hashtags that co-occur with them, and without high frequency hashtags, can provide insight into the groups using those hashtags and their discussions.

The co-hashtag network, in this context, consists of nodes representing accounts linked with undirected weighted edges when the accounts use the same hashtag, regardless of the timeframe. The edge weights are the sums of the product of the number of uses each account made of a given hashtag, for each hashtag they both used. So, for example, if accounts $\{v_i, v_j \in V\}$ use a set of common hashtags, $\{h_1, h_2, \ldots h_k \in H\}$, we create an undirected edge $\{v_i, v_j\}$. If $h_i^{v_i}$ indicates how often user v_i used hashtag h_i, the weight of the new edge is the given by

$$\omega(v_i, v_j) = \sum_{i=0}^{k} h_i^{v_i} \cdot h_i^{v_j}. \tag{3.12}$$

Others (e.g., Magelinski et al., 2021) use the minimum of v_i and v_j's usages of each hashtag, but their aim was to reduce computational overheads, but targeted use with small datasets can mean this consideration can be avoided. Instead, our weight calculation emphasises links from quiet (i.e., those with a small number of uses of a hashtag) accounts to loud accounts (i.e., ones with many uses), highlighting links that might otherwise be obscured or filtered out.

Some hashtags appear frequently in social media datasets, especially ones used as query terms to create the dataset in the first place (in which case it may appear in every single post). Creating a co-hashtag network using such popular hashtags will result in a very dense network in which many edges may lack any significant meaning, as they refer back to the query hashtags. Instead, we can examine the distribution of hashtag use in a dataset and remove the most widely used hashtags. Doing this removes terms discussed widely in the dataset, but serves to reveal the more community-specific discussion topics that would otherwise be obscured.

Further meaningful filtering can be employed by considering the content of the hashtags themselves; in political datasets, partisan hashtags are usually indicative of (1) an opinion on an issue (2) that potentially creates an axis of polarisation depending on how strongly it divides accounts, and (3) an association with one of the polarised groups. The example in Figure 3.13 shows how the popular common hashtags (in green) dominate the use counts (each is used six times) but tell us nothing of communities, but the use of partisan hashtags (in red and blue), each of which is only used three times, neatly describe the red and blue communities. As such, focusing on popular partisan hashtags gives us an opportunity to identify further polarised communities in a dataset. If no relevant partisan hashtags occur in the dataset (e.g., in our AFL dataset), then it is still possible to use a hashtag co-use network as a confirmation tool by using *faux* partisan hashtags based on the partisan groups discovered by other means: we choose our set of faux hashtags as those most used by the accounts in the known partisan groups, but that are also unique to each group. Of course, these faux partisan hashtags will be likely appear in posts (e.g., tweets) alongside other hashtags not unique to the partisan groups (i.e., they are used by members of the broader network). Despite this, the faux hashtags will still form a strong basis to judge whether the polarisation found in other, e.g., interaction networks also appears in the co-hashtag network.

3.3 Comparison and Characterisation

In many circumstances, the characterisation of a dataset requires comparison with another dataset, rather than examining it in isolation. Comparison methods are useful for examining datasets that are expected to be the same if not identical (Part I) and for comparing the behaviour and structure of polarised communities (Part II).

3.3.1 Contrasting datasets

Where possible, it is valuable to compare results against a ground truth dataset. This is necessary for supervised machine learning systems (as discussed below in Section 3.4), which are trained on the basis of labelled data. In the context of social media datasets, ground truth is often hard to obtain, partially due to limitations on exchanging datasets (discussed in Section I.1) but also because of the difficulty in defining the kinds of information sought in social media data. To label an account a bot, for example, the best way to be sure is to have written the bot in the first place, as there is significant overlap between some repetitive human online behaviour and that of genuine automated accounts (McKew, 2018; Bellutta et al., 2021). Concepts such as trolls and information campaigns are even more difficult to strictly define and typically are only meaningful in particular contexts.

That said, there are contexts in which ground truth data can be found, and examples often involve the use of activity of known accounts as exemplars of behaviours. Keller et al. (2017) and Keller et al. (2019) made use of court records to identify the accounts of South Korean secret service employees, used to influence the 2012 national elections at the direction of the incumbent president. Vargas et al. (2020) make use of four 'baseline' datasets, two political (from the US and UK) and two non-political (from academics and a random selection of accounts), in their exploration of detecting foreign influence campaigns. Some OSNs provide datasets of accounts they have banned, such as those available through Twitter's "Transparency Report",[7] last updated in December 2021. These datasets, though often rich, are missing the surrounding social media discussion of which they are a part and which they have the potential to influence.

The value of using a ground truth dataset is that it can confirm that a detection method works with a dataset in which the desired behaviour is known to exist. Then, if that the same kind of behaviour is detectable in a second, non-ground truth dataset, where it was not certain to exist, it suggests that the behaviour in question is detectable and engaged in by others. This provides confidence in the detection method.

A further useful dataset comparison tool is that of random datasets. Cao et al. (2015) demonstrates the value of this approach to confirm that the URL sharing behaviour identified by their classifiers was clearly distinct from random aggregated behaviour. Though this may be a coarse method of comparison, there are circumstances where it can be valuable as a simple check.

3.3.2 Comparing distributions

Binomial tests are used to compare distributions of values to determine if they are different to a statistically significant degree. Two sets of numbers may differ, but their distributions may be similar enough as to be indistinguishable. By choosing the *null hypothesis* that they are indistinguishable, a binomial test can be applied to determine if there is sufficient evidence to reject that hypothesis, and the degree of confidence in that conclusion, measured as a p-value. The p-value indicates the statistical likelihood that the two distributions provided are, in fact, the same. The smaller the p-value, the more unlikely it is that two sets of values have the same distribution. For this reason, it is common to consider p-value thresholds of 0.05, 0.01, 0.001, and 0.0001 to express the degree of confidence in rejecting the null hypothesis.

Another measure for comparing distributions, ones with repeated values, is *entropy*, a diversity measure. Cao et al. (2015) used entropy a measure of the diversity of URLs posted by groups of social media accounts. If a group posted the same URL repeatedly, it resulted in a very low entropy value. Similarly, a set of values which are all the same has zero entropy. In this way, it can be hypothesised that groups engaging in coordinated behaviour may use the same hashtags and URLs and retweet the same tweets than organic groups, and thus their entropy should be lower, whereas organic value should vary more and thus have higher entropy.

[7] https://transparency.twitter.com/en/reports/information-operations.html. Accessed 2021-12-15.

3.3.3 Comparing rankings

If, instead, the sets of numbers are ordered (i.e., are ranked lists), then we can compare the similarity of their rankings with Kendall's τ and Spearman's ρ coefficients. Both provide a value in the range $[-1, 1]$, with 1 indicating a perfect match in rankings and -1 indicating that one ranking is the exact reverse of the other. To classify the strength of the correlations, we follow the guidance of Dancey and Reidy (p.175, 2011), who posit that a coefficient of $0.0 - 0.1$ is uncorrelated, $0.11 - 0.4$ is weak, $0.41 - 0.7$ is moderate, $0.71 - 0.90$ is strong, and $0.91 - 1.0$ is perfect.

A visual method for comparing rankings is to create scatter plots of the elements common to each ranked list, in which the (x, y) position on the plot is determined by its rank in the first and second lists. We restrict ourselves to only using common elements, because there is no meaningful position to assign an element that only appears in one of the lists.

3.3.4 Comparing clusters

Cluster comparison is used to examine the results of different community detection methods or variations in parameter choices. Although comparing the number of nodes and edges in clusters as subgraphs will provide a degree of insight, a simple initial analysis is to consider the clusters as sets of values and then use measures associated with set membership, such as the Jaccard similarity and overlap coefficients (Verma and Aggarwal, 2020). In Part III, we use these measures to compare the groups of accounts identified as coordinating and render the results as heatmaps. The Jaccard similarity coefficient of two sets of items, X and Y, is:

$$J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} = \frac{|X \cap Y|}{|X| + |Y| - |X \cap Y|}. \tag{3.13}$$

If there is significant imbalance in the sizes of X and Y, then their similarity may be low, even if one is a subset of the other. An alternative measure, the Overlap coefficient (also known as the Szymkiewicz–Simpson coefficient, Verma and Aggarwal, 2020), takes this imbalance into account by using the size of the smaller of the two sets as the denominator:

$$overlap(X, Y) = \frac{|X \cap Y|}{min(|X|, |Y|)}. \tag{3.14}$$

The Jaccard and overlap coefficients can be used to quickly understand two facts about the sets of accounts:

- *Is one set a subset of the other?* If so, the overlap coefficient will reach 1.0, while the Jaccard coefficient will not if the two sets differ in size. If they are disjoint, the overlap coefficient will be 0.0 along with the Jaccard coefficient.

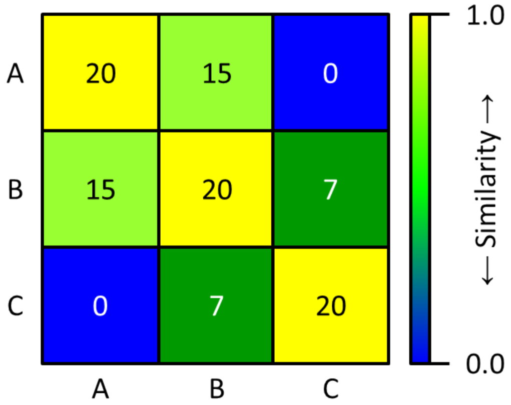


FIGURE 3.14. A similarity heatmap example, showing the common members of each pairing of three sets of items. Sets A, B, and C have 20 items each. Each cell represents the pairwise comparison of the labelled sets, hence the diagonal shows the maximum raw number (20 identical members) and is filled with the colour of the maximum similarity value (yellow). A scale to the right indicates how the cell colour relates to the similarity value.

- *Do the sets differ in size?* If the sets are different sizes, but one is a subset of the other, the overlap coefficient will hide this fact, while the Jaccard coefficient will expose it. If both coefficients have values close to 0.0, then the sets are clearly different in membership and potentially also in size. If the coefficient values are very close, then the sets are close in size, because the denominators are similar in size, meaning $|X \cup Y| \approx min(|X|, |Y|)$, but this will only occur if they share many members (i.e., $|X \cap Y|$ is high).

The heatmap representation presents pairwise comparison of the sets in question in a visually accessible manner. In the example shown in Figure 3.14, we can see that sets A and B share 15 members, and so the AB and BA cells are coloured bright green. Sets B and C only share 7 members, and so cells BC and CB are coloured a deeper green, while sets A and C have no members in common, and so their corresponding cells are blue. The diagonal represents each set compared with itself, and so its similarity value (whether it be the Jaccard or overlap coefficient) is 1.0, and so the cell is coloured yellow. Each cell includes the raw number of common members, to better inform the reader of the overall influence of the variations between sets. If each set is the result of progressively varying a particular parameter, it is possible to see the effect the degree of variation has.

Finally, a statistical measure to compare set membership en masse is the Adjusted Rand Index (Hubert and Arabie, 1985). This considers two networks of the same nodes that have been partitioned into subsets. When considered in pairs, there are nodes that appear in the same subset in both partitions (a), and there are (many) pairs of nodes that do not appear in the same subsets in either partition (b), and the rest appear in the same subset in one of the partitions but not in the other. Defining the total of possible pairings of the n nodes ($\frac{n(n-1)}{2}$) as c, the Rand index, R, is simply

$R = \frac{a+b}{c}$. The Adjusted Rand Index (ARI) corrects for chance and provides a value in the range $[-1, 1]$ where 0 implies that the two partitions are random with respect to one another and 1 implies they are identical.

3.3.5 Comparing the text of posts

To consider how the content produced by the members of groups compares not just within the group, but also with the broader population, we can perform a pairwise examination of the text each member produces. This will be most successful when the type of behaviour being sought relies on repetition, e.g., co-retweeting or copypasta. If a group is boosting a message, it is reasonable to assume the content posted by the members of the group will be more similar internally than when compared externally (i.e., to the content of non-members). To analyse this internal consistency of content, we treat the text of each group member's posts as a single document and create a doc-term matrix using 5-character n-grams for terms to maintain phrase ordering (which is lost with bag-of-word approaches). Comparing the members' document vectors using cosine similarity in a pairwise fashion creates a $n \cdot n$ matrix where n is the number of accounts in the coordination network. This approach was chosen for its performance with non-English corpora (Damashek, 1995), and because using individual tweets as documents produced too sparse a matrix in a number of tests we conducted. The pairwise account similarity matrix can be visualised as a heatmap, using a spectrum of colours to represent similarity. By ordering the accounts on both the x and y axes to ensure they are grouped together, if our hypothesis is correct that similarity within groups is higher than outside, then we should observe clear bright squares representing entire groups along the diagonal of the resulting similarity matrix. The diagonal itself will be the brightest because it represents each account's similarity with itself.

If groups contribute few posts, which are similar or identical to other groups, then bright squares may appear off the diagonal, and this would be evidence similar to clusters of account nodes around a small number of reason nodes in the 2-layer account/reason networks mentioned above in Section 3.2.2.4, as illustrated in Figure 3.12b.

This method offers no indication of how active each group or group member is, nor evidence of similar posting times, so displays of high similarity may imply low levels of coincidental activity as well as high content similarity. This is just because of the lower likelihood that highly active accounts are going to be highly similar in content (by contributing more posts, there are simply more opportunities for accounts' content to diverge). The use of the 5-character n-gram approach is designed to offset this because each tweet in common between two accounts will yield a large number of points (n-grams) of similarity, as will the case when the same two tweets are posted in the same order (i.e., two accounts both post tweet t_1 and then t_2), because the overlap between the tweets will yield at least four points of similarity (those being the n-grams across last four characters of t_1 and the first four of t_2).

3.3.6 Characterising group member connectivity

Groups that repost (e.g., retweeting on Twitter) or mention themselves create direct connections between their members, meaning if one is discovered, it may be trivial to find its collaborators. To be more inconspicuous (i.e., covert), therefore, it would be sensible to have a low *internal repost* and *mention ratios* (IRR and IMR, respectively). This concept is closely related to the homophily analyses mentioned in Section 3.2.1.6. Formally, if RT_{int} and M_{int} are the the sets of reposts and mentions of accounts within a group, respectively, and RT_{ext} and M_{ext} are the corresponding sets of reposts and mentions of accounts outside the group, then, for a single group

$$IRR = \frac{|RT_{int}|}{|RT_{int}| + |RT_{ext}|} \tag{3.15}$$

and

$$IMR = \frac{|M_{int}|}{|M_{int}| + |M_{ext}|}. \tag{3.16}$$

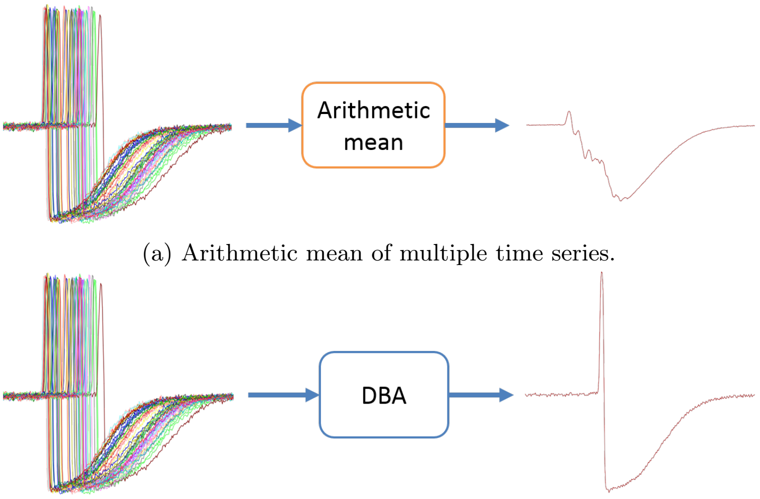
3.3.7 Temporal patterns in group posting behaviour

Online campaigns can exhibit different temporal patterns depending on their type. Temporal averaging techniques, such as the *dynamic time warping barycenter averaging* (DBA) method (Petitjean et al., 2011), can highlight this. In Part III, we use this technique to compare the daily activities of groups of accounts in ground truth and random datasets with those in the test datasets. The temporal averaging technique produces a single time series made by combining each account's activity time series (e.g., daily activities). The common approach of simply calculating the arithmetic mean of each account's activity at each time point (e.g., tweets per day) can result in sub-optimal results if the accounts' behaviour is off-phase (i.e., when accounts are active on different days). DBA avoids averaging out time series that are off-phase from one another by first aligning them before averaging them. The results of applying arithmetic mean and DBA techniques to a set of time series is shown in Figure 3.15.

Another aspect of temporal analysis is the comparison of each group's activity at different times in the datasets, including specifically exploring whether group members' timelines match and what the implications are for the behaviour of members whose activity aligns. It may reveal groups that should be merged or split. This is non-trivial for any moderately large dataset, but examination of the ground truth can provide insight into the behaviours exhibited by known collaborators.

3.4 Classification via Machine Learning

The primary benefit of machine learning (ML) techniques is that they are data-driven, meaning that the system itself builds statistical models to discriminate between samples with which it is presented. In other words, ML is "the science (and art) of



(a) Arithmetic mean of multiple time series.

(b) Applying DBA to multiple time series.

Figure 3.15. Temporal averaging of time series using the arithmetic mean and DBA. Open source images sourced from https://github.com/fpetitjean/DBA, under the GPLv3 licence.

programming computers so they can *learn from data*" (p.4, Géron, 2019). Only a brief introduction to ML concepts is provided here, as it is only used as one of many validation measures in Chapter 7, and our interest lies only with classifiers.

Domingos (2012) discusses classifiers specifically, clarifying the importance of algorithm choice, training set design, and selected evaluation measure. Though some ML algorithms are *unsupervised*, meaning they require no training and operate on only the data provided, classifiers are typically *supervised*, trained on a selection of labelled instances and then evaluated on how well they predict the labels (or *classes*) of test instances.

3.4.1 One class classifiers

Many classifiers are binary, able to distinguish instances between two classes, e.g., an image classifier may be trained to distinguish between photos of dogs and cats based on examples of the two. A specialisation of interest is one-class classifiers, able to find instances of a desired class in data that includes things that simply are and are not of the desired class. They can be used for anomaly detection (p.274, Géron, 2019) or information retrieval (e.g., gene ranking, Mordelet and Vert, 2014). An intuitive example is a library's book recommendation system. Based on a person's borrowing history, it has examples of the books the person likes (assuming they have liked everything they have borrowed), but no information about what do not like. The recommender's task is then to find other books in the library that are similar to those borrowed. If the person could list which books they had not enjoyed, that could form a negative training set to contrast with the positive one of books borrowed, but a one-class classifier, sometimes referred to as a positive-unlabelled, or PU, classifier, is intended to work without a clear negative training set (Mordelet and Vert, 2014).

3.4.2 Performance measures

A classifier's performance metrics include its accuracy, F_1 scores for each class, and the *precision* and *recall* measures that the F_1 scores are based upon. High precision implies the classifier is good at recognising samples correctly, and high recall implies that a classifier does not miss instances of the class they are trained on in any testing data. For example, a good apple classifier will successfully recognise an apple when presented with one, and when presented with a bowl of fruit, the classifier will successfully find all the apples in the bowl. The F_1 score combines these two measures:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{3.17}$$

and provides insight into to the balance between the classifier's precision and recall. The accuracy of a classifier is the proportion of instances in a test data set that the classifier labelled correctly. In this way, the accuracy is the most coarse of these measures, because it offers little understanding of whether the classifier is missing instances it should find (false negatives) or labelling non-matching instances incorrectly (false positives). It is particularly vulnerable when using *skewed* datasets, where only a few positive examples of a given class `X` exist, as every not-`X` response will be regarded as correct. The F_1 score begins to address this failing, but direct examination of the precision and recall provides the most insight into each classifier's performance.

Further discussion of performance measures can be found in (p.90, Géron, 2019).

3.5 Bot Analysis

Although coordinated behaviour in online campaigns is often conducted without automation (Starbird et al., 2019; Nimmo et al., 2020), automation is still commonly present in campaigns, especially in the form of social bots, which aim to present themselves as typical human users (Ferrara et al., 2016; Grimme et al., 2017; Cresci, 2020). The malicious use of such automation was discussed in Section 2.2.1. In this context, automation detection is a useful supporting tool for exposing teams of cooperating bot and social bot accounts. Detecting coordination amongst accounts engaging in automation simply bolsters the conclusion that they are being used for a particular purpose, whether that purpose is, for example, to promote a narrative or aggregate news headlines. News aggregators are unlikely to hide their identity, however.

We use the Botometer (Davis et al., 2016) service to evaluate selected accounts for bot-like behaviour. As discussed in Section 2.2.1, Botometer's primary summary measure is the Complete Automation Probability (CAP), provided as a value in $[0, 1]$ in two variants: one for predominantly English-speaking accounts and one language-agnostic. Other studies have relied on a CAP of 0.5 as a threshold for labelling an account as a bot, but there is a significant overlap between humans that act in a very bot-like manner and bots that are quite human-like, so we adopt the practice of Rizoiu et al. (2018) and regard scores below 0.2 to be human and above 0.6 to be bots.

Part I

The Information Environment: Social Media

The study of coordinated behaviour online could apply to any cooperative activity making use of internet-based communication. This could range from packet-based distributed denial of service (DDoS) attacks[8] on web sites and services to multi-year information operations conducted with highly curated false personas and fake news organisations (e.g., Secondary Infektion, Nimmo et al., 2020). Due to its increasing importance and use in daily life, social media is our focus, with the primary target being the groups of social media accounts engaging in computational propaganda and other attempts at inauthentic influence. Thus our *information environment* is not the internet at large or data obtained from web sites (e.g., blogs and media sites), but the data provided by the social media platforms regarding their users' activities.

In this Part, we focus on **TRQ1**, relating to trust in the data we obtain from social media platforms and in the results of analyses of those data.

Relying on OSN data for analysis introduces at least two specific challenges, both of which affect trust in different ways: availability and reliability. We address how the two relate to each other in brief below, including contributions made specifically to the question of benchmarking, the foundation of algorithmic comparison, of social media analytics. The issue of reliability relates to platform transparency and refers to the fact that it is unclear whether the data provided by OSNs is complete, the reasons to assume it would be incomplete, and consequences for social media research in general.

Beyond the discipline-level questions of trust in data, we then turn, in Chapter 4, to the effects that variations in data obtained from Twitter have upon the results of SNA. To explore these effects, we establish a systematic methodology and validate it through several case studies.

I.1 On the Importance of Open Data

Availability of transparent, trustworthy data is vital for any research, but is particularly important for social media analytics, the complexities of which were addressed in section 2.3. The significance for OSN research is the fact that the raw social media data it relies upon is encumbered. Once collected, it remains owned by the OSN it came from and use of it is permitted only under the OSN's specific and unique terms and conditions (T&Cs, Bruns, 2019a). Availability affects benchmarking, which facilitates the fair comparison of algorithm performance. Access to fixed (i.e., immutable) and widely available datasets is a core element of this practice. The second aspect, reliability, arises when retrieving data from information systems: it is important to know that the same data will be provided when the same query or filter criteria are applied, i.e., that the data provided by any data gathering mechanism (including APIs, specifically) can, in fact, be relied upon. Reliability, in this way, affects repeatability,

[8] "A distributed denial of service (DDoS) attack is an attempt to make an online service unavailable by overwhelming it with traffic." Source: https://www.cyber.gov.au/acsc/view-all-content/threats/denial-service. Posted 2020-05-22. Accessed 2022-02-01.

and benchmarking requires that experiments on data be repeatable. Trust will be impacted, however, even if the data are reliable, if there is any question of whether the data are complete. In other words, has the OSN provided *all* the data relevant to a query.

I.1.1 Data for benchmarking

In Publication **IV**, we have argued there is currently a 'crisis' in benchmarking of social media analytics, similar to other 'replication crises' observed in other areas of science (e.g., Baker, 2016; Amrhein et al., 2017; Cockburn et al., 2020). We explored the issue, highlighting how OSN T&Cs limit the extent to which researchers can share their datasets. Benchmarking is the practice of using the same datasets and execution environments (including hardware and software configurations) to run algorithms or analytics (implementations of algorithms), thus providing a common basis on which to compare their performance. The comparison can include execution time and resource usage (e.g., disk and memory) in addition to the correctness of computed results. By ensuring fair comparisons, benchmarking engenders trust in the results, the algorithms and systems that produce them, and in the skills of those who devise them. The 'replication crisis', namely that published results have been difficult to reproduce (Baker, 2016), is particularly important in the field of social media analytics due to the constrained availability of data imposed by OSN policies (Bruns, 2019a; Assenmacher et al., 2021).

The primary issue with social media datasets as a source for benchmarking is how OSN T&Cs constrain their distribution (Bruns, 2019a; Freelon, 2019). For example, Twitter's conditions require that, except under certain conditions (including when datasets are very small), only the IDs of tweets can be shared.[9] Using the tweet ID, the data for the tweet can then be retrieved ('rehydrated' is the preferred term), but only if it is still available. Tweets that were valid at the time of collection may no longer be available in the future for a number of reasons: the tweet may have been deleted by the account which posted it; the account may have been set to be 'private' or 'protected', thereby hiding its tweets from the general public; or the account may have been suspended or deleted, making its tweets inaccessible in the process. The account may be reinstated or made public again, in which case the tweet may return, but there is no guarantee.

The effect of this is that a later researcher may not be able to faithfully reconstitute a dataset, and thus will be unable to compare their algorithm fairly with the original work. One way to ensure fair comparisons is for the second researcher to run the first's algorithm (i.e., an implementation thereof) on a dataset collected by the second

[9] Twitter's terms permit the sharing of up to only 50,000 hydrated tweets per day per individual recipient, or 1.5m tweet IDs in a 30-day period, so only on-request sharing is feasible. Hydrated tweet datasets (even those with fewer than 50,000 tweets) should not be posted publicly (e.g., on a blog or public GitHub repository). Source https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases. Accessed 2022-01-06.

researcher. This also may not be possible, as an implementation may not be available: it may be embargoed for commercial reasons or may not have been sufficiently well described to be re-implemented (re-implementing it would also be a waste of the second researcher's time). It is unclear what T&Cs Twitter bot datasets (e.g., Feng et al., 2021) are released under, but Dodds (2017)'s analysis of Twitter's terms in mid-2017 suggested it can only be by specific arrangement (often a commercial one). Twitter has introduced new research-specific terms since then, but they still do not clearly permit publication of arbitrary datasets of complete (i.e., fully hydrated) tweets or other Twitter data, such as user profile information.

I.1.2 Trust in data

The second aspect relates to variations in and the completeness of the data returned by OSN APIs (Bruns, 2019a). Some OSN APIs may not produce the same results when used multiple times, or by different clients simultaneously, such as Twitter's 1% Sample API (Joseph et al., 2014; Paik and Lin, 2015). Further, similar to Twitter's Sample API, it is unclear whether any filter- or search-based OSN API will provide complete results, but instead may provide a sample of the 'most relevant' results, as determined by the OSN. The sampling methods used are rarely made public (Morstatter et al., 2013), but it is reasonable to assume they are designed to maintain engagement and maximise profitability (given the OSNs are all run by for-profit companies). Other motivations emerge at the geopolitical level, where concerns have been raised regarding improper governmental influence from countries in which OSN-owning companies are based.[10] To account for the commercial interests, as mentioned above, many OSNs provide different levels of access to their data, usually providing greater access at a price,[11] but use of these raises questions of transparency (Ruths and Pfeffer, 2014).

The result is that researchers will gravitate towards OSNs whose data are most accessible, such as Twitter's, causing a platform bias in social media research (Persily and Tucker, 2020). Some OSNs are therefore over-represented in the literature. We observed that this will cause results to suffer from the 'Streetlight Effect', in which someone who has lost their keys searches for them under the streetlight where they can see, rather than nearer to wherever they dropped them. Care must be taken to ensure that algorithms designed using, e.g., Twitter's data model, can still be applied to other relevant platforms by abstracting out and only relying on common data model elements.

[10] E.g., government interference was suspected when the Australian Prime Minister's WeChat account was suddenly overtaken and rebranded. WeChat is owned and hosted within China. The Chinese Communist Party (CCP) denied the accusation. Source: https://thenewdaily.com.au/news/politics/australian-politics/2022/01/24/scott-morrison-we-chat-hack/. Posted 2022-01-24. Accessed 2022-01-24.

[11] E.g., Twitter's enterprise API: https://developer.twitter.com/en/products/twitter-api/enterprise. Accessed 2021-11-24.

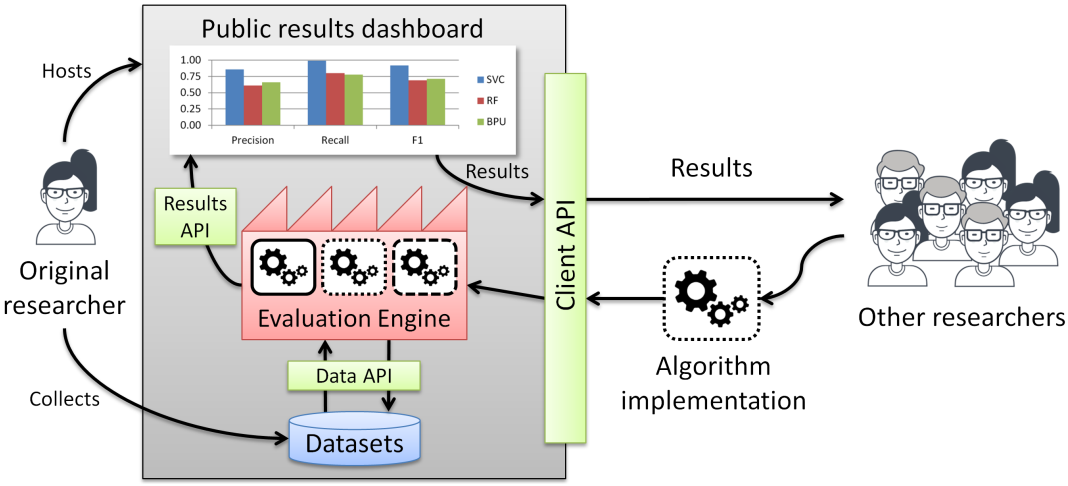


Figure 3.16. A conceptual interpretation of the benchmarking framework proposed in Publication IV. The original researcher stores datasets they have collected, making them available to a standardised analytic execution environment ("Evaluation Engine") via a fixed but Data API. Analytics are submitted by other researchers via a Client API, wrapped in self-contained deployment images (using, e.g., docker). These images are run in the Evaluation Engine and the results are published to a publicly visible dashboard via the Results API, and back to the submitting researcher via a callback in the Client API. To prevent results being displayed on the public dashboard, the submitting researcher can flag them as private or not conforming to the dashboard's preferred data model, but all results will be returned to the submitting researcher.

I.1.3 Towards a solution

To address these issues, we proposed a software framework and an exemplar implementation, which enable a researcher with a dataset to provide a stable execution environment to secondary researchers. This framework is presented in Figure 3.16. The framework ensures the dataset remains within the possession of the original researcher (who has permission to hold it), while allowing secondary researchers to run their analytics against it and obtain results for comparison with the analytic(s) of the original researcher(s). In fact, secondary researchers may run completely different analytics against the datasets, analysing the data in different ways, without ever needing to hold the raw data. This approach is limited in that it necessarily constrains the execution environment and how the data can be analysed, and leaves the cost of the data and computation hosting to the original researcher, but it was intended to be a starting point for future investigation.

The remainder of this Part presents a process for systematic exploration of variations in data provided by OSN streaming APIs, specifically with regard to their effects on SNA. The process is demonstrated through several Twitter-based case studies with systematically varied starting conditions, and the implications of the variations are discussed in detail.

Chapter 4

Variations in Social Media Data and SNA

To study the effects of Online Social Network (OSN) activity on real-world offline events, researchers need access to OSN data, the reliability of which has particular implications for social network analysis. This relates not only to the completeness of any collected dataset, but also to constructing meaningful social and information networks from them. In this multidisciplinary study, we consider the question of constructing traditional social networks from OSN data and then present several measurement case studies showing how variations in collected OSN data affect social network analyses. To this end, we developed a systematic comparison methodology, which we applied to five pairs of parallel datasets collected from Twitter in four case studies. We found considerable differences in several of the datasets collected with different tools and that these variations significantly alter the results of subsequent analyses.

Our results lead to a set of guidelines for researchers planning to collect online data streams to infer social networks.

*The content of this chapter was originally published in Publication **III** and expanded in Publication **V**.*

4.1 Introduction

Online activities can be associated with dramatic offline effects, such as voter fraud misinformation contributing to the 6 January 2021 riots and invasion of the US Capitol building in Washington DC (Scott, 2021), COVID-19 misinformation leading to panic buying of toilet paper (Yap, 2020), online narratives incorrectly attributing Australia's "Black Summer" bushfires to arson amplifying public attention to it via the media (see Chapter 5), and attempts to influence domestic and foreign politics (Ratkiewicz et al., 2011; Woolley, 2016; Morstatter et al., 2018; Woolley and Howard, 2018). For researchers to successfully analyse online activity and provide advice about protection from such events, they must be able to reliably analyse data from online social networks (OSNs).

As we discussed in Chapter 2, Social Network Analysis (SNA) facilitates exploration of social behaviours and processes. OSNs are often considered convenient proxies for offline social networks, because they seem to offer a wide range of data on a broad spectrum of individuals, their expressed opinions and inter-relationships. It is assumed that the social networks present on OSNs can inform the study of information dissemination and opinion formation, contributing to an understanding of offline community attitudes. Though such claims are prevalent in the social media literature, there are serious questions about their validity due to an absence of SNA theory on online behaviour, the mapping between online and offline phenomena, and the repeatability of such studies. Many of these issues were introduced in Section 2.3. In particular, the issue of reliable data collection is fundamental. Collection of OSN data is often prone to inaccurate boundary specifications due to sampling issues, collection methodology choices, as well as platform constraints. The establishment of datasets in which the research community can have confidence, as well as the ability for the replication of studies, including through common benchmarks, is vital for the validation of research findings. We discussed the importance of this, and recent contributions to it, in the introduction to this Part.

Previous work has considered the question of data reliability from a variety of perspectives. Broadly speaking, questions of how to reason about data quality appeared in the late 1960's in statistics but were picked up by management research in the 1980's and computer science in the 1990's as part of database and data warehouse research (Scannapieco et al., 2005). The dimensions described by Scannapieco et al. (2005) provide a structured way to reason about data quality in terms of accuracy, completeness, time-related measures and consistency. It is increasingly apparent that data heavy disciplines, such as ML, cannot rely on their techniques and a simple abundance of data to overcome these issues (Roccetti et al., 2020). Even if data is available, some ML techniques can still struggle if its distribution is uneven (Sun et al., 2009) and the 'cleanliness' of data can be a significant factor in the performance of ML systems (Breck et al., 2019; Roccetti et al., 2020). Data quality is also especially important for modern *Big Data* systems (Emani et al., 2015), including those underpinning OSNs, but those using OSN Application Programming Interfaces (APIs) can be assured of high quality data, at least with regard to the completeness of the schemas and validity of the values they provide.

Turning to OSN data specifically, relevant research into reliability has explored sampling (Morstatter et al., 2013; González-Bailón et al., 2014; Joseph et al., 2014; Paik and Lin, 2015), biases (Ruths and Pfeffer, 2014; Tromble et al., 2017; Pfeffer et al., 2018; Olteanu et al., 2019) and the danger of making invalid generalisations while relying on the promise of *Big Data* without first developing a nuanced understanding of the data (Lazer et al., 2014; Tufekci, 2014; Falzon et al., 2017; Venturini et al., 2019). Analyses of incomplete networks exist (Holzmann et al., 2018), but this paper specifically considers the questions of data reliability for SNA, considering not only the significance of online interactions to discover meaningful social networks, but also how sampling and boundary issues can complicate analyses of the networks constructed. Through an exploration of modelling and collection issues, and a measurement study examining the reliability of simultaneously collected, or *parallel*, datasets, this multi-disciplinary study addresses the following research questions:

- *To what extent do datasets obtained with social media collection tools differ, even when the tools are configured with the same search settings?*
- *How do variations in collections affect the results of social network analyses?*

Our work in this chapter makes the following contributions:

1. Discussion of the challenges mapping OSN data to meaningful social and information networks;
2. A methodology for systematic dataset comparison;
3. Recommendations for the use and evaluation of social media collection tools; and
4. Five original social media datasets collected in parallel, and relevant analysis code.

Five sections follow from this point: Section 4.2 briefly recaps the concepts of SNA and the challenges involved in obtaining and modelling OSN data for SNA purposes; Section 4.3 describes our methodology for systematic parallel dataset comparison; Section 4.4 presents results from using our methodology in a number of case studies; Section 4.5 discusses our findings and provides an exploration of the notion of a measure of reliability; and finally Section 4.6 offers recommendations for social media researchers and analysts, plus directions for future research.

4.2 Background

In Section 2.3, we introduced the concepts behind SNA and challenges involved in not just applying SNA to OSN data, but also with obtaining OSN data in the first place. Due to the relatively recent emergence of social media, social theory on how to build social networks from the constrained data provided by OSNs, and the effects on the meaning of analyses such as centrality and community detection, is still lacking. The only direct analogy to the traditionally long-standing relationships (e.g., familial, friendship, supervisory or collaboration, Borgatti et al., 2009) is follower links, which are cheap to create, easy to forget, unidirectional, and often computationally expensive to retrieve.

Instead, we use online interactions to build the relational links in social networks, as they provide evidence of direct connections between accounts at particular times. Further, they can offer insight into the strength of the connection based on the frequency, and also any direction of flow of information or influence. Compared with traditional SNA data collection methods, which often rely on interviewing subjects directly and then manually entering data, social media-based SNA can rely on highly structured and clean data from the OSNs' *Big Data* infrastructure, which is at least consistent and 'clean', avoiding some of Foidl and Felderer (2019)'s 'Data Smells'.

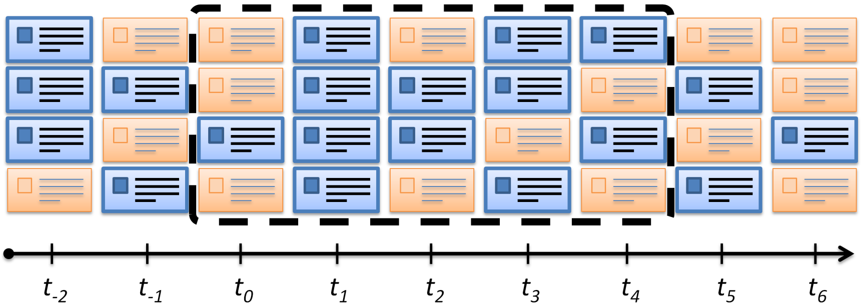


Figure 4.1. Given a stream of timestamped posts, our research question requires the blue posts. The boundary of a collection activity, defined by its filter criteria, such as filter terms, seed accounts, and when it starts and stops, is represented by the black dashed line, starting at time t_0. Target posts may be missed due to poor filter terms, starting or stopping too late or early, or due to OSN-imposed rate limits. Irrelevant posts may be captured due to filter term clashes, pollution from spammers, or language clashes (where a filter term is meaningful in non-target languages). Careful collection activity planning can address some of these concerns, but not necessarily all of them.

Careful consideration is required to determine how best to use timestamped interactions to build relations, however, and this will depend on the research question under consideration. *Is a single retweet enough to connect account A to B? What about three retweets? Do they need to be reciprocal? Should a mention or reply be treated differently?* We are necessarily limited to what the OSN provides us: we have no knowledge if a single person is using multiple accounts in the data under inspection, or knowledge of lurkers who closely observe and are influenced by specific other accounts, but never interact and never leave a digitial trace provided by the OSN.[1] Careful consideration of the research questions will also guide collection activities, as OSNs introduce specific complications to defining the network boundary. Figure 4.1 shows how only part of the data required to address a research question may be obtained, limited through factors such as missing filter terms, pollution from irrelevant content that match filter terms, and language clashes where a filter term may be meaningful in non-target languages. The data finally obtained may need cleaning and can only be regarded as a subset of the true desired dataset, and its representativeness is unclear.

So far, the following has been established:

[1] It is reasonable to assume that OSNs note which posts an account receives and has onscreen for what length of time, as this could easily guide personalised recommendation algorithms for the user.

- The OSN information selected and used to form ties in social networks requires careful consideration to ensure meaningfulness;
- Uncertainty regarding the completeness of OSN data (due to rates of access, accessibility of data models, query construction and OSN owner commercial or other priorities) must be accounted for; and
- Because OSNs maintain *Big Data* systems as infrastructure, researchers can rely on them to have carried out many tasks associated with data quality by the time they request data from the APIs (e.g., ensuring schema consistency and valid values)—these are tasks that other SNA researchers, such as those collecting data through direct community interaction, must do themselves.

We are now in a position to empirically examine more closely the issue of repeatability, by comparing simultaneously retrieved collections.

4.3 Methodology

Our initial hypothesis was that if the same collection strategies were used at the same time, then each OSN would provide the same data, regardless of the collection tool used. Consequently, social networks built from such data using the same methodology should be highly similar, in terms of both network- and node-level measurements. Our methodology, using techniques and measures defined in Sections 3.2.1 and 3.3, consisted of these steps:

1. Conduct simultaneous collections on an OSN using the same collection criteria with different tools.[2]
2. Compare statistics across datasets.
3. Construct sample social networks from the data collected and compare network-level statistics.
4. Compare the networks at the node level.
5. Compare the networks at the cluster level.

Examining the parallel datasets in each of these ways provides the opportunity for the analyst to develop a well-rounded understanding of the participants in an online discussion, their behaviour, how they relate to each other and the communities they form.

4.3.1 Scope

The scope of this chapter's work is limited to datasets obtained via streaming APIs filtered with keywords. Other collection styles may start with seed accounts, and collect their data and the data of accounts connected to them, either through interaction

[2] Different credentials are used to avoid any effects of account-based rate limiting.

(e.g., via comments, replies or mentions) or via follower links, as mentioned above. Such collections (especially follower networks) often require the collection of data that is prohibitive to obtain, is immediately out of date, and provides no real indication of strength of relationships (as discussed in Section 2.3.2). Additionally, in the absence of a domain-focused research question to inform the choice of seed accounts, no particular accounts would make sensible seeds, so here we rely on keyword-based collections.

4.3.2 Data collection

Twitter was chosen as the source OSN due to the availability of its data, the fact that the data it provides was thought to be highly regular (Joseph et al., 2014), and because it has similar interaction primitives to other major OSNs. Twitter is also widely used in academia for research that makes predictions, in particular predictions about population-level events, behavioural patterns and information flows, such as studies of predicting social unrest (Tuke et al., 2020) or misinformation (Wu et al., 2016). The validity of these predictions is fundamentally based on the consistency of the underlying (accessible) data. Two very different collection tools were chosen:

Twarc[3] Twarc is an open-source library which wraps Twitter's API, and provided the baseline for the study.

RAPID RAPID (Real-Time Analytics Platform for Interactive Data Mining, Lim et al., 2019) is a social media collection and data analysis platform for Twitter and Reddit. It enables filtering of OSN live streams, as well as dynamic *topic tracking*, meaning it can update filter criteria in real-time, adding terms popular in recent posts and removing unused ones.

Both tools facilitate filtering Twitter's Standard version 1.1 live stream[4] with keywords, providing datasets of tweets as JSON objects.

4.3.3 Constructing social networks

A social network is constructed from dyads of pairwise relations between nodes, which in our case are Twitter accounts. The node ties denote intermittent relations between accounts, inferred from observed interactions (Nasim, 2016; Borgatti et al., 2009). Like any choice of knowledge representation, different networks can be constructed to address different research questions. For example, a network to study information flow could draw an arc from node A to B if account B retweets A's tweet (implying B has read and perhaps agreed with A's tweet); alternatively, the same interaction could be used to draw an arc from B to A if the relation is to imply an attribution of status or influence (A has influence because B has supported it through a retweet). Networks can be constructed based on direct or inferred relations, including

[3] https://github.com/DocNow/twarc. Accessed 2022-01-10.

[4] https://developer.twitter.com/en/docs/twitter-api/v1/tweets/filter-realtime/overview. Accessed 2022-01-14.

retweeting, replying or mentioning, which we discuss below, or through the shared use of hashtags or URLs, reciprocation or minimum levels of interaction activity, or friend/follower connections. Morstatter et al. (2018) constructed networks of accounts based on retweets and mentions to discover communities active during the 2017 German election, valuing mentions and retweets equally to mean one account reacting to another. URL sharing behaviour is often studied in the detection and classification of spam and political campaigns (Cao et al., 2015; Wu et al., 2018; Giglietto et al., 2020b). Some require more complex calculation such as linking accounts through their participation in detected events (Nasim et al., 2018). Of course, applications for social network analysis exist outside the online sphere, e.g., in narrative analysis (Edwards et al., 2020), and require similar considerations with regard to network design. In the absence of clear alternative research questions, we will examine the social relationships implied by direct interactions and retweet networks (due to their frequency in the literature), and thus we will focus only on the three types of network construction described in Section 3.2.2.1: mention, reply and retweet networks.

4.3.4 Analyses

At this point, comparative analysis can be applied to the parallel tweet datasets, initially by examining OSN-specific features and then the mention, reply and retweet networks constructed from them. An overview of data comparison methods is provided in Section 3.3. When analysing these networks, it is relevant to note that SNA posits two important axioms on which most network measures are based: network structure affects collective outcomes; and positions within networks affect actor outcomes (Robins, 2015). Furthermore, we should expect minor differences in collections to be amplified in resulting social networks (Holzmann et al., 2018).

4.3.4.1 Dataset statistics

To compare the parallel datasets, we directly compare a number of features, their frequencies and several maximums, as listed in Section 3.3.1. Though the features specified are specific to Twitter data (e.g., number of retweets and frequency of most retweeted tweet), to the greater extent they have analogies on other major OSNs (as shown in Table 2.1), and if analogies are not available, the feature sets can be adjusted to the data provided by the OSN data under study accordingly.

Using these figures, we account for major discrepancies between the datasets, which can guide post-processing (e.g., spam filtering). Depending on the application domain, it may be appropriate to also consider comparing the distributions of particular features, rather than just their maximum values (*cf.*, the use of value distributions in bot classifier feature sets, described in Section 2.6.1).

4.3.4.2 Network statistics

A variety of graph-level statistics are introduced in Section 3.2.1.3. Applying these directly to the constructed social networks, the following statistics are used to assess their differences: number of nodes, edges, average degree, density, mean edge weight, component count and the size and diameter of the largest, Louvain (Blondel et al., 2008) cluster count and the size of the largest, reciprocity, transitivity, and maximum k-cores. These measures provide us with an understanding of the 'shape' of the networks in terms of how broad and dense they are and the strength of the connections within.

4.3.4.3 Centrality values

Centrality measures offer a way to consider the importance of individual nodes within a network, and are discussed in Section 3.2.1.4. The centrality measures considered here include: *degree* centrality, indicating how many other nodes one node is directly linked to; *betweenness* centrality, referring to the number of shortest paths between all pairs of nodes in the network that a node is on and thus to what degree the node is able to control information flowing between other nodes; *closeness* centrality, which provides a sense of how topologically close a node is to the other nodes in a network; and *eigenvector* centrality, which measures how connected a node is to other highly-connected nodes.

Only centrality measures for mention and reply networks are considered, as edges in retweet networks are not necessarily direct interactions (Ruths and Pfeffer, 2014).

Given the set of nodes in each corresponding pair of networks is not guaranteed to be identical, it is not possible to directly compare the centrality values of each node. Instead we make use of the rank comparison methods introduced in Section 3.3.3. This includes rank scatter plots of common nodes for visual interpretation, and the Kendall τ coefficient for statistical interpretation, with the Spearman's ρ coefficient used as a confirmation measure. Strength of statistical correlations is judged following the guidance of Dancey and Reidy (p.175, 2011): a coefficient of $0.0 - 0.1$ is uncorrelated, $0.11 - 0.4$ is weak, $0.41 - 0.7$ is moderate, $0.71 - 0.90$ is strong, and $0.91 - 1.0$ is perfect.

4.3.4.4 Cluster comparison

The final step is to consider the clusters discoverable in the mention, reply and retweet networks and compare their membership using methods introduced in Section 3.3.4. We first compare the distribution of the sizes of the twenty largest Louvain clusters (Blondel et al., 2008) visually, and then the ARI across all clusters.

Table 4.1. Summary of data collection conditions.

Case Study	Collection	Duration	Tool 1	Tool 2	Tool 3
1	Q&A Part 1	4 hours	Twarc	RAPID (topic tracking)	—
	Q&A Part 2	15 hours	Twarc	RAPID (topic tracking)	—
2	AFL1	3 days	Twarc	RAPID (no topic tracking)	—
3	AFL2	6 days	RAPID (no topic tracking)	RAPID (no topic tracking)	—
4	Election	1 day	Twarc	RAPID (topic tracking)	Tweepy

Table 4.2. Summary statistics for the datasets used in this chapter.

Collection	Dataset	Tweets	Accounts
Q&A Part 1	Twarc	27,389	7,057
	RAPID	15,930	4,970
	RAPID-E	17,675	5,547
Q&A Part 2	Twarc	15,490	5,799
	RAPID	11,719	4,708
	RAPID-E	23,583	8,854
AFL1	Twarc	44,470	16,821
	RAPID	21,799	11,573
AFL2	RAPID1	30,103	14,231
	RAPID2	30,115	14,232
Election Day	Twarc	39,297	10,860
	Tweepy	36,172	10,242
	RAPID	39,556	10,893
	RAPID-E	46,526	12,696

4.4 Evaluation Case Studies

Several case studies were conducted to evaluate the comparison methodology, the requirements for which developed progressively, each new case study's requirements informed by lessons from the previous. The collections used different tools to carry out the parallel collections. As mentioned, Twarc was employed as a baseline, while RAPID was used with topic tracking enabled and disabled, and the tool Tweepy[5] was used in only one case study as a second baseline. The first case study consisted of two parallel Twitter datasets relating to an Australian panel discussion television programme with a prominent online community (Q&A); the first datasets were collected over the running of the programme (4 hours) and the second covered the following day's discussion (15 hours), both employing RAPID's topic tracking feature to broaden the conversation. The second case study examined discussion surrounding the national Australian Rules Football competition (the Australian Football League, or AFL) over a longer period (3 days), without RAPID's topic tracking. The third also examined the same online sports discussion, but over a longer period again (6 days)

[5] Tweepy is another open source library which provides a thin wrapper around the TwitterAPI: https://github.com/tweepy/tweepy. Accessed 2022-01-29.

and only made use of RAPID without topic tracking. The final case study incorporated a third tool to act as a further baseline and covered a regional but large election day, during which a significant amount of activity was expected. These conditions are summarised in Table 4.1.

A summary of the corpora collectedis presented in Table 4.2. As noted above, when topic tracking was employed with RAPID, some of the tweets it collected did not contain any of the initial keywords. These datasets are given the label 'RAPID-E'. Prior to comparison with the corresponding Twarc datasets, the RAPID-E datasets were filtered to retain only tweets containing at least one of the original keywords. The AFL2 case study used RAPID with no topic tracking expansion with two sets of Twitter credentials simultaneously; in this case the datasets are labelled 'RAPID1' and 'RAPID2'. The third collection tool, Tweepy, was included in the Election Day case study to act as a second baseline.

4.4.1 Case study 1: Q&A, `#qanda` and the effect of *topic tracking*

Initially, to obtain a moderately active portion of activity, we collected data from Twitter's Standard live stream relevant to an Australian television panel show, Q&A, that invites its viewers to participate in the discussion live.[6] A particular broadcast in 2018 was chosen due to the expectation of high levels of activity given the planned discussion topic. As a result, the filter keywords used were 'qanda'[7] and two terms that identified a panel member (available on request). We collected two parallel datasets over two periods:

Q&A Part 1: Four hours starting 30 minutes before the hour-long programme, to allow for contributions from the country's major timezones; and

Q&A Part 2: From 6am to 9pm the following day, capturing further related online discussions.

Twarc acted as the baseline collection as it provides direct access to Twitter's API, while RAPID was configured to use topic tracking via *co-occurrence keyword expansion* (Lim et al., 2019), meaning it would progressively add keywords to the original set if they appeared sufficiently frequently (five times in ten minutes). Expanded datasets such as these are referred to as 'RAPID-E'; it was filtered back to just the tweets containing the original keywords and labelled 'RAPID' to enable fair comparison with the 'Twarc' dataset. We expected the moderate activity observed would not breach rate limits, and thus, RAPID should capture all tweets captured by Twarc. This was not the case.

[6] The Australian Broadcasting Commission's "Q&A" observes the hashtag **`#QandA`**, which Twitter treats as equivalent to `#qanda`.

[7] The '#' was omitted to catch mentions of '@qanda', the programme's Twitter account.

Table 4.3. Summary statistics for the Q&A Parts 1 and 2 datasets.

	Dataset	All Tweets	Unique Tweets		Retweets		All Accounts	Unique Accounts	
Q&A Part 1	Twarc	27,389	11,481	(41.9%)	14,191	(51.8%)	7,057	2,090	(29.6%)
(20:00-00:00)	RAPID	15,930	22	(0.1%)	8,744	(54.9%)	4,970	3	(0.1%)
	RAPID-E	17,675	1,767	(10.0%)	9,767	(55.3%)	5,547	527	(9.5%)
Q&A Part 2	Twarc	15,490	4,089	(26.4%)	10,988	(70.9%)	5,799	1,128	(19.5%)
(06:00-21:00)	RAPID	11,719	318	(2.7%)	8,051	(68.7%)	4,708	37	(0.8%)
	RAPID-E	23,583	12,180	(51.6%)	13,679	(58.0%)	8,854	4,007	(45.3%)

4.4.1.1 Comparison of collection statistics

The first striking difference between the datasets was the number of tweets collected and the effect on the number of contributors (Table 4.3). RAPID collected fewer tweets by fewer accounts, but the datasets were close to subsets of the Twarc datasets. Between 26 and 42% of the tweets collected by Twarc were missed by RAPID, but the proportion of retweets in each part is similar (52% and 55% for Part 1 and 69% and 71% for Part 2). In both parts, very few accounts appear in only the RAPID collections. Discussions with RAPID's developers revealed it dumps tweets that miss the filter terms from the textual parts of tweets (e.g., the body, the author's screen name and the author's profile description). The extra tweets RAPID collected were relevant and in English[8] (based on manual inspection) but posted by different accounts (unique to RAPID-E). Of the tweets that RAPID collected which contained the keywords, they were posted by almost the same accounts as Twarc, but simply did not contain the same tweets.

The benefit of topic tracking via keyword expansion is yet to be strongly evaluated, but this study indicates there are benefits (relevant tweets that omit the original filter terms are picked up once related terms are added) as well as costs (tweets that include the original filter terms but are not collected). RAPID's expansion strategies are modifiable to optimise data collection; however, we chose not to make use of this capability to prevent obscuring the current comparative study. The rest of this analysis explores how much of a difference the keyword expansion makes with regard to SNA.

Table 4.4 reveals that although feature counts vary significantly, many of the most common values are the same (e.g., most retweeted tweet, most mentioned account, most used hashtags). Many are approximately proportional to corpus size (Twarc is 1.72 and 1.32 times larger than RAPID for Parts 1 and 2, respectively), but with notable exceptions and no apparent pattern. Some values are remarkably similar, despite the size of the corpora they arise from being so different. For example, Twarc picked up nearly 8,000 more hashtag uses than RAPID in Part 1, but fewer than 200 more in Part 2. Notably, although the most prolific account is different in Part 2, the

[8] Sometimes short or obscure filter terms, like 'qanda', have meanings in non-target languages.

TABLE 4.4. Detailed statistics of Q&A Parts 1 and 2.

	Q&A Part 1		Q&A Part 2	
	RAPID	Twarc	RAPID	Twarc
Tweets	15,930	27,389	11,719	15,490
Quotes	325	1,203	498	1,232
Replies	1,446	2,067	1,715	1,731
Tweets with hashtags	10,043	15,591	3,912	3,961
Tweets with URLs	2,470	4,029	3,106	4,074
Most prolific account	Account a_1	Account a_1	Account a_2	Account a_3
Tweets by most prolific account	103	146	57	68
Most retweeted tweet	Tweet t_1	Tweet t_1	Tweet t_2	Tweet t_2
Most retweeted tweet count	260	288	385	385
Most replied to tweet	Tweet t_3	Tweet t_3	Tweet t_4	Tweet t_4
Most replied to tweet count	55	121	58	58
Tweets with mentions	11,314	18,253	10,472	13,514
Most mentioned account	Account a_4	Account a_4	Account a_4	Account a_4
Mentions of most mentioned account	2,883	3,853	2,753	2,752
Hashtags uses	15,700	23,557	7,672	7,862
Unique hashtags	1,015	1,438	960	1,082
Most used hashtag	`#qanda`	`#qanda`	`#qanda`	`#qanda`
Uses of most used hashtag	10,065	15,644	2,545	2,549
Next most used hashtag	`#auspol`	`#auspol`	`#auspol`	`#auspol`
Uses of next most used hashtag	1,381	2,103	1,652	1,349
URLs uses	913	1,650	1,602	2,411
Unique URLs	399	560	658	790
Most used URL	http://wp.me/p2WW3S-Gg	http://wp.me/p2WW3S-Gg	Tweet t_5 URL	Tweet t_6 URL
Uses of most used URL	49	128	71	81

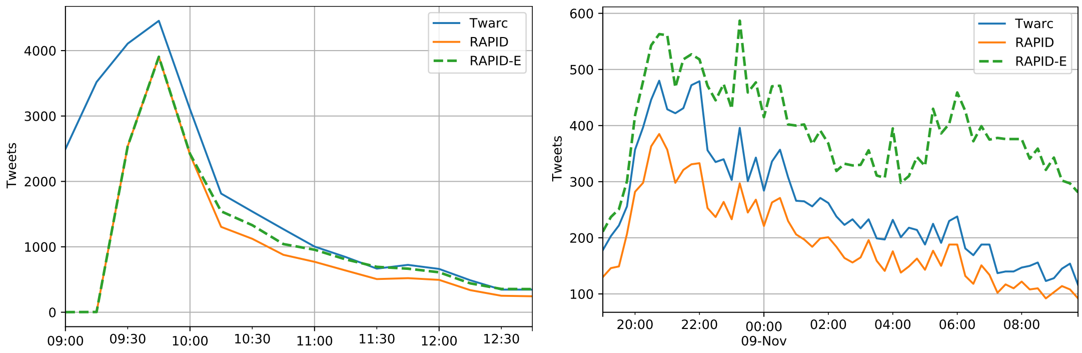
TABLE 4.5. The top ten most used hashtags in the Q&A datasets (ignoring case and anonymising names).

Q&A Part 1				Q&A Part 2			
RAPID		Twarc		RAPID		Twarc	
15,930 tweets		*27,389 tweets*		*11,719 tweets*		*15,490 tweets*	
Hashtag	Count	Hashtag	Count	Hashtag	Count	Hashtag	Count
qanda	10,065	qanda	15,644	qanda	2,545	qanda	2,549
auspol	1,381	auspol	2,103	auspol	1,652	auspol	1,349
ulurustatement	179	nbn	223	Surname of a_4	179	Surname of a_4	179
nbn	178	ulurustatement	187	nbn	135	nbn	133
Surname of a_4	137	Surname of a_4	179	breaking	85	ulurustatement	73
marriageequality	125	marriageequality	145	ulurustatement	72	pmlive	71
felizjueves	114	felizjueves	128	qldpol	65	qldpol	64
climate	73	ssm	80	nswpol	65	nswpol	63
8kasımdünyadelilergünü	61	climate	77	pmlive	64	marriageequality	53
ssm	60	libspill	76	marriageequality	53	springst	49

most mentioned account is the same for both Parts 1 and 2, potentially implying that account has had similarly high influence in both parallel datasets. Furthermore, both datasets shared almost all the same top ten hashtags, though in different orders (see Table 4.5). Approximately 5,000 of the extra hashtag uses are of ‘`#qanda`’. In Part 2, again, the top ten hashtags are nearly the same, but this time the usage counts are similar, except for ‘`#auspol`’ being used 22% more often in RAPID (1,652 times compared with 1,349), which would account for the overall difference of 190 uses when combined with the noise of lesser used hashtags. The most used URL in Part 1 is a shortened form of a link to a political party policy comparison resource prepared by an account prominent in the `#auspol` Twitter discussion.[9] In the longer collection, the most prominent URL is overtaken by retweets, one by `@QandA` (Tweet t_5) and one

[9] https://otiose94.wordpress.com/2015/05/30/nett_news-by-otiose94/ The site's most recent post was on 2020-04-25, but content the content for this particular post is missing as of 2022-01-29.

by @SkyNewsAust, an official news media account (Tweet t_6).



(a) Q&A Part 1.

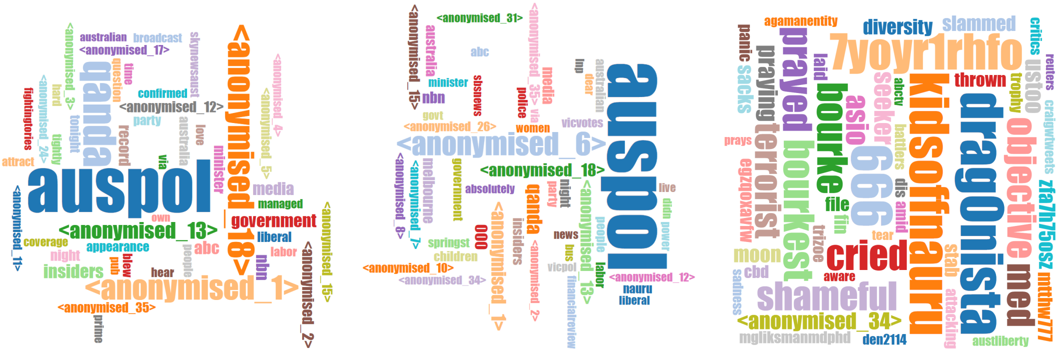
(b) Q&A Part 2.

FIGURE 4.2. Twitter activity in the Q&A Parts 1 and 2 dataset over time (in 15 minute blocks).

Moving beyond the bare statistics, the timelines shown in Figures 4.2a and 4.2b show the clear differences in tweets retrieved. Though the Twarc and pared back RAPID timelines appear at least proportionately similar, it is firstly notable that the RAPID-E dataset captured so much less data in Part 1 (Figure 4.2a) and so much more in Part 2 (Figure 4.2b), particularly from approximately 4 a.m. onwards (UTC). One possible explanation for this is that the discussion on the night of the episode was far more directly focused on the episode themes and had less opportunity to drift to other issues, especially while informed and guided by what was being broadcast at the time. In contrast, those discussing the episode the following day would have had more opportunity to broaden the discussion to other topics, and RAPID's topic tracking attended to that, apparently at the cost of tweets matching the exact filter terms. Word clouds of the terms drawn from the first and last 5,000 tweets of the RAPID-E dataset appear to offer mild support for this (Figure 4.3). Terms are sized according to their frequency. The discussion across the day focuses on the #auspol hashtag, but #qanda is more prominent early on. Mentions of anonymised IDs 1 and 18 are prominent early but shift to ID 6 later. All of these IDs refer to the same individual[10], but by Twitter handle and first name early on and by surname later in the day. Figure 4.3c, showing the top terms unique to the evening discussion, indicates that the discussion shifts to humanitarian concerns (e.g., "kidsoffnauru", "[asylum] seeker", "shameful", "cried", "sadness"), perhaps due to events of the day. The early discussion (Figure 4.3a) seems to mention individuals much more than later, as indicated by the greater size of anonymised IDs. This fact alone implies that the early discussion was focused more directly on the Q&A episode, as the topics it covered related to particular relationships and events involving those individuals.

The second notable feature is that the RAPID tool appeared to miss many of the available tweets in the first half an hour of the Part 1 collection. RAPID-E's first half hour includes only six tweets, the first of which was at 9 a.m. (UTC), while RAPID's

[10] Variants of this individual's name were used as filter terms.



(a) From the first 5,000 tweets. (b) From the last 5,000 tweets. (c) Terms unique to the last 5,000 tweets.

FIGURE 4.3. Word clouds of the 50 most used terms (anonymised consistently across the subfigures) in the first and last 5,000 tweets of the Q&A Part 2 RAPID-E dataset, and the top 100 terms unique to the last 5,000 tweets.

only includes four tweets, the first of which was at 9:15 a.m. It is unclear why the tool missed the tweets that Twarc captured, but a discrepancy such as this suggests it was not by design. The reason that the RAPID-E included tweets without the key terms early on in the specified timeframe is that the collection was running prior to the cut-off at 9 a.m. (UTC), tracking topics while it ran, as a 'burn-in' period, and we have extracted just these specific periods (UTC 0900 to 1300, and UTC 1900 to 1000 the next day) to study, post collection.

4.4.1.2 Comparison of network statistics

Given the differences in datasets, we expect differences in the derived social networks (Tables 4.6 and 4.7) (Holzmann et al., 2018). We also present the proportional balance between each dataset's statistics in Figures 4.4 and 4.5. Each network is dominated by a single large component, comprising over 90% of nodes in the retweet and mention networks and around 70% in the reply networks. The distributions of component sizes appear to follow a power law, resulting in corresponding high numbers of detected clusters.

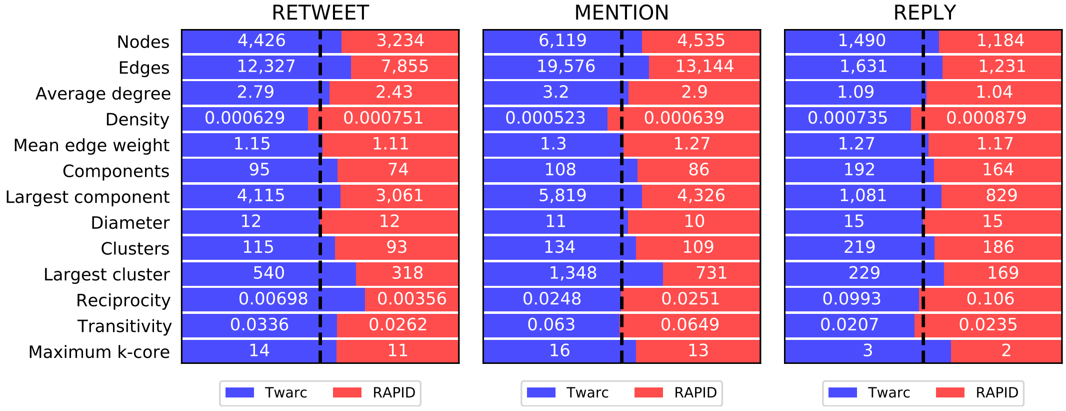


FIGURE 4.4. The proportional balance between Twarc and RAPID statistics of the retweet, mention and reply networks built from the Q&A Part 1 datasets.

Table 4.6. Q&A Part 1 network statistics.

	RETWEET		MENTION		REPLY	
	RAPID	Twarc	RAPID	Twarc	RAPID	Twarc
Nodes	3,234	4,426	4,535	6,119	1,184	1,490
Edges	7,855	12,327	13,144	19,576	1,231	1,631
Average degree	2.429	2.785	2.898	3.199	1.040	1.095
Density	0.001	0.001	0.001	0.001	0.001	0.001
Mean edge weight	1.113	1.151	1.268	1.300	1.175	1.267
Components	74	95	86	108	164	192
Largest component	3,061	4,115	4,326	5,819	829	1,081
- Diameter	12	12	10	11	15	15
Clusters	93	115	109	134	186	219
Largest cluster	318	540	731	1,348	169	229
Reciprocity	0.004	0.007	0.025	0.025	0.106	0.099
Transitivity	0.026	0.034	0.065	0.063	0.024	0.021
Maximum k-core	11	14	13	16	2	3

Table 4.7. Q&A Part 2 network statistics.

	RETWEET		MENTION		REPLY	
	RAPID	Twarc	RAPID	Twarc	RAPID	Twarc
Nodes	3,594	4,591	5,198	6,205	1,492	1,507
Edges	7,344	10,110	14,802	18,184	1,560	1,576
Average degree	2.043	2.202	2.848	2.931	1.046	1.046
Density	0.001	0.000	0.001	0.000	0.001	0.001
Mean edge weight	1.096	1.087	1.245	1.222	1.099	1.098
Components	118	176	123	179	196	201
Largest component	3,308	4,085	4,854	5,612	1,073	1,080
- Diameter	12	11	10	10	15	15
Clusters	138	197	158	210	221	226
Largest cluster	471	727	1,090	1,513	122	123
Reciprocity	0.004	0.004	0.024	0.025	0.072	0.071
Transitivity	0.027	0.026	0.084	0.079	0.016	0.016
Maxmium k-core	9	10	11	14	3	3

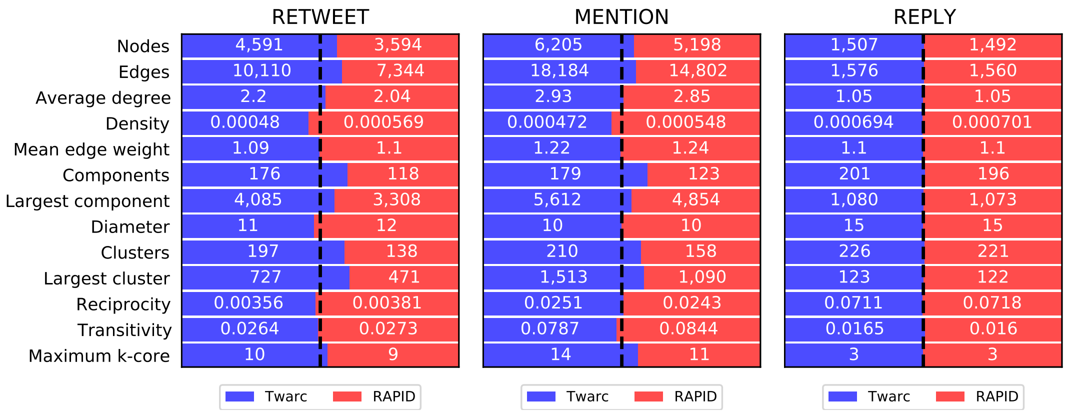


Figure 4.5. The proportional balance between Twarc and RAPID statistics of the retweet, mention and reply networks built from the Q&A Part 2 datasets.

Network structure statistics like density, diameter (of the largest component in disconnected networks), reciprocity and transitivity may offer insight into social behaviours such as influence and information gathering. The high component counts in all networks lead to low densities and correspondingly low transitivities, as the potential number of triads is limited by the connectivity of nodes. That said, the largest components were consistently larger in the Twarc datasets, but the diameters of the corresponding largest components from each dataset were remarkably similar, implying that the extra nodes and edges were in the components' centres rather than on the periphery. This increase in internal structures improves connectivity and therefore the number of nodes to which any one node could pass information (and therefore influence) or, at least, reduces the length of paths between nodes so information can pass more quickly. The similarities in transitivity imply the increase may not be significant, however, with networks of these sizes. Reciprocity values may provide insight into information gathering, which often relies on patterns of to-and-fro communication as a person asks a question and others respond. Interestingly, the only significant difference in reciprocity is in the Part 1 retweet networks, with the Twarc dataset having a reciprocity nearly double that of the RAPID dataset (though still small). The Twarc dataset includes 60% more retweets than the corresponding RAPID dataset and 40% more accounts (Table 4.3), which may account for the discrepancy. Given the network sizes, the reciprocity values indicate low degrees of conversation, mostly in the reply networks. Interestingly, mean edge weights are very low (1.3 at most), implying that most interactions between accounts in all networks happen only once, despite these being corpora of issue-based discussions.

The proportional statistical differences between the corresponding datasets are highlighted in Figure 4.4 for Part 1 and Figure 4.5 for Part 2. Part 1's Twarc networks were larger, both in nodes and edges, but less dense, than the RAPID ones, and the largest component in each network is larger by a significant proportion of the extra nodes (it is not clear what portion of the extra nodes are members of the largest components, however). An increase in components also led to an corresponding increase in detected clusters, and an increase in the size of the largest detected cluster. As mentioned earlier, the increase in internal structures leads to a higher maximum k -ore value. Though the proportional differences in reciprocity in the retweet networks are high, the values themselves remain low. Part 2's reply networks are remarkably similar despite the Twarc dataset having 26% more tweets. The differences in Part 2's retweet and mention networks are similar to those of Part 1.

That the differences in retweet and mention networks are so proportionately similar across both Parts 1 and 2 is notable because the retweet network is not based on direct interactions, while the mention network is. Retweeting a tweet links a retweeter, X, back to the original author, Y, of a tweet, rather than any intermediate account, even if the retweet passed through several accounts on its way between Y and X. It is possible that these datasets were sufficiently constrained both in size and timespan and focus of the participants (by which we mean they engaged in the discussion by following the `#qanda` hashtag), that there was little opportunity to build up chains of retweets.

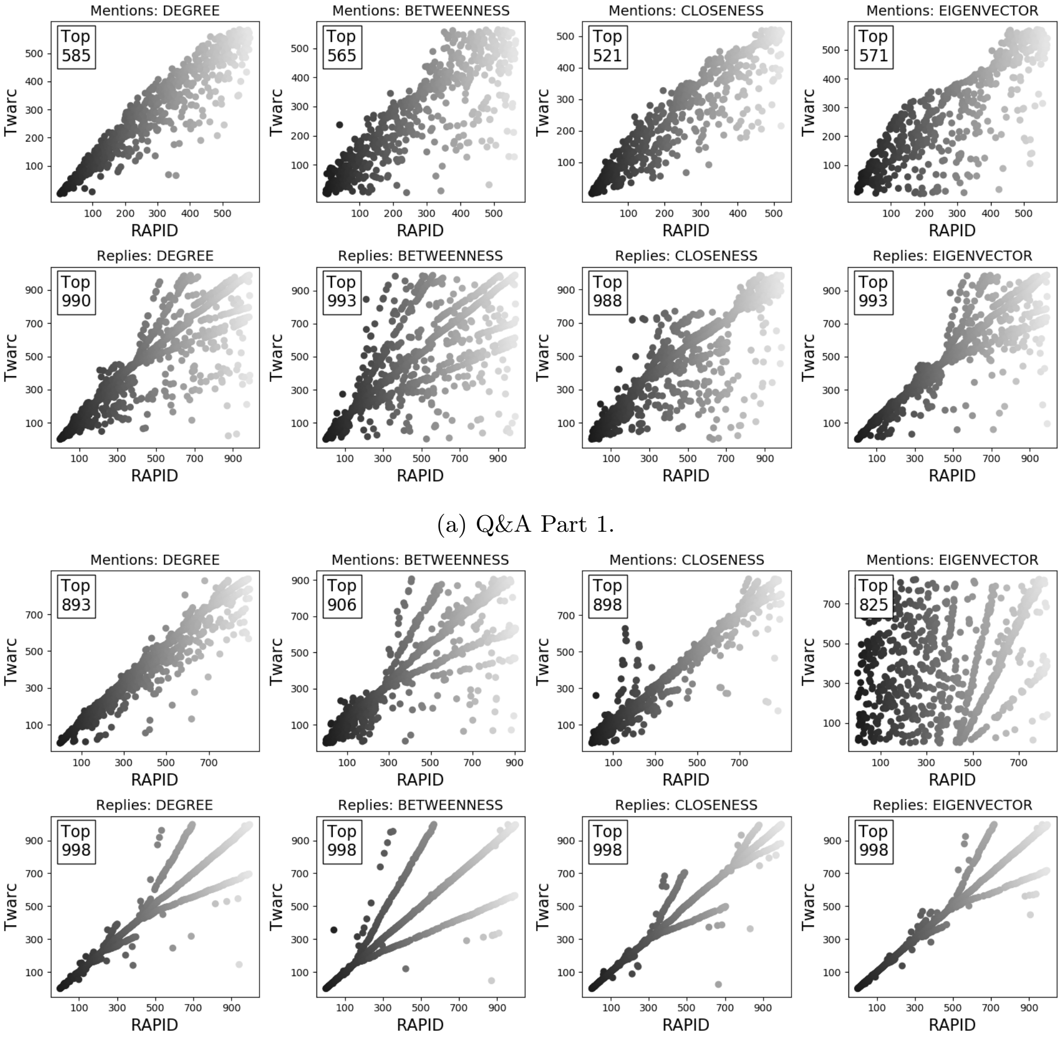
Next we look at two major categories of network analysis: *indexing*, for the computation of node-level properties, such as centrality, and *grouping*, for the computation of specific groups of nodes, such as clustering.

4.4.1.3 Comparison of centralities

Centrality measures can tell us about the influence an individual has over their neighbourhood, though the timing of interactions should ideally be taken into account to get a better understanding of their dynamic aspects (e.g., Falzon et al. 2018). If networks are constructed from partial data, network-level metrics (e.g., radius, shortest paths, cluster detection) and neighbourhood-aware measures (e.g., eigenvector and Katz centrality) may vary and not be meaningful (Holzmann et al., 2018).

We compare centralities of corresponding networks using scatter plots of node rankings, as per Section 4.3.4 (Figure 4.6). The symmetrical structures come from corresponding shifts in order: if an item appears higher in one list, then it displaces another, leading to the evident fork-like patterns. There is considerable variation in most centrality rankings for both mention and reply networks in Part 1 (Figure 4.6a) but much less in Part 2 (Figure 4.6b), apart from the ranking of eigenvector centralities for the mention networks, which lacks almost any alignment between the RAPID and Twarc node rankings, despite the high number of common nodes (825). This implies that the neighbourhoods of nodes differ between the Twarc and RAPID mention networks, but the top-ranked nodes are similar though their orders differ greatly. Furthermore, the relatively few common nodes in Part 1's Twarc mention networks (521 to 585) and greater edge count (Tables 4.6) could indicate that the extra edges significantly affect the node rankings. However, Part 2's Twarc mentions networks also had many more edges, but many more nodes in common (approximately 900). Thus it must have been how the mentions were distributed in the datasets that differed, rather than simply their number. It is not clear that Part 1's four-hour duration (*cf.*, Part 2's 15 hours) explains this. Instead, if we look at the 11,480 tweets unique to Twarc in Part 1 (*cf.*, fewer than 4,000 are unique to Twarc in Part 2, Table 4.3), only 622 are replies, whereas 6,915 include mentions. There are also 34% more unique accounts in the Part 1 Twarc dataset, but only 19% more in the Part 2 Twarc dataset (Table 4.3). Each mention refers to one of these accounts and forms an extra edge in the mention network, thus altering the network's structure and the centrality values of many of its nodes; this is likely where the variation in rankings originates.

The Kendall τ and Spearman's ρ coefficients were calculated comparing the corresponding lists of nodes, each pair ranked by one of the four centrality measures (Figure 4.7). Although somewhat proportional, it is notable how different the coefficient



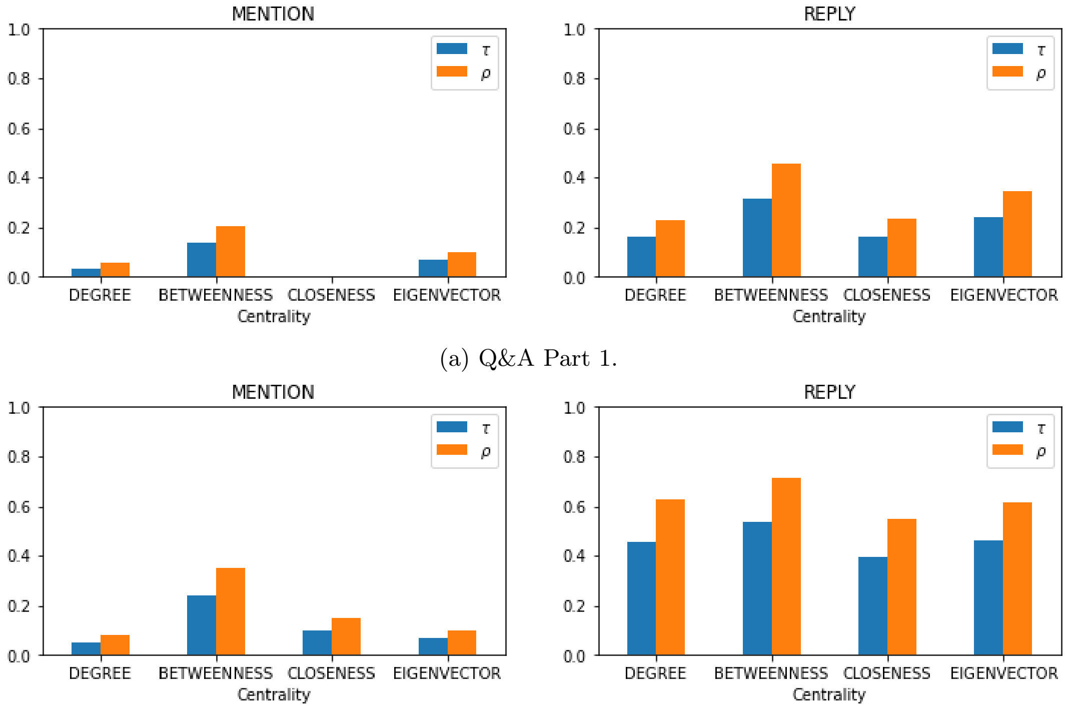
(a) Q&A Part 1.

(b) Q&A Part 2.

Figure 4.6. Centrality ranking comparison scatter plots of the mention and reply networks built from the Q&A Parts 1 and 2 datasets. In each plot, each point represents a node's ranking in the RAPID and Twarc lists of centralities (common nodes amongst the top 1,000 of each list). The number of nodes appearing in both lists is inset. Point darkness indicates rank on the x axis (darker = higher).

values are, especially in Part 2. While Twarc produced more tweets than RAPID (Table 4.3), and more unique accounts, the corresponding mention and reply node counts are not significantly higher (Tables 4.6 and 4.7). In fact, the node counts in the Part 1 reply networks are correspondingly lower than in the Part 2 reply networks, even though both Part 2 datasets were smaller. Edge counts in the mention networks were very different (Twarc had many more) but were quite similar in the reply networks.

The biggest variation was in the mention networks from Part 1 (Figure 4.6a and Table 4.6), due to the large number of extra mentions from Twarc. It is notable that the Kendall's τ was low for all mention networks (Figure 4.7), especially for degree and closeness centrality. It is worth noting the minor differences in the degree and



(a) Q&A Part 1.

(b) Q&A Part 2.

FIGURE 4.7. Centrality ranking comparisons using Kendall τ and Spearman's ρ coefficients for corresponding mention and reply networks made from the Q&A Parts 1 and 2 datasets.

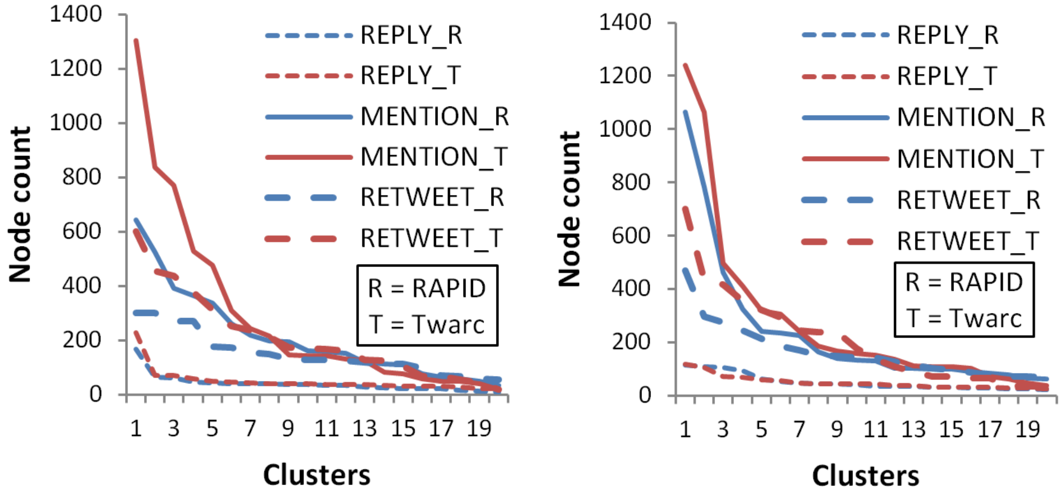
immediate neighbours of nodes impacts degree and closeness centralities significantly, and, correspondingly, their relative rankings. In contrast, rankings for betweenness and eigenvector centrality, which rely more on global network structure, remained relatively stable.

4.4.1.4 Comparison of clusters

We finally compare the networks via largest clusters (Figure 4.8). The reply network clusters are relatively similar, and the largest mention and reply clusters differ the most. The ARI scores (Table 4.8) confirm that the reply clusters were most similar for Parts 1 and 2 (0.738 and 0.756, respectively), possibly due to the small size of the reply networks. The mention and retweet clusters for Part 2 were more similar than those of Part 1 (0.437 and 0.468 compared to 0.320 and 0.350), possibly due to the longer collection period. In Part 1, there is a chance the networks are different due to RAPID's expansion strategy. Changes to filter keywords may have collected posts of other vocal accounts not using the original keywords, at the cost of the posts which did.

4.4.1.5 Summary of findings

Overall, Twarc and RAPID provided very different views into the Twitter activity surrounding the Q&A episode in question, both on the evening of and the day after.



(a) Q&A Part 1.

(b) Q&A Part 2.

FIGURE 4.8. The largest retweet, mention and reply clusters built from the Q&A Parts 1 and 2 datasets.

TABLE 4.8. ARI scores for the clusters found in the corresponding retweet, mention and reply networks built from the Q&A Parts 1 and 2 datasets.

	RETWEET	MENTION	REPLY
Q&A Part 1	0.320	0.350	0.738
Q&A Part 2	0.437	0.468	0.756

This includes variations in basic collection statistics, network statistics for retweet, mention and reply networks built from the collected data, centrality measures of the nodes in the networks and comparison of detected clusters. The extra tweets collected by the Twarc collections appear to have resulted in greater numbers of connections internal to the largest components, which may have implications for the analysis of influence, as reachability correspondingly increases. Deeper study of reply content is required to inform patterns of information gathering.

We are left with the open question of how reliable social media can be as a data source, if conducting simultaneous collection activities with the same query criteria can provide such different networks. Is the variation due to the platform providing a random sample of the overall data or an effect of the tool being used?

We next considered a more tightly controlled comparison of Twarc and RAPID, disabling RAPID's expansion strategies so that the tools performed as similarly as possible.

4.4.2 Case study 2: A weekend of AFL without topic tracking

RAPID's topic tracking feature broadens the scope of of the collection at the cost of strictly matching tweets, resulting in distinctly different corresponding corpora. Although the rankings of the most central nodes in networks built from the corpora appear relatively stable, the question remains of why the corpora were so different in size. In this section, we discuss a case study in which we disabled RAPID's topic tracking feature, expecting the resulting corresponding corpora to increase in similarity, especially over a longer period collection. Figure 4.9 indicates that again, initially at least, it appeared that Twarc and RAPID produced very different, but proportional over time, datasets. Constraining the datasets to only those tweets with a "lang" property of "en" or "und" resulted in much more similar datasets.

TABLE 4.9. Summary dataset statistics of the AFL1 collection.

Dataset	All Tweets	Unique Tweets		Retweets		All Accounts	Unique Accounts	
Twarc	44,461	22,731	(51.1%)	11,482	(25.8%)	16,821	5,274	(31.4%)
RAPID	21,799	69	(0.3%)	7,047	(32.3%)	11,573	26	(0.2%)
Twarc-en	25,231	4,065	(16.1%)	8,531	(33.8%)	12,399	1,187	(9.6%)
RAPID-en	21,235	69	(0.3%)	6,849	(32.3%)	11,238	26	(0.2%)

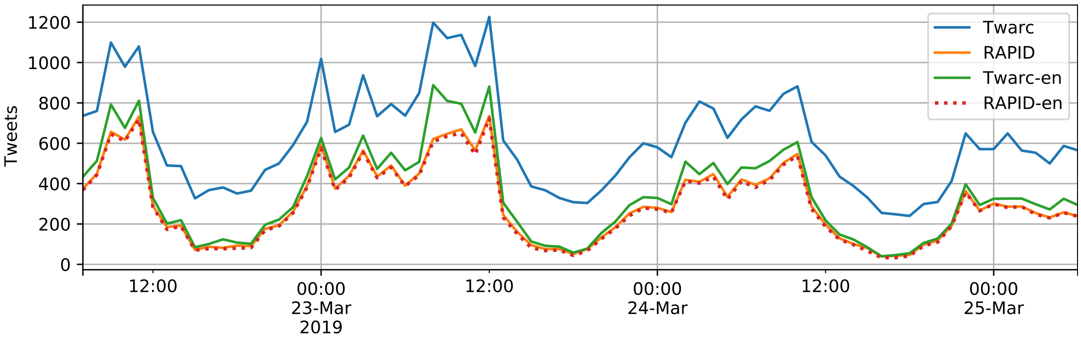


FIGURE 4.9. Twitter activity in the AFL1 dataset over time (in 60 minute blocks).

4.4.2.1 Comparison of collection statistics

We conducted two parallel collections under the term "afl" over a three-day period in March 2019 (the start of the AFL season) using RAPID without topic tracking and Twarc. This collection is labelled "AFL1" in Tables 4.1 and 4.2, and further details are offered in Table 4.9. The datasets obtained appear to be dramatically different: RAPID collected just shy of 22,000 tweets while Twarc found approximately twice that number with around 45,000 tweets, with 21,730 in common. Interestingly, as can be seen in Figure 4.9, the extra tweets appear to occur relatively evenly and consistently over time, rather than spiking. On closer inspection, it became apparent that the balance in languages was different, with 36% of the Twarc collection having `lang` property of 'jp' (Japanese) and 52% 'en' (English), while RAPID consisted of 94% English tweets (Figure 4.10). When both collections were trimmed to tweets with a `lang` property of 'en' or 'und' (undefined), they reduced to 25,231 tweets (Twarc) and 21,235 tweets (RAPID), with 21,166 in common, which still leaves more than 4,000 tweets specific to Twarc (Figure 4.11). The "AFL1" dataset, reduced to only posts with a `lang` property of 'en' or 'und' is referred to as "AFL1-en" henceforth.

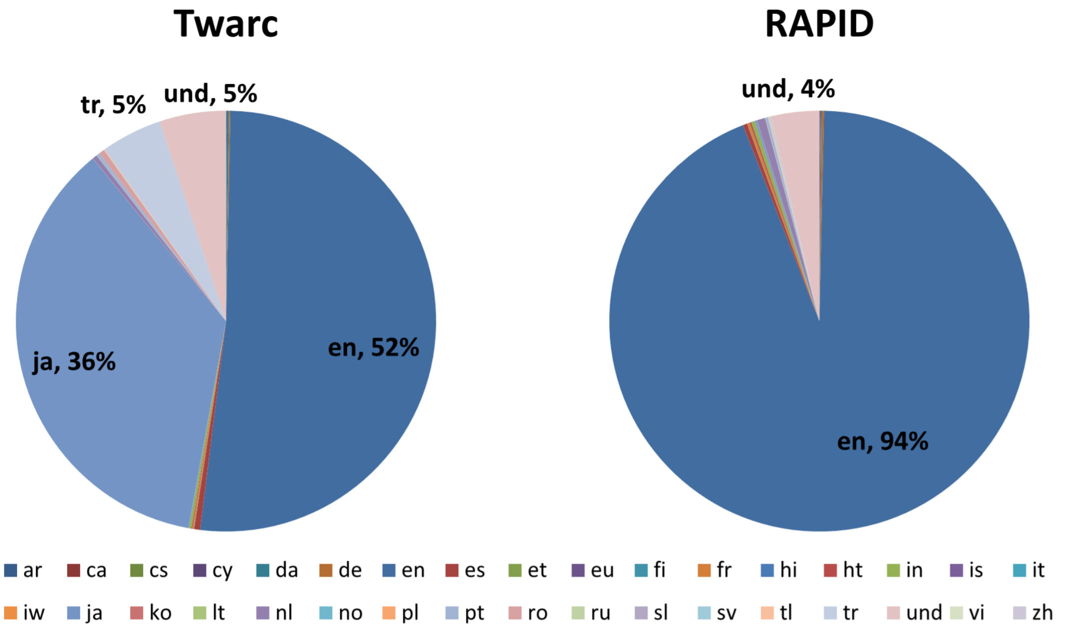


Figure 4.10. Distributions of tweet language values (specified in the `lang` property) in the RAPID and Twarc datasets, collected using the filter term "afl".

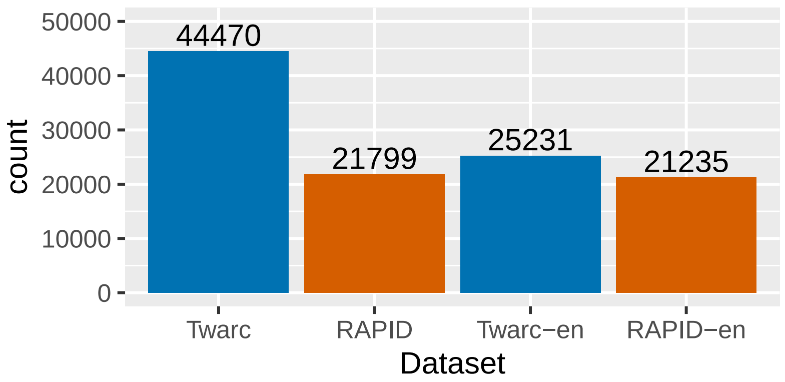


Figure 4.11. Tweet counts of the RAPID and Twarc datasets in the AFL1 and AFL1-en collections, obtained using the filter term "afl".

As previously mentioned, RAPID does not retain tweets which do not contain filter terms in text-related portions of the tweets. In the Twarc collection, the term 'afl' appeared in the domain of a website that many of the Japanese tweets referred to, belonging to an online marketplace. These tweets were dropped by RAPID and did not appear in the final collection.

Only 69 tweets were unique to the RAPID AFL1-en dataset, and they appear to be AFL-related sports discussions. The 4,065 tweets unique to the Twarc dataset comprise 2,595 English tweets and 1,470 with "und" for the `lang` value. This field is populated by Twitter based on language detection algorithms. When a language cannot be detected, such as when there is not sufficient free text to analyse, the value "und" is used. Inspection of the tweets indicates the reason for this: the undefined tweets include 884 retweets, 1,366 tweets with URLs, 116 with hashtags, and 916 with

TABLE 4.10. Statistics of the AFL1-en RAPID and Twarc datasets.

Property	Twarc-en	RAPID-en
Tweets	25,231	21,235
Accounts	12,399	11,238
Retweets	8,531	6,849
Quotes	2,291	1,615
Replies	6,185	5,936
Tweets with hashtags	7,606	6,911
Tweets with URLs	10,266	7,345
Most prolific account	Account a_5	Account a_5
Tweets by most prolific account	363	362
Most retweeted tweet	Tweet t_5	Tweet t_5
Most retweeted tweet count	529	529
Most replied to tweet	Tweet t_6	Tweet t_6
Most replied to tweet count	141	141
Tweets with mentions	17,467	15,230
Most mentioned account	Account a_6	Account a_6
Mentions of most mentioned account	7,131	7,130
Hashtag uses	17,352	15,886
Unique hashtags	2,381	2,249
Most used hashtag	#AFL	#AFL
Most used hashtag count	4,523	4,522
Next most used hashtag	#AflPiesCats	#AflPiesCats
Uses of next most used hashtag	1,575	1,482
URL uses	6,557	3,552
Unique URLs	2,843	2,043
Most used URL	http://watchrugby.net/AFL/	http://watchrugby.net/AFL/
Uses of most used URL	494	251

mentions. Of the "und" tweets with URLs, the vast majority (1,188) refer to a Japanese online electronics marketplace (771) and a Japanese online media platform (417). The next largest group refer to 38 retweets, some of the official @AFL account (9), though there are 16 and 5 retweets of two accounts that Botometer (Davis et al., 2016) scored at 4.2 and 4.4 out of 5, respectively, as bot-like, and both refer to the previously mentioned Japanese electronics marketplace. The top 12 most used hashtags in the English subset relate to the AFL, while the top 14 for the "und" subset are all Japanese terms, except for "iphone" (at number 9). The top term (in Japanese) is the name of the marketplace. The English tweets are mostly related to the AFL, though there is considerable obvious content from bot-like accounts, with several accounts posting the same content (offers of live streams of the matches) repeatedly within a short space of time (their messages appear adjacent in the timeline).

Once reduced to a relatively comparable state, the "AFL1-en" parallel datasets can be examined in more detail. It is understood that the tweets they consist of will differ, given that rate-limiting constraints may have caused each to receive different tweets. The statistics in Table 4.10 bear this out with the Twarc dataset statistics being approximately proportionately larger when compared with the RAPID dataset statistics. The author IDs have been anonymised, but the most mentioned account is the official @AFL account, while the most prolific author appears to be automated to

TABLE 4.11. Comparative statistics for networks generated from the RAPID and Twarc datasets for the ALF1-en collection.

	RETWEET		MENTION		REPLY	
	RAPID-en	Twarc-en	RAPID-en	Twarc-en	RAPID-en	Twarc-en
Number of nodes	5,584	6,430	11,566	12,525	4,705	4,759
Number of edges	5,881	6,977	22,310	23,937	4,928	5,005
Average degree	1.053	1.085	1.929	1.911	1.047	1.052
Density	0.000	0.000	0.000	0.000	0.000	0.000
Mean edge weight	1.165	1.223	1.308	1.329	1.205	1.236
Components	494	536	666	713	791	798
Largest component	3,946	4,233	9,416	9,789	2,951	3,017
- Diameter	16	16	17	17	16	16
Clusters	544	579	736	781	861	863
Largest cluster	659	753	2,177	2,125	851	844
Reciprocity	0.004	0.005	0.057	0.055	0.139	0.131
Transitivity	0.035	0.038	0.143	0.152	0.039	0.034
Maximum k-core	4	4	9	11	4	4

some degree, having posted nearly 35,000 tweets in two years and a Botometer (Davis et al., 2016) Complete Automation Probability (CAP[11]) of 68%, many seemingly promote the AFL, tennis, and a singer. The most replied to tweet was posted by an Australian NBA[12] player and the most retweeted tweet was of an amusing video of an AFL supporter.

4.4.2.2 Comparison of network statistics

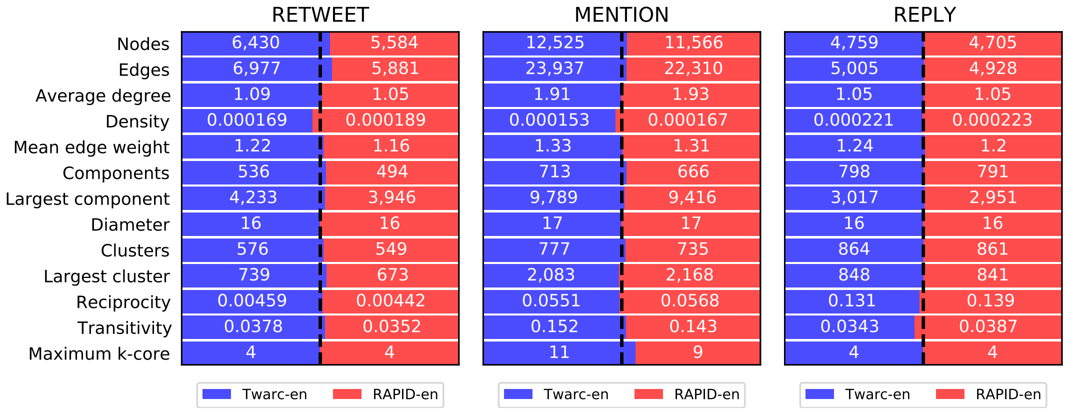


FIGURE 4.12. The proportional balance between Twarc and RAPID statistics of the retweet, mention and reply networks built from the AFL1-en datasets.

The network statistics Table 4.11 indicate that the networks were much more similar than in the Q&A case study, though there are still notable differences. The largest components of the retweet, mention and reply networks are, at most, 15% larger by node count, and the largest components are correspondingly similar, though their diameters and densities indicate they are much more sparse than the corresponding

[11] See https://botometer.osome.iu.edu/faq#which-score. Accessed 2022-01-26.
[12] United States National Basketball Association.

Q&A ones, with corresponding implications for the opportunity to influence. In contrast, in sporting discussions, there is less motivation to attempt to convert fellow sports fans to cheer for one's team than there is in a political discussion. Certainly in this study, politics has tended to generate more discussion than sports in general, and the nature of the discussions is also different. The reciprocity values here are much higher than in the Q&A case study, implying the presence of more communication among the communities that do exist. Another difference that lends weight to this interpretation is the average degree of nodes in the networks. In the Q&A retweet and mention networks, the average degrees were around 2-2.5 and 3, respectively, implying some repetition in connectivity, whereas in the sporting discussing the average degrees are around 1 and 2, respectively, implying much less continued interaction. As indicated in Figure 4.12, the degree values of the Twarc and RAPID networks are highly similar.

The number of tweets and accounts in the AFL1-en datasets (Table 4.10), coupled with the number of nodes and edges in the derived mention and reply networks (Table 4.11), indicates that although the AFL1-en collections differed by nearly 4,000 tweets, the number of accounts was not significantly different (approximately 10% more in Twarc) with a corresponding increase in nodes and edges in the mention network (8.3% and 7.3%, respectively) but only 54 and 77 (1.1% and 1.6%, respectively) more in the reply network.

4.4.2.3 Comparison of centralities

Considering the similarity of interaction networks constructed from the respective AFL1-en datasets, we compare the relative ranking of the top network nodes by various centrality values (with an upper bound of 1,000 nodes). Figure 4.13 shows scatterplots of the relative rankings of nodes common to corresponding networks, and Figure 4.14 plots the Kendall τ and Spearman's coefficients of the corresponding relative rankings. As with the Q&A collection, the centralities of nodes in the reply networks show more similarity than those in the mention networks, which is likely due to their relative size; Table 4.11 indicates a significant discrepancy in the reply and mention network sizes and average degree. Closeness is notably low in similarity, though the high component count would account for that. It is apparent the most central nodes in both network types mostly maintain their ordering for the first several hundred nodes, but all begin to diverge at some point. A few isolated nodes change their ranking significantly, such as those in the top left of the mentions betweenness and closeness plots, degrading their rankings (appearing above the diagonal), and those in the reply closeness and eigenvector plots, improving their rankings (appearing below the diagonal), but the majority diverge in a trident pattern, implying lower-ranked nodes improve their rankings swapping out higher-ranked nodes at progressively greater distances. The reason for the consistency is unclear. Minor variations would ensure that nodes' centrality values varied, and thus, their rankings could easily vary significantly, especially due to

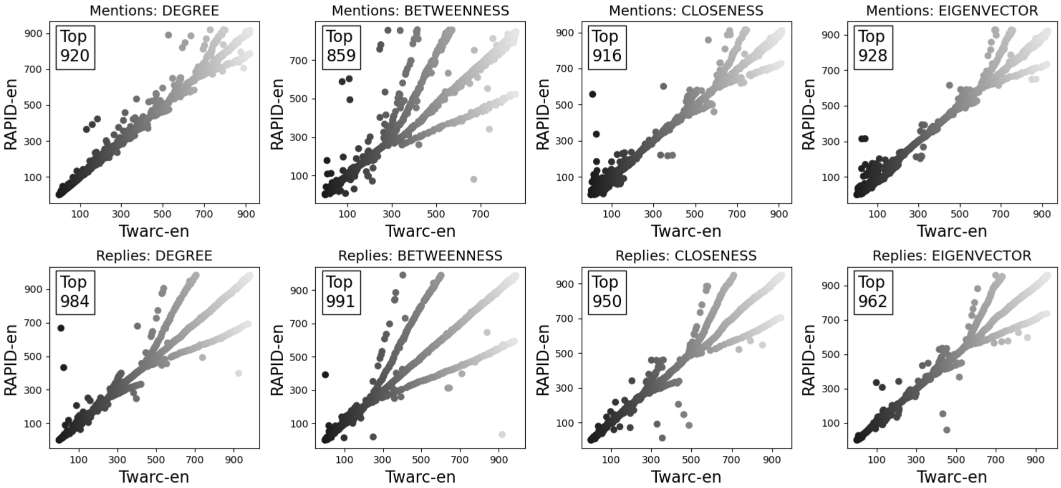


FIGURE 4.13. Centrality ranking comparison scatter plots of the mention and reply networks built from the AFL1-en datasets. In each plot, each point represents a node's ranking in the RAPID-en and Twarc-en lists of centralities (common nodes amongst the top 1,000 of each list). The number of nodes appearing in both lists is inset. Point darkness indicates rank on the x axis (darker = higher).

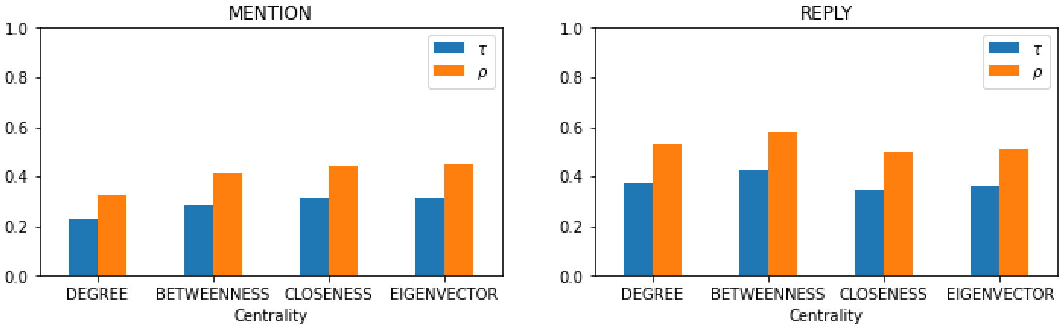


FIGURE 4.14. Centrality ranking comparisons from the RAPID and Twarc datasets of the AFL1-en collection using Kendall τ scores and Spearman's coefficients.

the high number of components. The high k-core values for the mention networks are likely to explain the high betweenness and eigenvector centrality values, as the highest ranked of these will reside in the larger components, which will have the greater likelihood of being similar across the networks.

4.4.2.4 Comparison of clusters

Comparing the clusters detected with the Louvain method (Blondel et al., 2008) in the retweet, mention and reply networks results in ARI values in Table 4.12. This implies that although the networks consisted of many components, the clusters they formed were highly similar for retweet and reply networks, and only slightly less so for the mention networks, despite the fact that the Twarc mention network included more than 2,000 more mention edges.

Table 4.12. ARI scores for the clusters found in the networks built from the RAPID and Twarc datasets for the AFL1-en collection.

RETWEET	MENTION	REPLY
0.818	0.675	0.853

Table 4.13. Summary dataset statistics of the AFL2 collection.

Dataset	All Tweets	Unique Tweets		Retweets		All Accounts	Unique Accounts	
RAPID1	30,103	-	(0.0%)	9,215	(30.6%)	14,231	-	(0.0%)
RAPID2	30,115	12	(0.0%)	9,215	(30.6%)	14,232	1	(0.0%)

4.4.2.5 Summary of findings

This case study makes it clear that the tool used for collection can have a significant effect on the data collected and the resulting analytic results. It was serendipitous that the filter term chosen was “afl”, because a more specific term or set of terms is unlikely to have captured the non-English content that Twarc did. This highlighted the fact that RAPID was post-processing and filtering the tweets it collected, and raises general questions for social media data collection: Do other collection tools, especially commercial ones, do this post-processing too, as a “convenience” or “value-add” to their users? Do they make it clear if and when they do? The validity of evidence-based conclusions rests on these details. Even when both datasets were filtered to ensure some degree of consistency, there remained large differences in the networks constructed from them. Minor differences in datasets may result in amplified differences in analyses.

A further, even more fundamental, question remained after this case study, which is addressed by the next subsection: *Does the same tool provide the same data over two simultaneous collections with identical filter terms?*

4.4.3 Case study 3: Tracking AFL Twitter activity with RAPID

Given it appeared that different collection tools could produce different results using the same inputs, the question of whether APIs are delivering consistent content for all clients remained. A second collection (Table 4.13) was initiated over a longer period (six days) using the same filter term and tool (RAPID), but with different API credentials. One set of credentials belonged to a relatively new and unused account (created in 2018 having posted only 3 tweets) and the other to a well-established account (created in 2009 and having posted approximately 17,000 tweets). This resulted in two highly similar, but not quite identical, datasets, with sizes 30,103 and 30,115 tweets; their timeline is shown in Figure 4.15. The first dataset was a proper subset of the second, so the difference of 12 posts can be regarded as due to noise or minor differences in timing. A brief examination revealed these extra tweets (shown in blue in the Figure) were all about AFL or other sports in Australia, and their

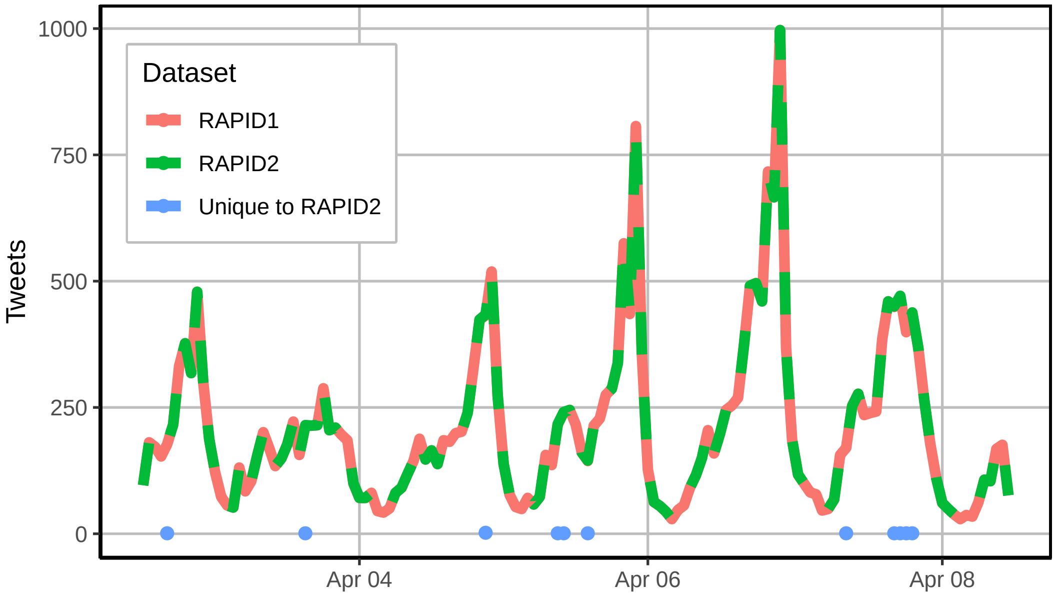


Figure 4.15. Twitter activity in the AFL2 dataset over time (in 60 minute blocks). Thick and dashed lines are used here to highlight how the timeseries overlap almost exactly. The timestamps of tweets unique to RAPID2 are shown as blue points.

timing appeared random. Further confirmation of the similarity between datasets can be seen in Table 4.14 where the most prolific account, most retweeted tweet, most replied to tweet and most mentioned account details are all identical. Again, the most mentioned account is the official @AFL account.

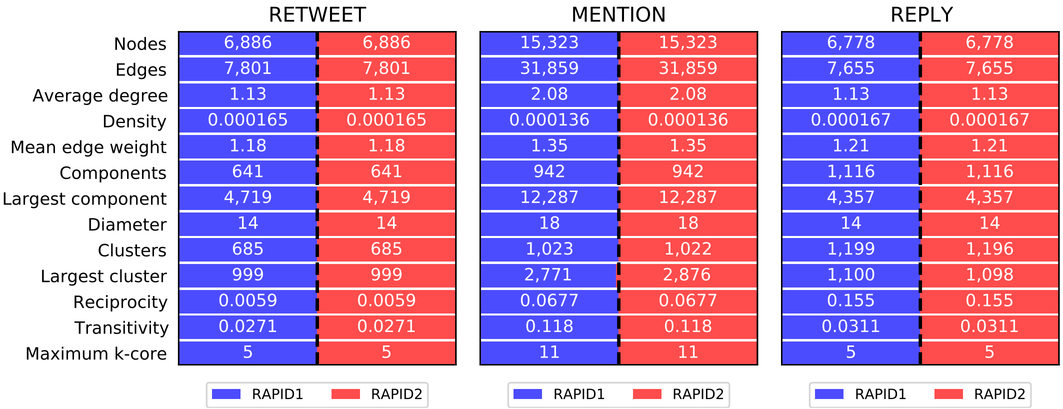
4.4.3.1 Comparison of network statistics

Due to the similarity of the datasets, the retweet, mention and reply networks generated from them were almost identical, and only a summary of the structures is provided in Table 4.15. Details of the networks are provided in Figure 4.16, which show that the only differences occur in the detected clusters. In particular, the largest cluster detected in the RAPID2 mention network is around 3% larger than the corresponding cluster from the RAPID1 mention network. This is likely due to an element of randomness used in the Louvain algorithm (Blondel et al., 2008).

4.4.3.2 Comparison of centralities and clusters

The similarity of the networks based on their statistics is further confirmed by a comparison of their centrality rankings, which indicates that their structures are all but identical. A visual inspection of their respective rankings in Figure 4.17 reveals no major differences, and the Kendall τ and Spearman's coefficients indicate their rankings are, in fact identical (Figure 4.18).

Interestingly, the high degree of similarity does not extend to the membership of detected clusters using the ARI measure (Table 4.16). Presumably, the sensitivity of the measure indicates that these scores must be as close to the maximum as we could



	RETWEET		MENTION		REPLY	
	RAPID1	RAPID2	RAPID1	RAPID2	RAPID1	RAPID2
Nodes	6,886	6,886	15,323	15,323	6,778	6,778
Edges	7,801	7,801	31,859	31,859	7,655	7,655
Average degree	1.13	1.13	2.08	2.08	1.13	1.13
Density	0.000165	0.000165	0.000136	0.000136	0.000167	0.000167
Mean edge weight	1.18	1.18	1.35	1.35	1.21	1.21
Components	641	641	942	942	1,116	1,116
Largest component	4,719	4,719	12,287	12,287	4,357	4,357
Diameter	14	14	18	18	14	14
Clusters	685	685	1,023	1,022	1,199	1,196
Largest cluster	999	999	2,771	2,876	1,100	1,098
Reciprocity	0.0059	0.0059	0.0677	0.0677	0.155	0.155
Transitivity	0.0271	0.0271	0.118	0.118	0.0311	0.0311
Maximum k-core	5	5	11	11	5	5

FIGURE 4.16. The proportional balance between the RAPID1 and RAPID2 statistics of the retweet, mention and reply networks built from the AFL2 datasets.

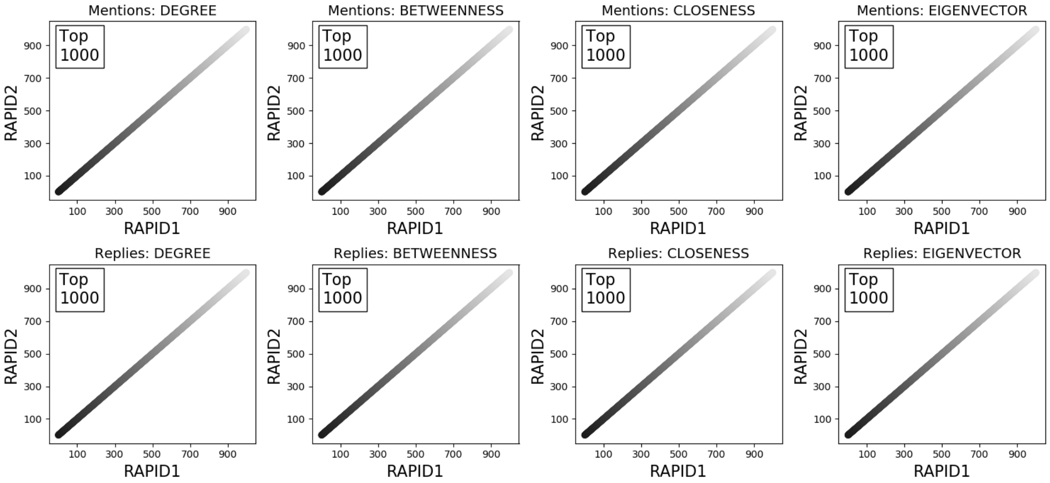


FIGURE 4.17. Centrality ranking comparison scatter plots of the mention and reply networks built from the AFL2 datasets. In each plot, each point represents a node's ranking in the RAPID1 and RAPID2 lists of centralities (common nodes amongst the top 1,000 of each list). The number of nodes appearing in both lists is inset. Point darkness indicates rank on the x axis (darker = higher).

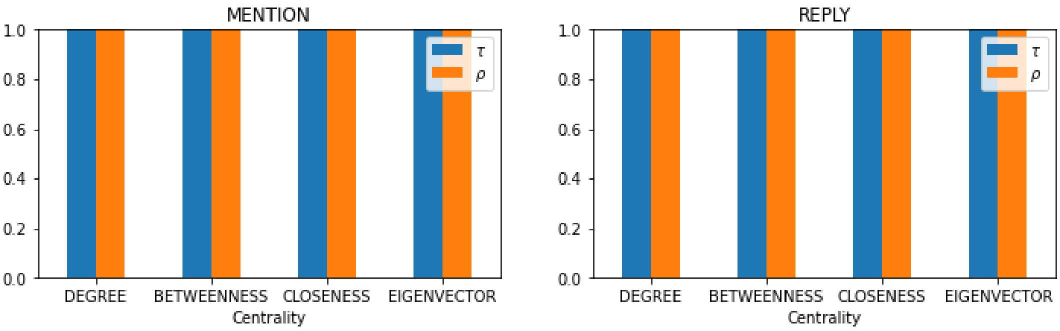


FIGURE 4.18. Centrality ranking comparisons from the two RAPID datasets of the AFL2 collection using Kendall τ scores and Spearman's coefficients.

Table 4.14. Statistics of two parallel datasets collected using RAPID with the filter term “afl” over a six-day period with different API credentials.

Property	RAPID1	RAPID2
Tweet count	30,103	30,115
Retweet count	9,215	9,215
Account count	14,231	14,232
Quote count	2,340	2,341
Reply count	9,229	9,229
Tweets with hashtags	8,623	8,627
Tweets with URLs	11,467	11,474
Most prolific account	Account a_7	Account a_7
Most prolific account tweet count	612	612
Most retweeted tweet	Tweet t_7	Tweet t_7
Most retweeted tweet count	269	269
Most replied to tweet	Tweet t_8	Tweet t_8
Most replied to tweet count	206	206
Tweets with mentions	22,083	22,083
Most mentioned account	Account a_6	Account a_6
Most mentioned account count	10,468	10,468
Hashtag uses	20,136	20,140
Unique hashtags	3,337	3,337
Most used hashtag	#AFL	#AFL
Most used hashtag count	5,096	5,096
Next most used hashtag	#AflDeesDons	#AflDeesDons
Uses of next most used hashtag	759	759
URL uses	5,702	5,709
Unique URLs	3,580	3,587
Most used URL	http://watchrugby.net/AFL/	http://watchrugby.net/AFL/
Most used URL count	341	341

expect due to the degree of randomness inherent in Louvain clustering. For a score of 1.0 each pair of detected clusters would need to match perfectly, across the thousands of nodes in the networks, so any minor variation will reduce that score.

4.4.3.3 Summary of findings

This evidence suggests that the results provided by the Twitter API (if not other platforms’ APIs) are consistent, regardless of the consumer. It is clearly important that a researcher understand how their collection tool works to guarantee their understanding of the results returned. In this regard, open-source solutions are, as the name implies, more transparent than closed-source solutions. The benefit gained as a result of more tailored filtering must be balanced against the initial effort required

Table 4.15. Selected comparative statistics for networks generated from the two RAPID datasets for the AFL2 collection.

			MENTION		REPLY	
Dataset	Tweets	Accounts	Nodes	Edges	Nodes	Edges
RAPID1	30,103	14,231	15,323	31,859	6,778	7,655
RAPID2	30,115	14,232	15,323	31,859	6,778	7,655

Table 4.16. ARI scores for the clusters found in the networks built from the RAPID and Twarc datasets for the AFL2 collection.

RETWEET	MENTION	REPLY
0.916	0.808	0.865

Table 4.17. Summary dataset statistics of the Election Day collection.

Dataset	All Tweets	Unique Tweets		Retweets		All Accounts	Unique Accounts	
Twarc	39,293	3	(0.0%)	26,412	(67.2%)	10,860	1	(0.0%)
RAPID	39,556	285	(0.7%)	26,612	(67.3%)	10,893	36	(0.3%)
RAPID-E	46,526	7,255	(15.6%)	30,735	(66.1%)	12,696	1,839	(14.5%)
Tweepy	36,172	0	(0.0%)	24,276	(67.1%)	10,242	0	(0.0%)

to understand how the APIs are employed by the tool used and what modifications tools make to the data they collect.

4.4.4 Case study 4: Election day

This final case study highlights the importance of continuous network connectivity, and awareness of when that condition is not met. Given the social media researcher can offload many other aspects of data quality to the OSN (e.g., well-designed schemas, data consistency, value validity, Scannapieco et al., 2005; Foidl and Felderer, 2019), it is important to note that this is an aspect for which the researcher must retain responsibility.

To consider a more focused collection activity and to consider a second open-source collection tool (thus similar to the baseline tool, Twarc), a collection was conducted over an election day (24 hour period) in early 2019, using RAPID, Twarc and Tweepy, each configured with the same filter terms: `#NswVotes`, `#NswElection`, `#nswpol`, and `#NswVotes2019`. RAPID and Twarc collected slightly below 40,000 tweets each while Tweepy collected around 36,000 tweets, but suffered from network outages on two occasions for approximately 110 and 96 minutes each time (see Figure 4.19). In the resulting datasets (highlights of which are shown in Table 4.17), 285 tweets were unique to RAPID, three to Twarc, and 19 were shared by Twarc and Tweepy but not RAPID. The vast majority of the Tweepy dataset's 36,172 tweets appeared in all three datasets, while Tweepy missed the 3,118 further tweets that appeared in both Twarc and RAPID datasets. In fact, by examining the periods where Tweepy lost its connection, around 6 p.m. (UTC) and again approximately six hours later, Twarc retained 3,036 tweets while RAPID retained 3,055 tweets (RAPID-E collected 3,918 during these periods), so it is possible that if Tweepy's connection had stayed up, the Tweepy dataset might have been very similar to Twarc and RAPID, especially as the remainder of the collection behaviour of the tools appears almost identical in the timeline.

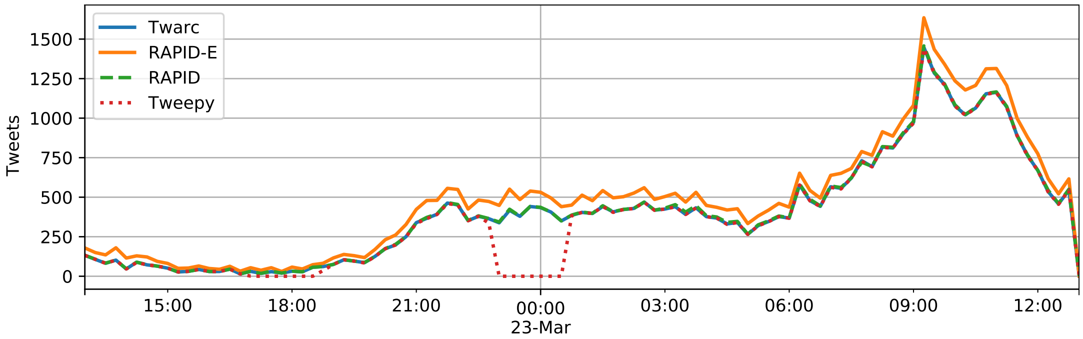


Figure 4.19. Twitter activity in the Election Day dataset over time (in 15 minute blocks). Dashed and dotted lines are used here to highlight how the timeseries overlap almost exactly.

4.4.4.1 Comparison of collection statistics

The collection statistics are highly similar and are provided primarily for completeness. The effect of Tweepy's disconnection is highlighted by the differences in its statistics from Twarc as the baseline. Although more than 3,000 tweets were missed, only a few hundred accounts, quotes, replies and tweets with URLs were missed. Several thousand retweets were missed as well as tweets with hashtags and mentions, but the effect on the features with the highest counts is limited. The most prolific account, most retweeted tweet, most replied to tweet, most mentioned accounts, hashtags and URLs are all the same (Table 4.18).

4.4.4.2 Comparison of network statistics

Continuing the similarities in the collection statistics, statistics drawn from retweet, mention and reply networks built from the Election Day datasets are also strikingly resilient, despite the Tweepy networks including several hundred fewer nodes (Table 4.19). This is borne out by the proportional differences between Twarc and RAPID in Figure 4.20, where the only significant difference is the size of the largest detected cluster (again, likely due to the randomness inherent in the Louvain algorithm, Blondel et al., 2008) and then in the proportional differences in all the statistics across the Twarc and Tweepy networks in Figure 4.21.

4.4.4.3 Comparison of centralities

Examining the centralities of the mention and reply networks built from the Election Day datasets, comparing RAPID and Tweepy against the Twarc baseline shows, as expected, only minor variations in the RAPID dataset which only occur among the lower ranked nodes (Figure 4.22) and more widespread differences with the Tweepy networks (Figure 4.23). Statistically, Twarc and RAPID's mention network centrality rankings, shown in Figure 4.24, had Kendall τ values around 0.35 to 0.4 and Spearman's coefficients around 0.45 to 0.6, while the reply networks' values were higher, with τ around 0.5 and Spearman's coefficient around 0.7, possibly due to the smaller

TABLE 4.18. Statistics of the Twarc, RAPID, and Tweepy datasets collected in parallel over a 24 hour period. *https://www.fiverr.com/s2/ee030ef08d This task is no longer active as of 2022-01-29.

Property	Twarc	RAPID	Tweepy
Tweets	39,297	39,556	36,172
Accounts	10,860	10,893	10,242
Retweets	26,412	26,612	24,276
Quotes	3,590	3,610	3,363
Replies	1,374	1,381	1,252
Tweets with hashtags	21,582	21,686	19,977
Tweets with URLs	7,829	7,860	7,194
Most prolific account	Account a_8	Account a_8	Account a_8
Tweets by most prolific account	212	211	212
Most retweeted tweet	Tweet t_9	Tweet t_9	Tweet t_9
Most retweeted tweet count	368	367	278
Most replied to tweet	Tweet t_{10}	Tweet t_{10}	Tweet t_{10}
Most replied to tweet count	25	26	24
Tweets with mentions	30,626	30,848	28,154
Most mentioned account	Account a_9	Account a_9	Account a_9
Mentions of most mentioned account	2,442	2,443	2,187
Hashtag uses	51,288	51,470	47,106
Unique hashtags	2,450	2,458	2,306
Most used hashtag	#NswVotes	#NswVotes	#Nswvotes
Most used hashtag count	11,739	11,731	10,901
Next most used hashtag	#NswVotes2019	#NswVotes2019	#NswVotes2019
Uses of next most used hashtag	7,606	7,602	6,968
URL uses	3,766	3,761	3,478
Unique URLs	1,374	1,374	1,258
Most used URL	URL 1*	URL 1*	URL 1*
Uses of most used URL	100	100	100

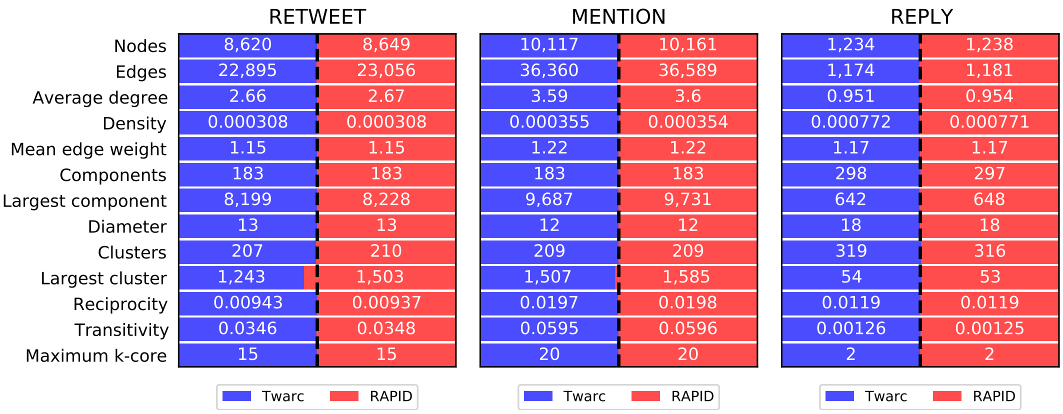


FIGURE 4.20. The proportional balance between Twarc and RAPID statistics of the retweet, mention and reply networks built from the Twarc and RAPID datasets.

Table 4.19. Comparative statistics for networks generated from the Twarc, RAPID and Tweepy datasets for the Election Day collection.

	RETWEET			MENTION			REPLY		
	Twarc	RAPID	Tweepy	Twarc	RAPID	Tweepy	Twarc	RAPID	Tweepy
Number of nodes	8,620	8,649	8,193	10,117	10,161	9,600	1,234	1,238	1,147
Number of edges	22,895	23,056	21,122	36,360	36,589	33,595	1,174	1,181	1074
Average degree	2.656	2.666	2.578	3.594	3.601	3.500	0.951	0.954	0.936
Density	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.001	0.001
Mean edge weight	1.154	1.154	1.149	1.217	1.218	1.209	1.170	1.169	1.166
Component count	183	183	183	183	183	178	298	297	286
Largest component	8,199	8,228	7,779	9,687	9,731	9,146	642	648	569
- Diameter	13	13	13	12	12	12	18	18	19
Clusters	207	210	210	209	209	204	319	316	304
Largest cluster	1,243	1,503	1,405	1,507	1,585	1,582	54	53	53
Reciprocity	0.009	0.009	0.009	0.020	0.020	0.019	0.012	0.012	0.011
Transitivity	0.035	0.035	0.032	0.060	0.060	0.058	0.001	0.001	0.001
Maximum k-core	15	15	14	20	20	19	2	2	2

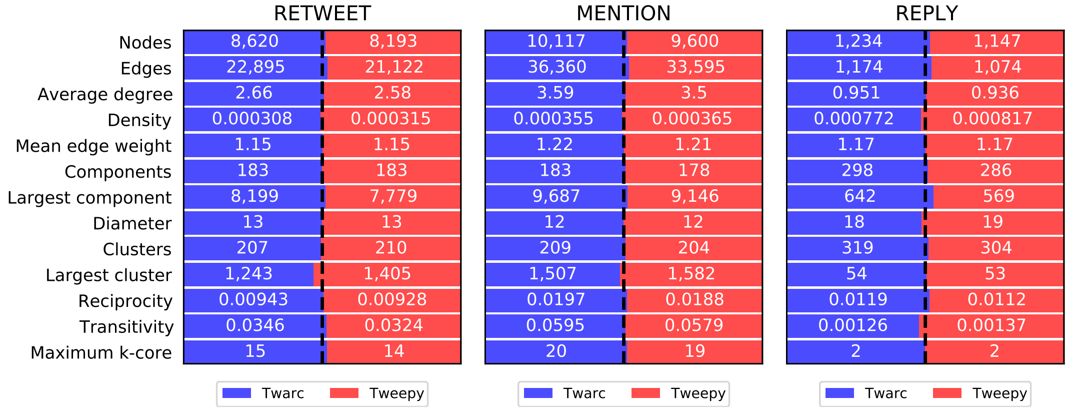


Figure 4.21. The proportional balance between Twarc and RAPID statistics of the retweet, mention and reply networks built from the Twarc and Tweepy datasets.

size of the reply networks. These values are all approaching or exceeding the τ value of 0.4 to 0.6 that was regarded as reasonably to highly similar, mentioned in Section 4.3.4. The ranking similarity statistics calculated by comparing the Twarc and Tweepy baselines are notably lower (Figure 4.25, though even the reply networks' betweenness and closeness comparisons are moderately similar with τ around 0.4 and Spearman's coefficient around 0.5 to 0.6.

4.4.4.4 Comparison of clusters

Despite the similarities between the Twarc and RAPID networks, the cluster membership still varies significantly, with the highest similarity being found amongst the (smaller) reply networks, as can be seen in the ARI scores in Table 4.20. The clusters found in the Twarc and Tweepy networks are less similar, almost in line with the differences in network sizes: the retweet networks had fewer nodes than the mention networks, and the ARI scores are less different, and the reply networks were the smallest and had the smallest difference between ARI scores.

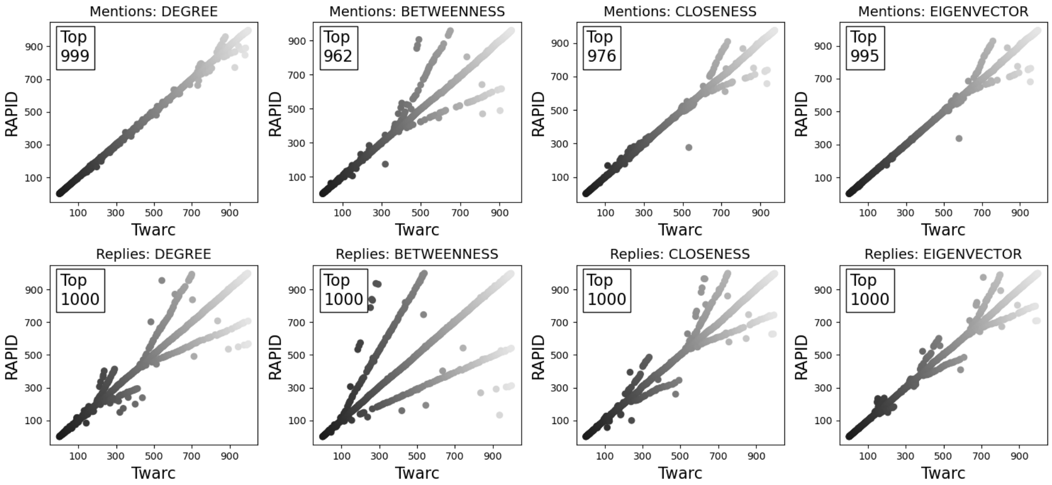


Figure 4.22. Centrality ranking comparison scatter plots of the mention and reply networks built from the Twarc and RAPID Election Day datasets. In each plot, each point represents a node's ranking in the Twarc and RAPID lists of centralities (common nodes amongst the top 1,000 of each list). The number of nodes appearing in both lists is inset. Point darkness indicates rank on the x axis (darker = higher).

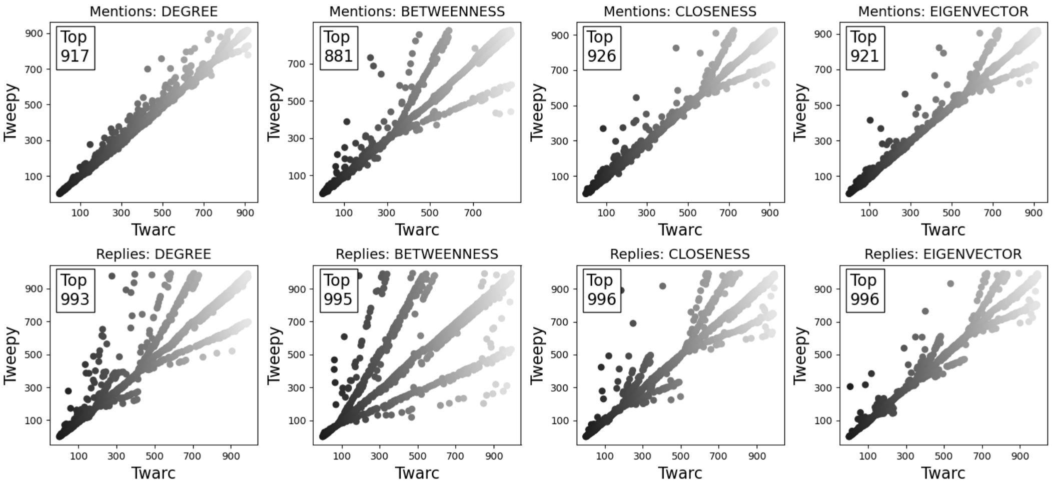


Figure 4.23. Centrality ranking comparison scatter plots of the mention and reply networks built from the Twarc and Tweepy Election Day datasets. In each plot, each point represents a node's ranking in the Twarc and Tweepy lists of centralities (common nodes amongst the top 1,000 of each list). The number of nodes appearing in both lists is inset. Point darkness indicates rank on the x axis (darker = higher).

Table 4.20. ARI scores for the clusters found in the corresponding retweet, mention and reply networks built from the Election Day datasets.

	RETWEET	MENTION	REPLY
Twarc/RAPID	0.547	0.656	0.737
Twarc/Tweepy	0.453	0.534	0.703

4.4.4.5 Summary of findings

This final case study provides us with further confidence that the differences observed early on in the Q&A datasets are primarily caused by enhancements provided by the

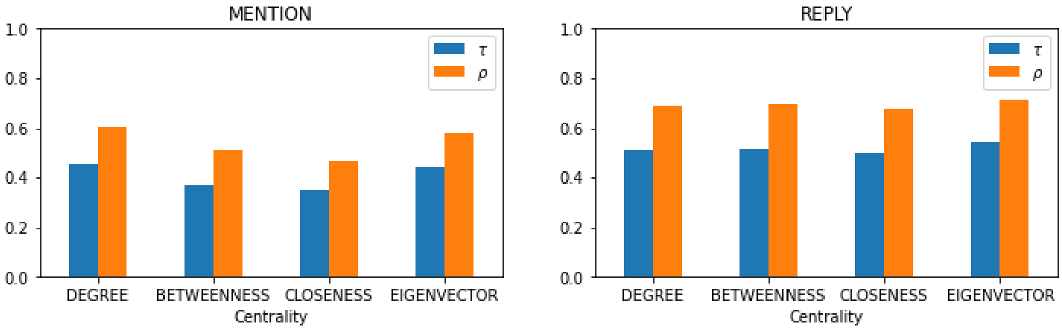


FIGURE 4.24. Centrality ranking comparisons from the Twarc and RAPID datasets of the Election Day collection using Kendall τ scores and Spearman's coefficients.

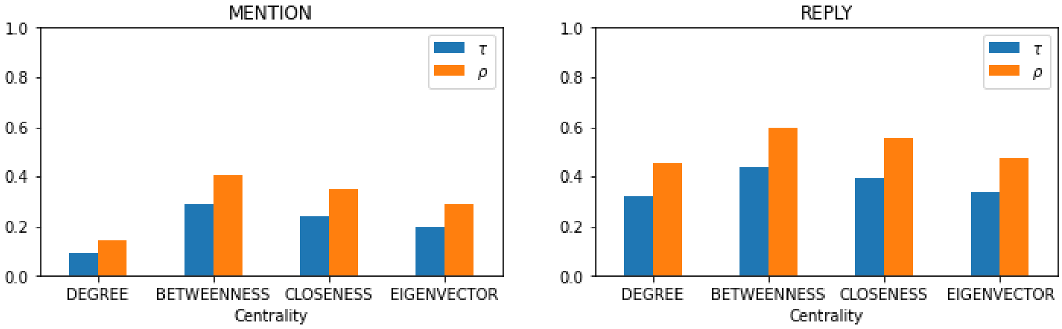


FIGURE 4.25. Centrality ranking comparisons from the Twarc and Tweepy datasets of the Election Day collection using Kendall τ scores and Spearman's coefficients.

RAPID platform and the differences in the AFL1 datasets were due, in part, to the choice of "afl" as the lone filter term. The Election Day collection used several specific filter terms and ran long enough to collect several tens of thousands of tweets, enough time to avoid minor differences in start and stop times. Even the differences that did occur did not result in significant effects on several networks constructed from the data or on network analysis measures calculated over those networks.

4.5 Discussion

A number of points worthy of further discussion have been raised by these case studies, and here we consider the statistical effect of the case study variations, specifically, but then also more general questions regarding the size of datasets, the effects of language and terminology, and the influence of the platforms.

4.5.1 Regarding statistics

The case studies presented highlight how decisions regarding collection specification, such as the filter terms used or their number, and the collection duration, can result in datasets that trigger features in complex collection tools, quite apart from configuration of such tools to dynamically change the collection specification (e.g., use of RAPID's topic tracking feature). The primary variations explored here involved filter terms although collection duration also varied, depending on the collection event.

The biggest variations in parallel datasets appeared when few filter terms were used and when they were short (i.e., having few characters), resulting in incidental noise from posts in unexpected languages (`#qanda`) or with unexpected acronyms and from elements in post metadata (`#AFL`). When multiple terms were used, and when those terms were not valid words in a language (e.g., variations on `#NswElec`), the parallel datasets were much more similar. Although it might be common sense to encourage careful design of collection specifications, these case studies highlight the value in (and danger in not) being more specific, by dictating the language of posts required as well as using multiple filter terms.

When variations in datasets occurred, the extra tweets resulted in the introduction of new nodes (accounts) in retweet, mention and reply networks, the majority of which were located within the largest connected network components (relatively few appeared as new, independent components). This consistently reduced the density of the retweet, mention and reply networks, but rarely affected the diameter of the largest component (Q&A Part 2's retweet network is an exception here), implying that the new nodes appear in the core of the components, rather than on the periphery. Consequently, the extra nodes increased reciprocity, transitivity, and sometimes maximum k-core values in retweet and mention networks, but rarely changed reply networks. Reply interactions occurred least frequently in all datasets, and so reply networks were the least different in raw size (nodes and edges).

The effects of collection variation were most prominent in centrality scores, particularly when the collection event involved direct interaction between participants (e.g., issue- or theme-based discussions such as during Q&A and over weekends of football) and less straightforward information dissemination (e.g., during an election campaign). The ranking of nodes by centrality varied most in the mention and reply networks of Q&A Part 1, even though more than half the top thousand ranked nodes in each pair of parallel networks were the same (an average of 560.5 for mentions and 991 for replies). The forking patterns appearing in scatter plots imply the presence of groups of nodes with adjacent centrality rankings, which then swapped when new nodes were added, possibly through impacting the internal topology of the largest components in some way. Spearman's ρ and Kendall τ correlation coefficients were consistently higher for reply networks than mention networks, possibly due to their smaller size. No particular patterns in differences between centrality types were observed, which implies the differences between pairs of parallel networks did not result in significantly different topologies.

A final lesson regarding network statistics can be drawn from the use of ARI scores is that even clustering of highly similar (e.g., almost identical) networks (in Case Study 4) result in ARI scores around 0.7, meaning that ARI scores around 0.4 can be seen as confirmation cluster membership is, in fact, quite similar.

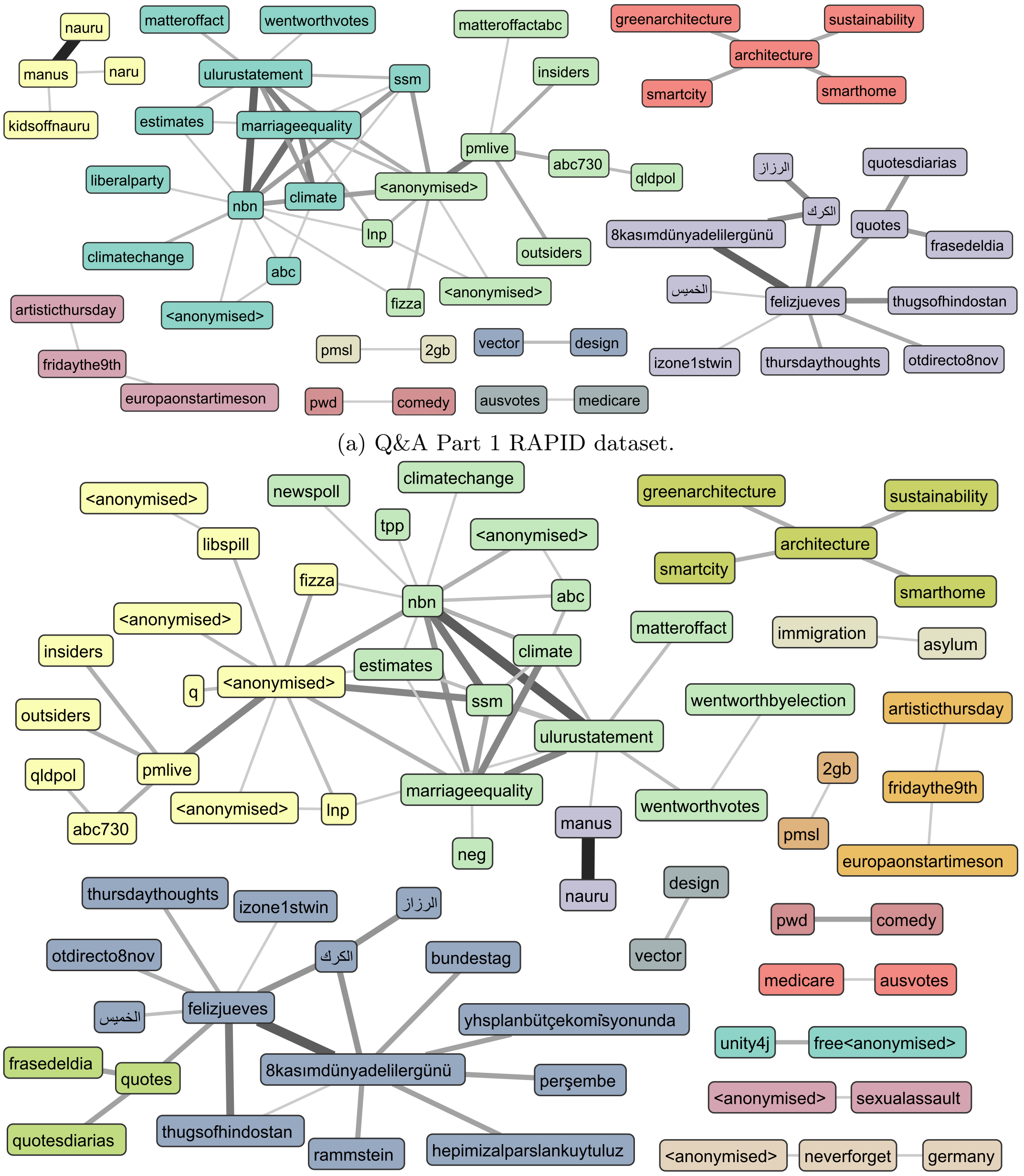
4.5.2 Regarding dataset size

Social media datasets analysed in the literature are often much larger than the datasets we have used in this study. For example, Cao et al. collected over a *billion* URLs sourced from Twitter alone in their study of URL sharing (Cao et al., 2015). There is significant power in such datasets to examine the flow of information and influence, but their scale can hamper more granular examinations focusing on accounts and the communities they form. The study of conversations can rely on direct interactions, such as replies or comments on posts and mentions, or indirect interactions such as the shared use of hashtags (e.g., Ackland, 2020). Such studies examine both the structure and dynamics of conversations and their prevalence, but those structures can be found in small, targeted datasets as well as larger ones. Information sharing via retweet or repost or URLs can reveal patterns of information dissemination and related research can certainly benefit from larger datasets, especially when relying on mathematical models of behaviour (e.g., Lee et al., 2013; Cao et al., 2015; Bagrow et al., 2019). Depending on researchers' access to privileged APIs and data access rates, generating large datasets can often encounter API rate limits, raising the question of completeness, which may or may not be an issue depending on the research questions under investigation. Assuming that collections activities are rate limited in a consistent way, we expect larger parallel datasets to exhibit many of the same patterns we have observed here, but this remains an open question for future research.

4.5.3 Regarding language and terminology clashes

Most popular OSNs have been developed in the English speaking world, primarily for an English-speaking audience (at least initially), and even though most now have significant non-English-speaking users (e.g., 56.5% of internet use originates in Western, Southern, Eastern and South Eastern Asia, Kemp, 2021, slide 27), English enjoys significant support. Though many OSNs support many languages and alphabets now, anglicised spelling variants for many non-English languages exist because the major mobile operating systems (Apple iOS and Google Android) originate in America. For these reasons, if terms (essentially combinations of letters) that are meaningful in multiple languages are used to filter streams or in queries, it is possible that non-English posts may be captured, especially if preferred languages are not specified as part of the filter or query. This was certainly the case for the Q&A case study (Figure 4.26). RAPID attempts to address this oversight by ensuring that filter terms appear in text-related fields in the posts it captures, but our experiences with the AFL datasets raise questions about what other terms can capture posts unexpectedly. Depending on how OSN queries are interpreted, using 'http', '#' or 'March' as filter terms could return every post including a URL, using a hashtag or posted in March—these are also questions for further research.

A secondary form of more common clash is semantic in nature, rather than lexicographical. A prime example of this is found in our Chapter 7 study of an Australian



(a) Q&A Part 1 RAPID dataset.

(b) Q&A Part 1 Twarc dataset.

FIGURE 4.26. *Semantic networks* (Radicioni et al., 2021) of co-mentioned hashtags (i.e., hashtags appearing in the same tweet) built from the RAPID and Twarc Q&A Part 1 datasets. The node for the hashtag `#qanda` has been excluded, as all tweets included it, and the minimum edge weight (i.e., times hashtags needed to co-occur in a tweet) was set to 3. Nodes are coloured according to Louvain clusters (Blondel et al., 2008), and labels identifying individuals have been anonymised. It is clear that non-English clusters of tweets have been captured due to a clash with the term 'qanda'.

election in which the filter term `#Liberals` (referring to a political party) clashed with the use of the term during student protests against gun violence in America, where the term refers more to political ideology, resulting in a spike of American tweets in a predominantly Australian discussion. Similarly, the hashtag `#VoteNo` clashed in a study of the 2017 Australian postal plebiscite on same sex marriage, drawing in American commentary on a healthcare Bill before the US Congress (see Chapter 6). To remove such 'Data Smells' (Foidl and Felderer, 2019), co-occuring hashtag networks, otherwise known as *semantic networks* (Radicioni et al., 2021), can be used to identify the out-of-scope content, but any use of automation is likely to require human oversight to avoid removing relevant content.

4.5.4 Regarding platform influences

Two of the biggest impediments to credibility in social media datasets are confidence that they are complete and, if they are known to be incomplete, knowledge of the sampling biases; both of these rely on openness and transparency on the part of the OSNs. Case study 3 (Section 4.4.3) at least confirms that different credentials, when used with the same collection tool against the same OSN with the same network boundaries (i.e., filter expression), result in approximately identical datasets (assuming some minor variation for the timing of network connections). OSNs are commercial entities and thus it stands to reason that they would bias samples to maintain users' attention, which could mean that if Case study 3 was repeated by running collections in different parts of the world, regional preferences (e.g., languages, topics of discussion) could influence the datasets, causing greater divergence. That said, studies of Twitter's 1% Sample API seem to offer evidence contrary to that (Joseph et al., 2014; Paik and Lin, 2015). The Sample API is different from a keyword-based query or stream filter, however, and is primarily designed to support research. Query term–based APIs might be more likely to exhibit regional differences, as the queries they service could originate from user-facing applications or market analysts, and not just academic researchers. Though these studies are all focused on Twitter, most other OSNs are under similar commercial pressures and regional popularity is vital to management of their brand.

4.5.5 Regarding measures for reliability and representativeness

The central purpose of this paper is to draw attention to unexpected variations in datasets collected from social media streams and the networks constructed from them. This is especially relevant when it is known that the stream is limited (either through platform rate-limiting or through platform algorithms, as occurs with, say, Twitter's 1% sample stream). An obvious follow-up question is whether or not an objective measure of reliability is feasible. This relates closely to the question of how representative samples provided by platforms are of their entire data holdings (e.g., as studied by Morstatter et al., 2013; González-Bailón et al., 2014; Joseph et al., 2014), but that question relies on examining the choices made by the platform in deciding what to include in the sample they offer. Here, similar to Paik and Lin (2015), our interest is in confirming that the data we request from a platform (with filter terms) matches what it has, or is at least representative of what it has (if rate-limits are encountered). Such a measure might rely on comparing the distributions of various features in our result dataset and the complete dataset (known only to the platform), such as the accounts and the number of tweets they post, the number of hashtags, URLs, and mentions used and replies, quotes and retweets made. Only the platform has sufficient information to calculate this measure, and there may be significant value in them providing it for the free or low cost streams they offer to researchers, analysts and other social media mass consumers. Providing a measure of representativeness (indicating reliability) alongside query and filter results could: (1) encourage consumers to pay for the higher cost streams, while also (2) providing consumers with more certainty in any conclusions they draw from the results they analyse. A measure of *reliability* rather than *representativeness* could, in fact, be more useful because there may be good reasons for results to not be truly representative – this would be the case when the complete dataset includes significant amounts of spam, pornography or other objectionable content.[13] The reliability measure would indicate how representative the provided results are when compared with the complete results.

4.6 Conclusion

Under a variety of conditions, the collection tools employed in several use cases provided different views of specific online discussions. These differences manifested as variations in collection statistics, and network-level and node-level statistics for retweet, mention and reply social networks built from the collected data. Extra tweets were most often collected by Twarc, and these appear to have resulted in more connections within the largest components without affecting their diameters. This may affect results of information diffusion analysis, as reachability correspondingly increases. Deeper study of reply content is required to inform discussion patterns.

How reliable social media can be as a source for research without deep knowledge of the effects of collection tools on analyses is an open question. If a tool *adds value* through analytics or data cleaning features, what is the nature of the effect? This paper provides a methodology to explore those effects. A canonical measure of the reliability of a dataset would be valuable to the research and broader social media analysis community. This measure would explain how complete the results of a search or filter of live posts is, and if it is not complete, how representative the provided sample results are of the complete results. Only the platforms have this information, however there would be benefits for them to provide such a measure, including as an enticement to consumers to pay for greater access to platform data holdings, as well as helping inform consumers of the degree to which they can depend on analyses of the data they receive. Twitter, in particular, has recently introduced changes as part of their API version 2.0 that facilitate academic research.[14]

We recommend the following to those using OSN data:

- Be aware of tool biases and their effects.

[13] This raises the question of how one could deliberately study such topics, however.

[14] https://developer.twitter.com/en/products/twitter-api/academic-research. Accessed 2022-01-14.

- Take care to specify filter and search conditions with keywords that capture relevant data and avoid irrelevant data, and make use of metadata filters to avoid unwanted content, e.g., constraining language codes. Beware of short filter terms and ones that are meaningful in non-target languages.
- Check the integrity of data. We observed gaps and minor inconsistencies in the Election case study due to connection failures as well as the appearance of duplicate tweets, identical in data and metadata.

Finally: Does it matter if a streamed collection is not necessarily either complete or representative? As long as a researcher makes clear how they conducted a collection and using what tools and configuration, does it not still result in an analysis of behaviour that occurred online? The answer is that it very much depends on the conclusions being drawn. Yes, the collection represents real activity that occurred, but the potential for its incompleteness may cause conclusions drawn from it to be unintentionally misinformed and lacking in nuance. This is especially important for benchmarking efforts. We have seen that variations in collections have an impact on network size and structure. This may result in different community compositions and affect centrality analyses, consequently misleading influential account identification and expected diffusion patterns. A firm understanding of the data and how it was obtained is therefore vital.

4.7 Part Summary

In this Part, our focus has been on **TRQ1**, relating to the information environment, in which our target CIB occurs. We have considered the extent to which we can trust the data we obtain from social media platforms, and explored the effects of variations in data obtained from OSNs on social network analyses on the data, finding pitfalls and determining that care must be taken to avoid them. As long as collection conditions are maintained (including the collection tool), we can have reasonable confidence that the data obtained is, if not complete, at least mostly consistent with the data that someone else might collect under the same circumstances with the same tools, and that certain analytics will be more sensitive to any variations present (e.g., centrality measures). We now turn to examine communication environments vulnerable to CIB in Part II.

Part II

The Danger: Polarisation

On the 20th anniversary of the 9/11 terror attacks, former US President George W. Bush commented on the growing divide amongst people of different political persuasions:

> "A malign force seems at work in our common life that turns every disagreement into an argument, and every argument into a clash of cultures. So much of our politics has become a naked appeal to anger, fear, and resentment," Bush said. "That leaves us worried about our nation and our future together."[15]

His comments implicitly referred to the riots and storming of the US Capitol building on the the 6th of January earlier that year, but the phenomenon is not unique to the United States. Prominent Australian political journalist Leigh Sales recently described how working on social media, which now forms a important part of their duties, is increasingly difficult due to the growing bullying, harassment and accusations of bias from both sides of Australian politics.[16] More recent still is emerging evidence of the abuse leveled at health workers during the global COVID-19 pandemic, particularly at UK COVID advisors.[17]

Though this behaviour is unfortunately not limited to online interactions only,[18] social media provides mechanisms that seem to facilitate and exacerbate the shift to extreme opinions. Apart from anything, isolated radicalised individuals now have easily accessible ways to find like-minded others and radicalise them, resulting in increasing fears for former bastions of democracy.[19] For this reason, an understanding of the mechanisms underlying conflict and particularly ideologically-driven polarisation on social media is vital to inform how it is addressed.

In this Part, to tackle **TRQ2**, we first consider the case study of a Twitter discussion on `#ArsonEmergency` at the height of the Australian "Black Summer" bushfires during the Southern Hemisphere 2019-2020 summer. In Chapter 5, we find two distinct polarised groups arguing over the primary causes of the bushfires and characterise their behaviour, particularly within the context of the broader discussion. Following this, in Chapter 6, across three datasets covering nearly a year we conduct a longitudinal study confirming the presence of the same polarised accounts, as well as accounts polarised over a different, but equally contentious, social issue (same sex marriage). Further, we examine their activities in the datasets, which relate to the bushfires,

[15] https://edition.cnn.com/2021/09/11/politics/george-w-bush-9-11-speech-domestic-violent-extremism/index.html. Posted 2021-09-11. Accessed 2022-01-11.

[16] https://www.abc.net.au/news/2021-09-14/twitter-social-media-bullies-political-journalism/100458714. Posted 2021-09-14. Accessed 2022-01-10.

[17] https://www.theguardian.com/world/2021/dec/31/uk-governments-covid-advisers-enduring-tidal-waves-of-abuse. Posted 2021-12-31. Accessed 2022-01-06.

[18] E.g., anti-vaccine crowd attacked a COVID-19 test and trace centre in the UK. Source: https://www.bbc.com/news/uk-england-beds-bucks-herts-59836172. Posted 2021-12-31. Accessed 2022-01-11.

[19] https://edition.cnn.com/2022/01/09/opinions/canadians-fear-us-democracy-collapse-obeidallah/index.html. Posted 2022-01-10. Accessed 2022-01-11.

federal elections and sport, and consider whether they remain polarised and to what degree. This longitudinal study addresses **TRQ3**.

Chapter 5

#ArsonEmergency: A Case Study of Polarisation

During Australia's unprecedented bushfires in 2019-2020, misinformation blaming arson resurfaced on Twitter using #ArsonEmergency. The extent to which bots were responsible for disseminating and amplifying this misinformation has received scrutiny in the media and academic research. Here we study Twitter communities spreading this misinformation during the population-level event, and investigate the role of online communities and bots. Our in-depth investigation of the dynamics of the discussion uses a phased approach – before and after reporting of bots promoting the hashtag was broadcast by the mainstream media. Though we did not find many bots, the most bot-like accounts were *social bots*, which present as genuine humans. This suggests automated influence remains a concern.

Further, we distilled meaningful quantitative differences between two polarised communities in the Twitter discussion, resulting in the following insights. First, *Supporters* of the arson narrative promoted misinformation by engaging others directly with replies and mentions using hashtags and links to external sources. In response, *Opposers* retweeted fact-based articles and official information. Second, Supporters' were embedded throughout their interaction networks, but Opposers obtained high centrality more efficiently despite their peripheral positions. By the last phase, Opposers and unaffiliated accounts appeared to coordinate, potentially reaching a broader audience. Finally, the introduction of the bot report changed the discussion dynamic: Opposers responded immediately only, while Supporters countered strongly for days, but new unaffiliated accounts drawn in shifted the dominant narrative from arson misinformation to factual and official information. This foiled Supporters' efforts, highlighting the value of exposing misinformation campaigns.

We speculate that the communication strategies observed here could be discoverable in other misinformation-related discussions and could inform counter-strategies.

*This chapter expands upon the material presented in Publication **I** and is under review by Social Network Analysis and Modelling as Publication **VIII**.*

5.1 Introduction

People share an abundance of useful information on social media during crises (Bruns and Liang, 2012; Bruns and Burgess, 2012). This information, if analysed correctly, can rapidly reveal population-level events such as imminent civil unrest, natural disasters, or accidents (Tuke et al., 2020). Not all content is helpful, however: different entities may try to popularise false narratives using sophisticated social bots and/or engaging humans. The spread of such misinformation and disinformation not only makes it difficult for analysts to use Twitter data for public benefit (Nasim et al., 2018) but may also encourage large numbers of people to adopt the false narratives causing social disruption and polarisation, which may then influence public policy and action, and thus can be particularly destabilising during crises (Singer and Brooking, 2019; Kušen and Strembeck, 2020; The Soufan Center, 2021b; Scott, 2021).

This paper expands previous work (Weber et al., 2020a) presenting deeper analysis of a case study of the dynamics of misinformation propagation, and the communities which promote or counter it, during one such crisis. We demonstrate that polarised groups can communicate/use social media in very different ways even when they are discussing the same issue, and in effect these can be considered communication strategies, as they are promoting their narrative and trying to convince others to accept their position.

5.1.1 The "Black Summer" bushfires and misinformation on Twitter

The 2019-2020 Australian 'Black Summer' bushfires (a.k.a., wildfires) burnt over 16 million hectares, destroyed over 3,500 homes, and caused at least 33 human and a billion animal fatalities,[1] and attracted global media attention. During the bushfires, as in other crises, social media provided a mechanism for people in the fire zones to provide on-the-ground reports of what was happening around them, a way for those outside to get insight into the events as they occurred (including authorities and media), but also a way for the broader community to connect and process the imagery and experiences through discussion. The lack of the traditional information mediator or gatekeeper role played by the mainstream media on social media permits factual errors, mis-interpretation and outright bias to proliferate without check in a way it could not in decades past. Our analysis of online discussion at this time shows:

- Significant Twitter discussion activity accompanied the Australian bushfires, influencing media coverage.
- Clearly discernible communities in the discussion had very different interpretations of the ongoing events.

[1] https://www.abc.net.au/news/2020-02-19/australia-bushfires-how-heat-and-drought-created-a-tinderbox/11976134. Posted 2020-02-19. Accessed 2022-01-10.

- In the midst of the discussion, false narratives and misinformation circulated on social media, much of it seen during previous crises, including specific statements that:
  - the bushfires were mostly caused by arson;
  - preventative backburning efforts had been reduced due to green activism (previously presented in 2009[2]);
  - Australia commonly experiences such bushfires (previously put forward in 2013[3]); and
  - climate change is not related to bushfires.

All of these statements and their associated narratives were refuted officially, including via a State government inquiry which found that of 11,744 fires, only "11 were lit with intention to cause a bush fire" (p.29, NSW Bushfire Inquiry, 2020). In particular, the arson figures being disseminated online were incorrect,[4] preventative backburning has increasingly limited effectiveness,[5] its use has not been curbed to appease environmentalists,[6] the fires are "unprecedented",[7] and climate change is, in fact, increasing the frequency and severity of the fires (Jones et al., 2020). The Twitter discussion surrounding the bushfires made use of many hashtags, but according to research by Graham & Keller (Graham and Keller, 2020) reported on ZDNet (Stilgherrian, 2020), the arson narrative was over-represented on `#ArsonEmergency`, likely created as a counter to the pre-existing `#ClimateEmergency` (Barry, 2020). Furthermore, their research indicated that `#ArsonEmergency` was being boosted by bots and trolls. This attracted widespread media attention, with most coverage debunking the arson conspiracy theory.[8] This case thus presents an interesting natural experiment: the nature of the online narrative, and the communities that formed in the related discussions, before the publication of the ZDNet article and then after these conspiracy theories were debunked.

We present an exploratory mixed-method analysis of the Twitter activity using the term 'ArsonEmergency' approximately a week before and after the publication of the ZDNet article (Stilgherrian, 2020), making use of social network analysis (SNA),

[2] https://www.smh.com.au/national/green-ideas-must-take-blame-for-deaths-20090211-84mk.html. Posted 2009-02-12. Accessed 2022-01-10.

[3] https://www.theguardian.com/world/2013/oct/24/greg-hunt-wikipedia-climate-change-bushfires. Posted 2013-10-24. Accessed 2022-01-10.

[4] https://www.abc.net.au/radionational/programs/breakfast/victorian-police-reject-claims-bushfires-started-by-arsonists/11857634. Posted 2020-01-10. Accessed 2022-01-10.

[5] https://www.theguardian.com/australia-news/2020/jan/08/hazard-reduction-is-not-a-panacea-for-bushfire-risk-rfs-boss-says. Posted 2020-01-08. Accessed 2022-01-10.

[6] https://theconversation.com/theres-no-evidence-greenies-block-bushfire-hazard-reduction-but-heres-a-controlled-burn-idea-worth-trying-129350. Posted 2020-01-07. Accessed 2022-01-10.

[7] The Australian Academy of Science's statement: https://www.science.org.au/news-and-events/news-and-media-releases/statement-regarding-australian-bushfires. Posted 2020-01-10. Accessed 2022-01-10.

[8] The BBC's Ros Atkins's video on the matter was one of the most highly shared URLs: https://www.youtube.com/watch?v=aDvmAMsYwNY Posted 2020-01-09. Accessed 2022-01-10.

behavioural and content analyses. Comparisons are made with activity related to another prominent contemporaneous bushfire-related hashtag, `#AustraliaFire`, and a prominent but unrelated hashtag, `#Brexit`. A timeline analysis revealed two points in time that define three phases of activity. SNA of retweeting behaviour identifies two polarised groups of Twitter users: those promoting the arson narrative, and those exposing and arguing against it. These polarised groups, along with the unaffiliated accounts, provide a further lens through which to examine the behaviour observed. Analysis of the networks of different interactions in the data reveal how central these groups became and to what degree they connected to each other and the broader discussion. Content and co-activity analyses highlight how the different groups used hashtags, external articles and other sources to promote their narratives. Finally, an analysis of bot-like behaviour then seeks to replicate Graham & Keller's findings (Graham and Keller, 2020) and explores the most bot-like contributors in detail, including their contribution to the overall discussion.

5.1.2 Contribution

The contribution of this work includes:

1. a relevant focused dataset from Twitter at a critical time period covering two eras in misinformation spread, plus two contemporaneous datasets;
2. insight into the evolution of a misinformation campaign relating to the denial of climate change science and experience in dealing with bushfires;
3. characterisation of different distinct communities active in the discussion with different agendas; and
4. methods and approaches for examining the behaviour and interaction of polarised communities in the context of the broader discussion.

5.1.3 Related work

Research into the use of social media, particularly Twitter, is well-established as a means for authorities to gather information to address specific operational requirements, maintain awareness of the conditions from those experiencing the brunt of the crisis (e.g., those at the fronts of bushfires), as well as a mechanism to get messages and warnings out to many people quickly. It also provides an environment for those in the crisis to maintain connections with those outside, and for those outside to check on friends and loved ones. This area is discussed further in Section 2.4.

The dangers of information disorders are also well known, and are discussed in Section 2.1. Misinformation and rumour can be extremely damaging during times of crisis, let alone organised disinformation efforts. The distrust in authority figures fostered by false information is particularly dangerous at times when the power of a central authority is vital. A result of conflict in the information space is polarisation and the formation of echo chambers, both of which further hamper the cooperation of the public at times when it is most needed. Polarisation is discussed in detail in Section 2.5.

5.1.4 Research questions

We use the following research questions to guide our exploration of Twitter activity over an 18 day period during the 2019-2020 Australian "Black Summer" bushfires:

RQ1 To what extent can online misinformation campaigns be discerned? Are there discernible groups of accounts driving the misinformation, and if so how are they doing it?

RQ2 How did the spread of arson narrative-related misinformation differ between phases, and did the spread of the hashtag `#ArsonEmergency` differ from other emergent discussions (e.g., `#AustraliaFire` and `#Brexit`)?

RQ3 How did the online behaviour of those who prefer the arson narrative differ from those who refute or question it? How was it affected by media coverage exposing how the `#ArsonEmergency` hashtag was being used?

RQ4 How central were the communities to the discussion and how insular were they from each other and the broader discussion?

RQ5 How did the communities make use of retweets, hashtags and URLs to promote their narrative? What evidence is there of coordination?

RQ6 To what degree did the polarised groups receive support from outside Australia?

RQ7 To what degree was the spread of misinformation facilitated or aided by trolls and/or automated bot behaviour engaging in inauthentic behaviour?

In the remainder of this paper, we describe our mixed-method analysis and the datasets used. A timeline analysis is followed by the polarisation analysis. The revealed polarised communities are compared from behavioural and content perspectives, as well as through bot analysis. Answers to the research questions are summarised and we conclude with observations and proposals for further study of polarised communities.

5.2 The Data and the Timeline

The primary dataset was collected over an 18 day period at the height of the bushfires using the term 'ArsonEmergency' (see Table 5.1). For comparison, over the same time period, a second bushfire-related dataset was collected using the search term 'AustraliaFire', along with a non-bushfire-related dataset focused on `#Brexit`.

Broader searches using multiple related terms were not conducted due to time constraints and in the interests of comparison with the original findings (Stilgherrian,

Table 5.1. The datasets were collected from 31 December 2019 to 17 January 2020. Both Twarc and RAPID communicate with Twitter's standard Search and Streaming APIs. *https://github.com/DocNow/twarc

Dataset	Tweets	Accounts	Collection method
Primary			
- ArsonEmergency	27,546	12,872	Twarc* searches on 8, 12, and 17 January
Comparison			
- AustraliaFire	111,966	96,502	Twarc searches on 8 and 17 January
- `#Brexit`	187,792	78,216	Streamed with RAPID (Lim et al., 2019)

2020). Due to the use of Twint[9] in that study, differences in dataset were expected to be likely but minimal. Differences in datasets collected simultaneously with different tools have been previously noted (Chapter 4). Live filtering was also not employed for these bushfire-related collections, as the research started after Graham & Keller's findings were reported.

Twitter may have removed inauthentic content in the time between it being posted and us conducting searches as part of data cleaning routines. For these reasons, some of the content observed by Graham & Keller were expected to be missing from our dataset. This lack of consistency between social media datasets for comparative analyses is a growing challenge recently identified in the benchmarking literature (Publication **IV**).

Tweets by Graham and Keller, whose research was referred to in the ZDNet article (Stilgherrian, 2020) were not removed from the 'ArsonEmergency' dataset, as it was felt their effect was limited. Graham and Keller posted six and three retweets, respectively, all after the ZDNet article was published. As Graham and Keller were mentioned in tweets promoting the ZDNet article and, three days later, the Conversation article by them (Graham and Keller, 2020), their Twitter handles appeared in 106 retweets and 8 tweets posted between the 7th and the 11th of January, peaking on the days the articles were published.

5.2.1 The timeline

This study focuses on about a week of Twitter activity before and after the publication of the ZDNet article (Stilgherrian, 2020). Prior to its publication, the narratives that arson was the primary cause of the bushfires and that fuel load caused the extremity of the blazes were well known in the conservative media (Barry, 2020; Keller et al., 2020). The ZDNet article was published at 6:03am GMT (5:03pm AEST[10]) on 7 January 2020, and was then reported more widely in the mainstream media (MSM) morning news, starting around 13 hours later. We use these temporal markers to define three dataset phases:

- *Phase 1*: Before 6am GMT, 7 January, 2020;

[9] https://github.com/twintproject/twint. Accessed 2022-01-10.
[10] Australian Eastern Standard Time.

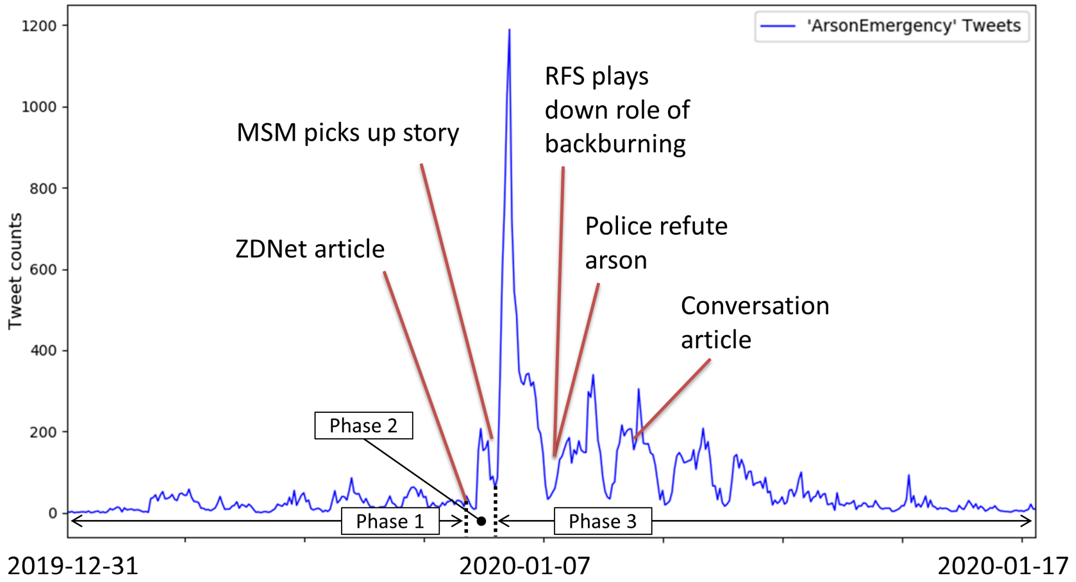


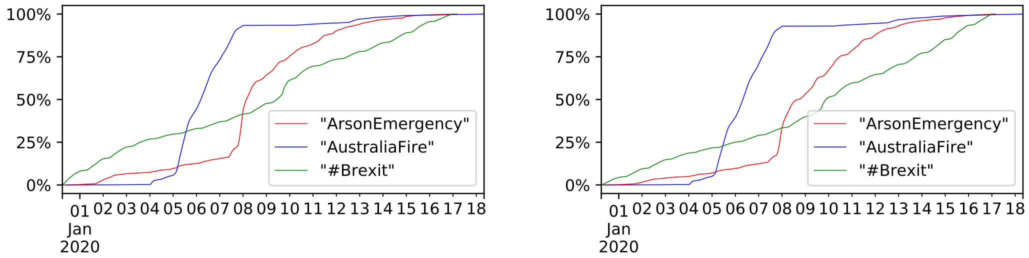
FIGURE 5.1. Tweet activity in the 'ArsonEmergency' dataset, annotated with notable real-world events and the identified phases.

- *Phase 2*: From 6am to 7pm GMT, 7 January, 2020; and
- *Phase 3*: After 7pm GMT, 7 January, 2020.

Figure 5.1 shows the number of tweets posted each hour in the 'ArsonEmergency' dataset, and highlights the phases and notable events including: the publication of the ZDNet article; when the story hit the MSM; the time at which the Rural Fire Service (RFS) and Victorian Police countered the narratives promoted on the #ArsonEmergency hashtag; the publication of the follow-up Conversation article (Graham and Keller, 2020); and the clear subsequent day/night cycle. The RFS and Victorian Police announcements countered the false narratives promoted in political discourse in the days prior.

Since late September 2020, Australian and international media had reported on the bushfires around Australia, including stories and photos drawn directly from social media, as those caught in the fires shared their experiences. No one hashtag had emerged to dominate the online conversation and many were in use, including #AustraliaFires, #ClimateEmergency, #bushfires, and #AustraliaIsBurning.

The use of #ArsonEmergency was limited in Phase 1, with the busiest hour having around 100 tweets, but there was an influx of new accounts in Phase 2. Of all 927 accounts active in Phase 2 (responsible for 1,207 tweets), 824 (88.9%) of them had not posted in Phase 1 (which had 2,061 active accounts). 1,014 (84%) of the tweets in Phase 2 were retweets, more than 60% of which were retweets promoting the ZDNet article and the findings it reported. Closer examination of the timeline revealed that the majority of the discussion occurred between 9pm and 2am AEST, possibly inflated by a single tweet referring to the ZDNet article (at 10:19 GMT), which was retweeted 357 times. In Phase 3, more new accounts joined the conversation, but the day/night



(a) Growth in accounts.

(b) Growth in tweets.

Figure 5.2. The growth of the ArsonEmergency, AustraliaFire, and `#Brexit` datasets in terms of accounts joining the discussion and the tweets posted.

cycle indicates that the majority of discussion was local to Australia (or at least its major timezones).

5.2.2 Growth of the discussions

To consider if the pattern of discussion growth in 'ArsonEmergency' is typical, we compared the discussion with two other contemporary discussions in terms of user growth (i.e., number of new accounts joining the discussion) and tweet growth (Figure 5.2). The similarity in the user and tweet growth lines indicates that as new accounts joined each discussion, they usually only posted a single tweet. The `#Brexit` discussion lacks an intervention event and so its growth is smooth and consistent.[11] In contrast, 'AustraliaFire' discussion appears to be a hashtag campaign instigated by people in Pakistan and Germany resulting in 45k retweets. Many of the retweeting accounts were suspended, so it is possible they were driven by botnets, and the campaign stops growing suddenly after a few days. The 'ArsonEmergency' dataset's growth pattern clearly shows the point of the ZDNet article's appearance, but it continues to grow for several more days after the initial response.

5.2.3 Meta-discussion: Avoiding promotion of the hashtag

The term 'ArsonEmergency' (without '#') was used for the Twarc searches, rather than '`#ArsonEmergency`', to capture tweets that did not include the hashtag symbol but were relevant to the discussion. This was done to capture discussions of the term, in which participants deliberately chose to avoid using the term in a way that would contribute to the hashtag discussion (i.e., by including the hashtag symbol). We refer to this as meta-discussion, i.e., discussion *about* the discussion. We sought to understand how much of the discussion relating to `#ArsonEmergency` was, in fact, meta-discussion. Of the 27,546 tweets in the 'ArsonEmergency' dataset, only 100 did not use it with the '#' symbol (0.36%), and only 34 of the 111,966 'AustraliaFire' tweets did the same (0.03%), so it is clear that very little of the discussion was meta-discussion. That said, there were several days on which tens of tweets seemed to be

[11] This collection occurred in the days prior to the passing of the Withdrawal Agreement Bill, on 22 January, 2020. Source: https://en.wikipedia.org/wiki/Timeline_of_Brexit. Accessed 2022-02-15.

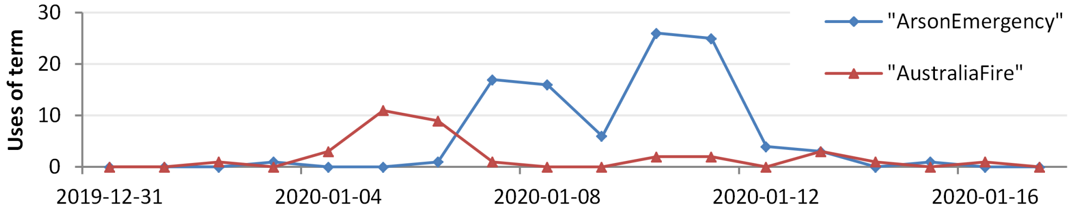


FIGURE 5.3. Counts of tweets using the terms 'ArsonEmergency' and 'AustraliaFire' without a '#' symbol from the period 2–15 January, 2020, in meta-discussion regarding each term's use as a hashtag (counts outside were zero).

involved in meta-discussion, as shown in Figure 5.3. These coincide with Phase 2, when the story reached the MSM, and then again a few days later, possibly as a secondary reaction to the story (commenting on the initial reaction to the story on the MSM).

The small number of uses in the meta-discussion imply that most use of the term 'ArsonEmergency' without the hash or pound symbol is a deliberate, rather than an incidental, part of the discussion. Examination of these particular tweets confirms this; we present examples in Table 5.2.

TABLE 5.2. Examples of meta-discussion referring to the **#ArsonEmergency** hashtag without including it directly by removing or separating the leading '#' character.

Research from QUT shows that 'some kind of a disinformation campaign' is pushing the Twitter hashtag # ArsonEmergency. There is no arson emergency. https://t.co/⟨URL⟩
@⟨ACADEMIC⟩ @⟨JOURNALIST⟩ Venn Diagram of "ArsonEmergency" with "Qanon" and "Agenda21" conspiracies could be interesting ⟨UNIMPRESSED EMOJI⟩
suggest @AFP @NSWpolice ,@Victoriapolice as this misinformation is likely to cause panic & distress in Bushfire hit communties. This link is US news but it contains saliant facts about arrests. https://t.co/⟨URL⟩ When retweeting, remove hashtag from 'arsonemergency' https://t.co/⟨URL⟩
@⟨JOURNALIST⟩ #!ArsonEmergency - a notag.

5.3 Finding and Characterising Polarised Communities

As our aim is to learn about who is promoting the **#ArsonEmergency** and its related misinformation, we first looked to the retweets. Retweets are the primary mechanism for Twitter users to reshare tweets to their own followers. Retweets reproduce a tweet unmodified, except to include an annotation indicating which account retweeted them. There is no agreement on whether retweets imply endorsement or alignment. Metaxas et al. (2015) studied retweeting behaviour in detail by conducting user surveys and studying over 100 relevant papers referring to retweets. Their findings conclude that when users retweet, it indicates interest and agreement as well as trust in not only the message content but also in the originator of the tweet. This opinion is not shared by some celebrities and journalists who put a disclaimer on their profile: "retweets $\neq$ endorsements". Metaxas et al. (2015) also indicated that inclusion of hashtags

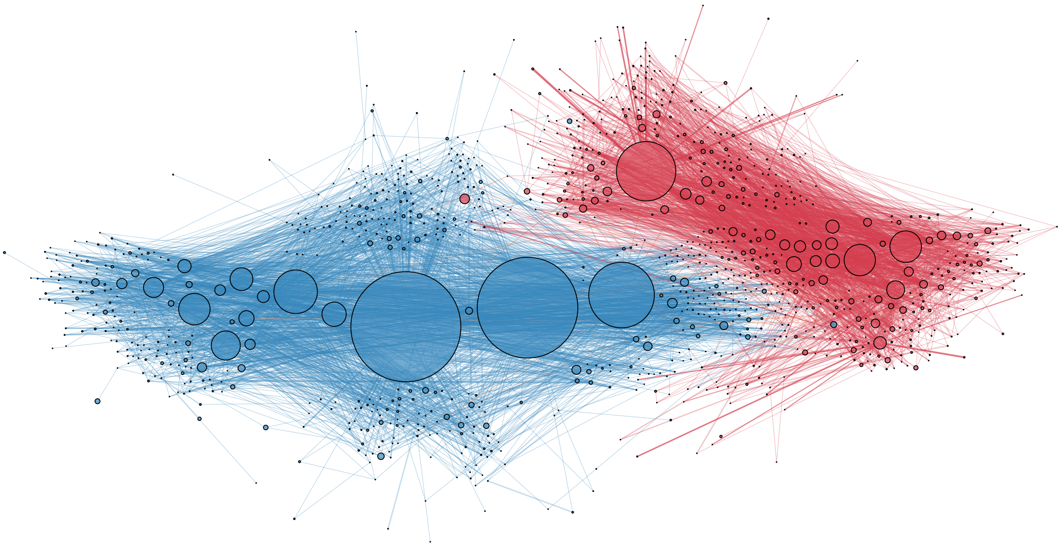


Figure 5.4. Retweet network of the **#ArsonEmergency** discussion showing clear clustering, which, when coupled with examination of each cluster's dominant narrative, provides evidence of

polarisation. On the left in blue is the Opposer community, which countered the arson narrative promoted by the red Supporter community on the right. Nodes represent users. An edge from one node to another means that the account represented by the first node retweeted one of the tweets by the account represented by the second. Node size corresponds to indegree centrality, indicating how often the account was retweeted.

strengthens the agreement, especially for political topics. Other motivations, such as the desire to signal to others to form bonds and manage appearances (Falzon et al., 2017), serve to further imply that even if retweets are not endorsements, we can assume they represent agreement or an appeal to like-mindedness at the very least.

Given the highly connected nature of Twitter data and our aim of exploring human social behaviour, using networks to model our data facilitate social network analysis is a logical step (Brandes and Erlebach, 2005). Using nodes to represent individuals, edges can be used to represent the flow of information and influence and the strength of those connections. We conducted an exploratory analysis on the retweet network built from the 'ArsonEmergency' dataset, which shown in Figure 5.4. The nodes represent Twitter accounts and are sized by indegree (i.e., frequency of being retweeted). An edge between two accounts shows that one retweeted a tweet of the other. Using conductance cutting (Brandes et al., 2008), we discovered two distinct well-connected communities, with a very low number of edges between the two communities. These communities are based on positive edges (likely endorsements) thus appear clear, cohesive and mostly isolated within the context of this discussion, suggesting the formation of echo chambers, depending on their predominant content. Next, we selected the top ten most retweeted accounts from each community, manually checked their profiles and the content they contributed, and hand labelled them as *Supporters* and *Opposers* of the arson narrative accordingly.[12] The accounts have been coloured accordingly in Figure 5.4: the 497 red nodes are accounts that promoted the narrative (the Supporters), while the 593 blue nodes are accounts that opposed them (the Opposers).

The term `#ArsonEmergency` had different connotations for each community. Supporters used the hashtag to reinforce and promote their existing beliefs about climate change, while Opposers used this hashtag to refute the arson theory. The arson theory was a topic on which people held strong opinions resulting in the formation of the two strongly connected communities. Such polarised communities sometimes do not admit much information flow between them, hence members of such communities are repeatedly exposed to similar narratives, which then further strengthens their existing beliefs. Completely closed-off communities are also known as *echo chambers*, and they limit people's information space. The retweets here tend to coalesce within and thus form the communities, as has also been seen in Facebook comments (Nasim et al., 2013).

These two groups, Supporters and Opposers, and those users *Unaffiliated* with either group, are used to frame the remainder of the analysis in this paper.

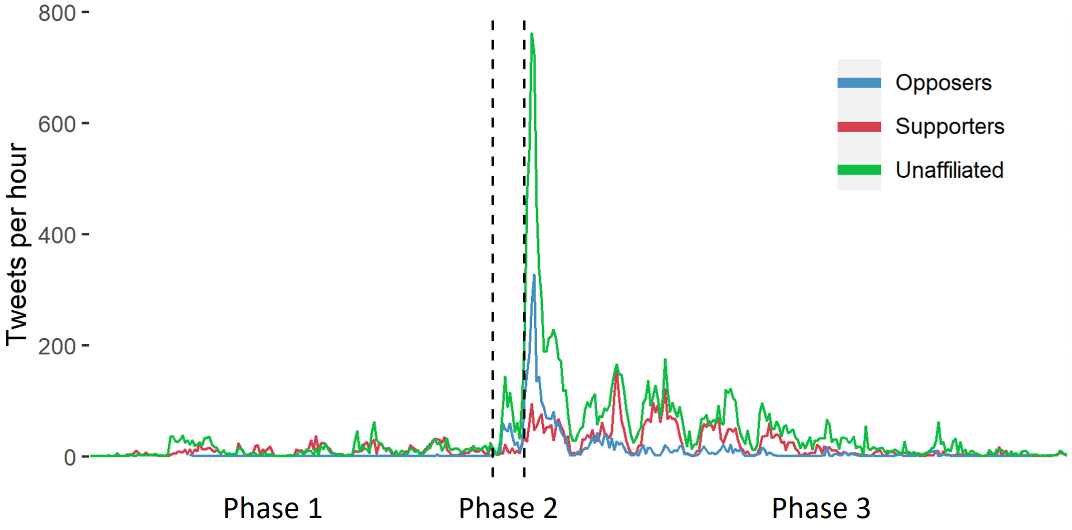


FIGURE 5.5. A timeline of each communities' activity over the collection period.

5.3.1 Community timelines

The relative behaviour of the communities over the collection period, shown in Figure 5.5, informs several key observations. The first is the impact of the story reaching the MSM: the peaks of both Opposer and Unaffiliated contributions is on the morning of Phase 3, immediately after the story appeared on the morning bulletins. Despite the much greater number of Unaffiliated accounts (11,782), their peak is only a little

[12] Labelling was conducted by the first two authors of Publication I independently and then compared.

more than twice that of the 593 Opposer accounts. Unaffiliated and Supporter accounts are active during the entire collection, but Supporters' activity is prominent each day in Phase 3, and peaks on the second day of Phase 3. That peak might have occurred as a response to the previous peak, as by that time the news would have had a full day to disseminate around the world. By reaching a broader audience via the MSM, more Supporter accounts may have been drawn into the online discussion.

Analysis confirms that the composition of the Unaffiliated participants did change across the phases. Of the 1,680 Unaffiliated accounts active in Phase 1, only 30 participated in Phase 2 (of 678 Unaffiliated accounts) and 427 in Phase 3 (of 10,074 Unaffiliated accounts). Furthermore, the contributions of those Phase 1 Unaffiliated accounts in the later phases were not markedly different from the other Unaffiliated accounts, with the Phase 1 accounts contributing 51 (of 759) and 561 (of 14,267) Unaffiliated tweets in Phases 2 and 3, respectively. This suggests that new accounts joined the discussion with similar enthusiasm, and neither the new nor old accounts dominated the Unaffiliated contribution (Figure 5.6).

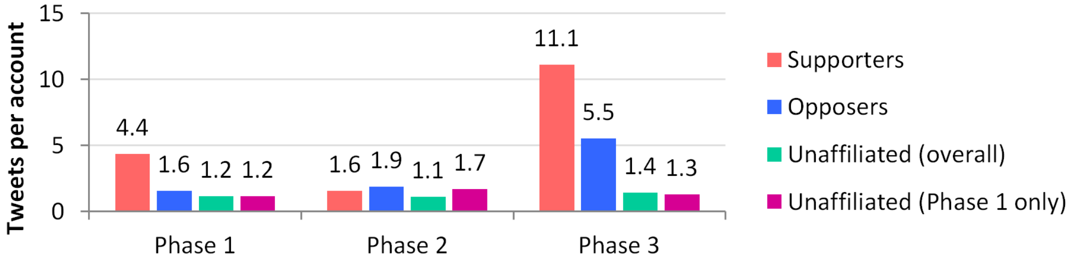
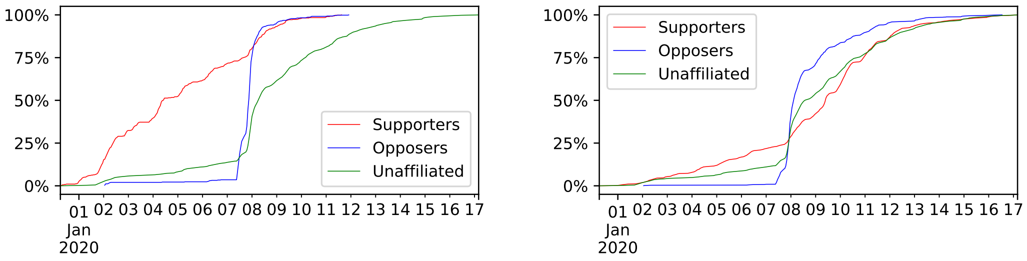


FIGURE 5.6. Tweets per account in each phase for the Supporters, Opposers, Unaffiliated accounts overall, and Unaffiliated accounts active in Phase 1.



(a) Growth in accounts.

(b) Growth in tweets.

FIGURE 5.7. The growth of the ArsonEmergency discussion in terms of accounts joining the discussion and the tweets they posted.

Examining the accumulation of new accounts (Figure 5.7a) and new tweets (Figure 5.7b) shows that #ArsonEmergency was steadily accruing Supporters until the ZDNet article (Stilgherrian, 2020), at which point the community was established and remained active for several days into Phase 3. The Opposer community joined almost entirely in Phase 2, and its activity was mostly confined to that phase, while the Unaffiliated continued to join the discussion well into Phase 3. The publication of the

ZDNet article appears to have drawn in large numbers of Opposers and Unaffiliated, while the Supporter growth immediately plateaued.

Finally, a clear diurnal effect can be seen in Figure 5.5 with daily peaks of activity occurring during Australian daytime hours, implying that the majority of the activity was domestic. Analysis of the 'lang' field in the tweets[13] confirmed that over 99% of tweets used 'en' (English, 90.5%) or 'und' (undefined, 8.7%).

5.3.2 Behaviour

User behaviour on Twitter can be examined through the features used to connect with others and through content. Here we consider how active the different groups were across the phases of the collection, and then how that activity manifested itself in the use of mentions, hashtags, URLs, replies, quotes and retweets.

TABLE 5.3. Activity of the polarised retweeting accounts, by interaction type in phases.

	Group		Tweets	Accounts	Hashtags	Mentions	Quotes	Replies	Retweets	URLs
Phase 1	Supporters	*Raw count*	1,573	360	2,257	1,020	185	356	938	405
		Per account	4.37	–	6.27	2.83	0.51	0.99	2.61	1.13
		Per tweet	–	–	1.43	0.65	0.12	0.23	0.60	0.26
	Opposers	*Raw count*	33	21	100	5	8	2	20	9
		Per account	1.57	–	4.76	0.24	0.38	0.10	0.95	0.43
		Per tweet	–	–	3.03	0.15	0.24	0.06	0.61	0.27
Phase 2	Supporters	*Raw count*	121	77	226	64	11	29	74	24
		Per account	1.57	–	2.94	0.83	0.14	0.38	0.96	0.31
		Per tweet	–	–	1.87	0.53	0.09	0.24	0.61	0.20
	Opposers	*Raw count*	327	172	266	34	7	14	288	31
		Per account	1.90	–	1.55	0.20	0.04	0.08	1.67	0.18
		Per tweet	–	–	0.81	0.10	0.02	0.04	0.88	0.09
Phase 3	Supporters	*Raw count*	5,278	474	7,414	2,685	593	1,159	3,212	936
		Per account	11.14	–	15.64	5.66	1.25	2.45	6.78	1.97
		Per tweet	–	–	1.40	0.51	0.11	0.22	0.61	0.18
	Opposers	*Raw count*	3,227	585	3,997	243	124	95	2,876	359
		Per account	5.52	–	6.83	0.42	0.21	0.16	4.92	0.61
		Per tweet	–	–	1.24	0.08	0.04	0.03	0.89	0.11
Overall	Supporters	*Raw count*	6,972	497	9,897	3,769	789	1,544	4,224	1,365
		Per account	14.03	–	19.91	7.58	1.59	3.11	8.50	2.75
		Per tweet	–	–	1.42	0.54	0.11	0.22	0.61	0.20
	Opposers	*Raw count*	3,587	593	4,363	282	139	111	3,184	399
		Per account	6.05	–	7.36	0.48	0.23	0.19	5.37	0.67
		Per tweet	–	–	1.22	0.08	0.04	0.03	0.89	0.11
	Unaffiliated	*Raw count*	16,987	11,782	22,192	3,474	615	1,377	14,119	1,790
		Per account	1.44	–	1.88	0.29	0.05	0.12	1.20	0.15
		Per tweet	–	–	1.31	0.20	0.04	0.08	0.83	0.11

In Phase 1, Supporters used **#ArsonEmergency** nearly fifty times more often than Opposers (2,086 to 43), which accords with Graham & Keller's findings that the false narratives were significantly more prevalent on that hashtag compared with others in

[13] The 'lang' field is automatically populated by Twitter based on language detection. If insufficient content is available (e.g., the tweet is empty, or only contains URLs or mentions, 'und' is used to mean 'undefined'.

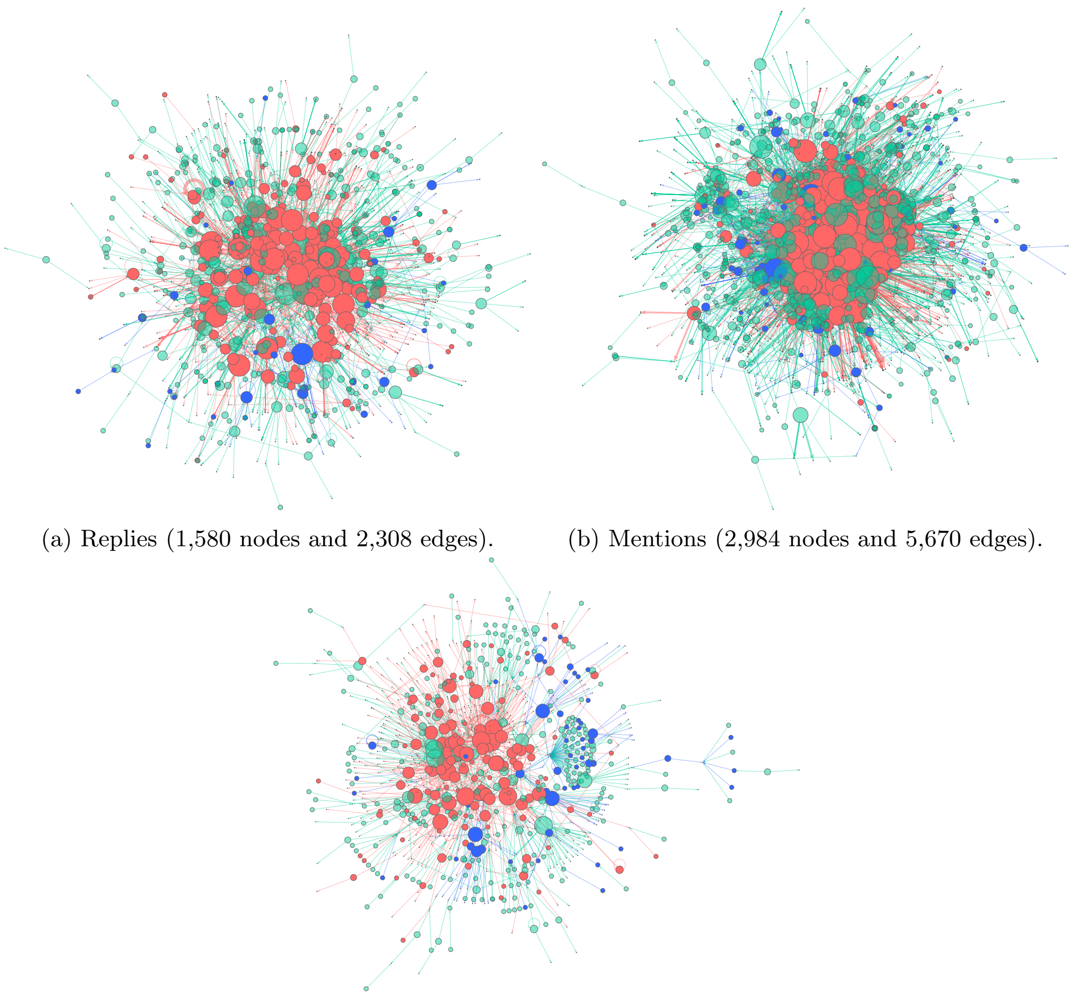
use at the time (Stilgherrian, 2020; Graham and Keller, 2020). This use is roughly proportional to the number of tweets posted by the two groups, however (Table 5.3). Overall in that Phase, Supporters used 22 times as many hashtags as Opposers. In Phase 2, during the Australian night, Opposers countered with three times as many tweets as Supporters, including fewer hashtags, more retweets, and half the number of replies, demonstrating different behaviour to Supporters, which actively used hashtags in conversations. Manual inspection and content analysis confirmed this to be the case. This is evidence that Supporters wanted to promote the hashtag as a way to promote the narrative. Interestingly, Supporters, having been relatively quiet in Phase 2, responded strongly, producing 64% more tweets in Phase 3 than Opposers. They used proportionately more of all interactions except retweeting, including many more replies, quotes, and tweets spreading the narrative with multiple hashtags, URLs and mentions. In short, Opposers tended to rely more on retweets, while Supporters engaged directly and were more active in the longer phases.

Overall, as shown in the bottom section of Table 5.3, Supporter accounts tweeted much more often than other accounts, and used more hashtags, mentions, quotes, replies and URLs, but retweeted less often than both Opposers and Unaffiliated accounts. This suggests that Supporters were generating their own content (not just retweeting it), and attempting to engage with others through the use of platform features, implying a high degree of motivation on their part.

5.3.2.1 Other interaction networks

If Supporters employed a variety of interaction mechanisms, while Opposers relied primarily on retweeting, then Supporters should be deeply embedded in networks constructed from those other interaction mechanisms. This is exactly what we find when we examine the largest components of networks constructed from replies (Figure 5.8a), mentions (Figure 5.8b), and quotes (Figure 5.8c). These largest components include 77.4%, 92.0%, and 72.2% of the reply, mention and quote networks' nodes, respectively. Supporters had more connections (correspondingly represented by larger nodes) and are clearly more active than Opposers using these interactions, engaging with each other and others in the network. They are particularly tightly and centrally clustered in the mention network, which is a reflection of their attempts to actively engage directly (rather than only indirectly, such as with hashtags). They are more diffusely located in the reply network, and the quote network, sharing similar network positions to Unaffiliated accounts. This is less to do with the amount of activity (i.e., the number of replies or tweets) and more to do with how they connect with others. The Opposer accounts that appear in the networks are not as centrally located nor as tightly clustered.

To provide a more objective analysis of the structural properties of these networks and the accounts within them, we employ a variety of centrality measures (discussed in Section 3.2.1.4) and k-core analysis (discussed in Section 3.2.1.3). We also use



(a) Replies (1,580 nodes and 2,308 edges).

(b) Mentions (2,984 nodes and 5,670 edges).

(c) Quotes (915 nodes and 1,230 edges).

FIGURE 5.8. The largest connected components from directed, weighted networks built from the replies, mentions, and quotes, linking from one account to another when it replied, mentioned, or quoted the other. Edges are sized by weight, indicating the frequency of connections, and coloured by source node affiliation. Thicker edges have greater weight. Nodes are sized by outdegree (indicating the replies, mentions and quotes they used) and coloured by affiliation: red nodes are Supporters, blue are Opposers, and green are Unaffiliated.

a variation of Krackhardt and Stern's E-I index (Krackhardt and Stern, 1988) as a measure of homophily (discussed in Section 3.2.1.6).

Centrality. Though the locations of Supporter and Opposer accounts in the networks in Figure 5.8 give the impression that Supporters are more central in each network, the statistics presented in Table 5.4 facilitate a more nuanced interpretation (see Section 3.2.1.4 for their definitions). In the reply, mention and quote networks, Supporters and Opposers make up only a small fraction of the overall networks (shown as a percentage in the Nodes column). Supporter betweenness scores are much higher than Opposers' in the reply and mention networks and even twice as high in the quote network (though still very low). Closeness scores are more weighted towards the Opposers, implying that even though they are not centrally positioned, they remain directly linked to more of the network than the Supporters. The mean degree centrality of Supporters is again higher than Opposers' for all networks, reflecting their tendency to directly reach out to a wider audience than Opposers, who relied mostly on retweets to disseminate their message. The eigenvector centrality scores are higher for Opposers in the reply and mention networks, suggesting they are more connected to important nodes in the network and perhaps were more efficient at selecting their interaction targets, while their lower scores for the quotes network is probably reflective of the fact they used them a lot less (139 uses to Supporters' 789). The centrality scores suggest that the Opposers were less centrally located, but well connected, while Supporters were more centrally positioned (reflected in their relatively high betweenness scores).

Table 5.4. Mean centrality scores for Supporter and Opposer nodes in the largest components of the reply, mention, and quote networks, omitting Unaffiliated node scores.

				Centrality			
Network	Group	Nodes		Betweenness	Closeness	Degree	Eigenvector
Replies	Supporters	231	(14.6%)	0.000181	0.002871	0.004551	0.001307
	Opposers	82	(5.2%)	0.000019	0.003453	0.002757	0.001811
Mentions	Supporters	284	(9.6%)	0.000304	0.005525	0.004207	0.006575
	Opposers	140	(4.7%)	0.000018	0.005067	0.001997	0.006625
Quotes	Supporters	169	(18.5%)	0.000012	0.001876	0.006170	0.016033
	Opposers	80	(8.7%)	0.000005	0.003334	0.004171	0.007302

***k*-core analysis.** The question of how tightly clustered the nodes are can be addressed with *k*-core analysis. This analysis progressively breaks a network down to sets of nodes that have at least *k* neighbours, so nodes on the periphery are discarded first, while highly connected nodes form the 'core' of the network. The result is that the higher the *k*-core for a particular node (i.e., the highest *k*-core of which they are a member), the more embedded in the network they are (see Section 3.2.1.3 for more detail). Figure 5.9 shows the proportions of each groups' members (of those present in each network) in each core. We can immediately see that across all networks, more Supporters have higher *k*-core values than both Opposers and the Unaffiliated. In fact, while the majority of Opposers and Unaffiliated are on the periphery of the networks, Supporters are relatively evenly spread throughout the networks' cores. This implies more of the Supporters were more active in reaching out to many other accounts, something that is also reflected in their higher use of mentions, replies and quotes per account, as shown in Table 5.3.

Homophily measures. The homophily measures (introduced in Section 3.2.1.6) provide an indication of how insular the groups were with their interactions, and here we also apply them to the retweet network for comparison (Table 5.5). Within the retweet network, both communities were highly insular, retweeting in-group accounts almost exclusively, both when considering only the polarised groups and the broader

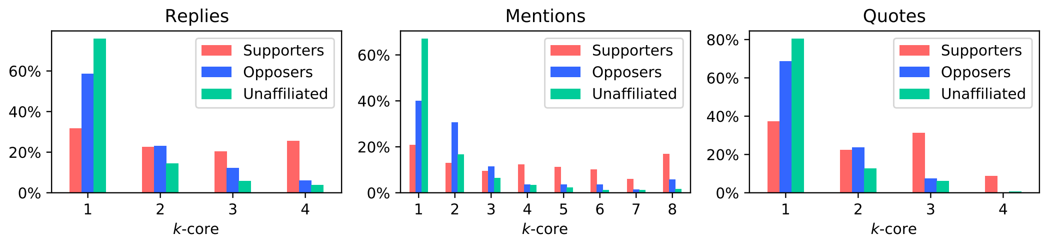


FIGURE 5.9. The distributions of k-core values for accounts in the reply, mention and quote networks. Nodes with higher k-core values are more deeply embedded in their network. The percentage refers to the proportion of each group's accounts with a given k-core value.

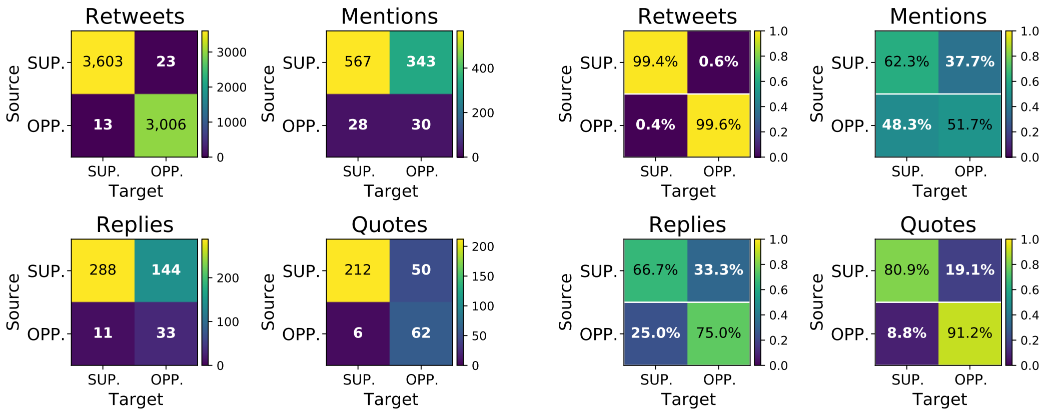
network. Insularity among the other interactions distinguished the groups. Though preferring in-group connections, Supporters engaged more with Opposers than vice versa, when considering just the two polarised groups, but both connected with the broader network much more than in-group members, with Supporters leading the outreach in replies and quotes, while Opposers mentioned others more. Examining the mixing matrix of raw interaction counts in Figure 5.10a emphasises the lower numbers of Opposer interactions and while the Opposer numbers were low, they very strongly preferred to reply and quote their own members. Other than when using mentions, Supporters clearly interacted with Opposers and Unaffiliated accounts more. Given Supporters opinions aligned with conservative politics (certainly with conservative news media, as we shall see later), this finding seems to go against other studies of political polarisation in which conservative-aligned groups have been observed to isolate themselves (Boutyline and Willer, 2016). Perhaps this is an indication that the Supporters were different from other conservative-aligned communities, in that their goal was less about simply discussing shared conservative opinions and more about promulgating a message and convincing others (i.e., outsiders) of their narrative (i.e., more proselytising than conversation).

TABLE 5.5. Homophily measures calculated with just Supporters and Opposers and then all nodes within interaction networks. Edge totals are the sums of the edge weights.

Network	**Polarised groups only**					**Broader network**			
	Nodes		Edges	E-I Index		Nodes	Edges	E-I Index	
	Supporters	Opposers	Total	Supporters	Opposers	All	Total	Supporters	Opposers
Retweet	493	592	6,645	-0.98731	-0.99139	12,076	21,526	-0.70961	-0.88997
Reply	247	105	476	-0.33333	-0.50000	2,041	3,031	0.62030	0.40541
Mention	288	149	968	-0.24615	-0.03448	3,206	7,523	0.69557	0.78723
Quote	190	104	330	-0.61832	-0.82353	1,268	1,542	0.45501	0.10791

These strong pointers to polarisation across both groups raise the question of whether there is a difference between the groups: for each group, what proportion of their ties are in-group or out-group? Figure 5.10 shows two mixing matrix representations of the interactions between Supporters and Opposers. The first (Figure 5.10a) shows the raw numbers of retweets, mentions, replies and quotes from a member of one group (the *source* of the interaction) to another account (the *target* of the interaction).

The second (Figure 5.10b) shows the proportion of interactions from each source that is directed to Supporters or Opposers, effectively presenting a normalised view of the source group's interactive behaviour. It is immediately clear that, outside of retweets, Supporters were much more active, and were biased towards connecting to members of their own group. The degree of activity is notable, because there were fewer Supporters (497) than Opposers (593) though their numbers were similar. Opposers were also heavily biased to connect to other Opposers via replies and quotes, but not so for mentions. The proportional view makes clear the bias in connectivity: while raw numbers of interactions may be low from Opposers, they strongly preferred to connect to themselves, while Supporter bias is less pronounced for mentions, replies and quotes, despite the raw numbers of interactions being much higher.



(a) Raw counts of outgoing connections.

(b) Proportion of outgoing connections.

FIGURE 5.10. The mixing matrices of Supporter (SUP.) and Opposer (OPP.) interactions (Newman, 2003). Figure 5.10a shows raw counts of interactions, while Figure 5.10b shows the proportions of interactions from each source group to each target group.

5.3.2.2 The concentration of voices

The concentration of narrative from certain voices requires attention. To consider this, Table 5.6 shows the degree to which accounts were retweeted by the different groups by phase and overall. Unaffiliated accounts relied on a smaller pool of accounts to retweet than both Supporters and Opposers in each phase and overall, which is reasonable to expect as the majority of Unaffiliated activity occurred in Phase 3, once the story reached the mainstream news, and therefore had access to tweets about the story from the media and prominent commentators. Of the top 41 accounts that were retweeted, each of which was retweeted 100 times or more in the dataset, 17 were Supporters and 20 Opposers. Supporters were retweeted 5,487 times (322.8 retweets per account), while Opposers were retweeted 8,833 times (441.7 times per account). Together, affiliated accounts contributed 93.3% of the top 41's 15,350 retweets, in a dataset with 21,526 retweets overall, and the top 41 accounts were retweeted far more often than most. This pattern was also apparent in the 25 accounts most retweeted by Unaffiliated accounts in Phase 3 (accounts retweeted at least 100 times): 8 were

TABLE 5.6. Retweeting activity in the dataset, by phase and group.

	Supporters			Opposers			Unaffiliated		
Phase	Retweets	Retweeted Accounts	Retweets per account	Retweets	Retweeted Accounts	Retweets per account	Retweets	Retweeted Accounts	Retweets per account
1	938	77	12.182	20	8	2.500	1,659	105	15.800
2	74	21	3.524	288	31	9.290	652	60	10.867
3	3,212	290	11.076	2,876	228	12.614	11,807	532	22.194
Overall	4,224	327	12.917	3,184	243	13.103	14,118	613	23.030

Supporters and 14 were Opposers. Thus Supporters and Opposers made up the majority of the most retweeted accounts, and arguably influenced the discussion more than Unaffiliated accounts.

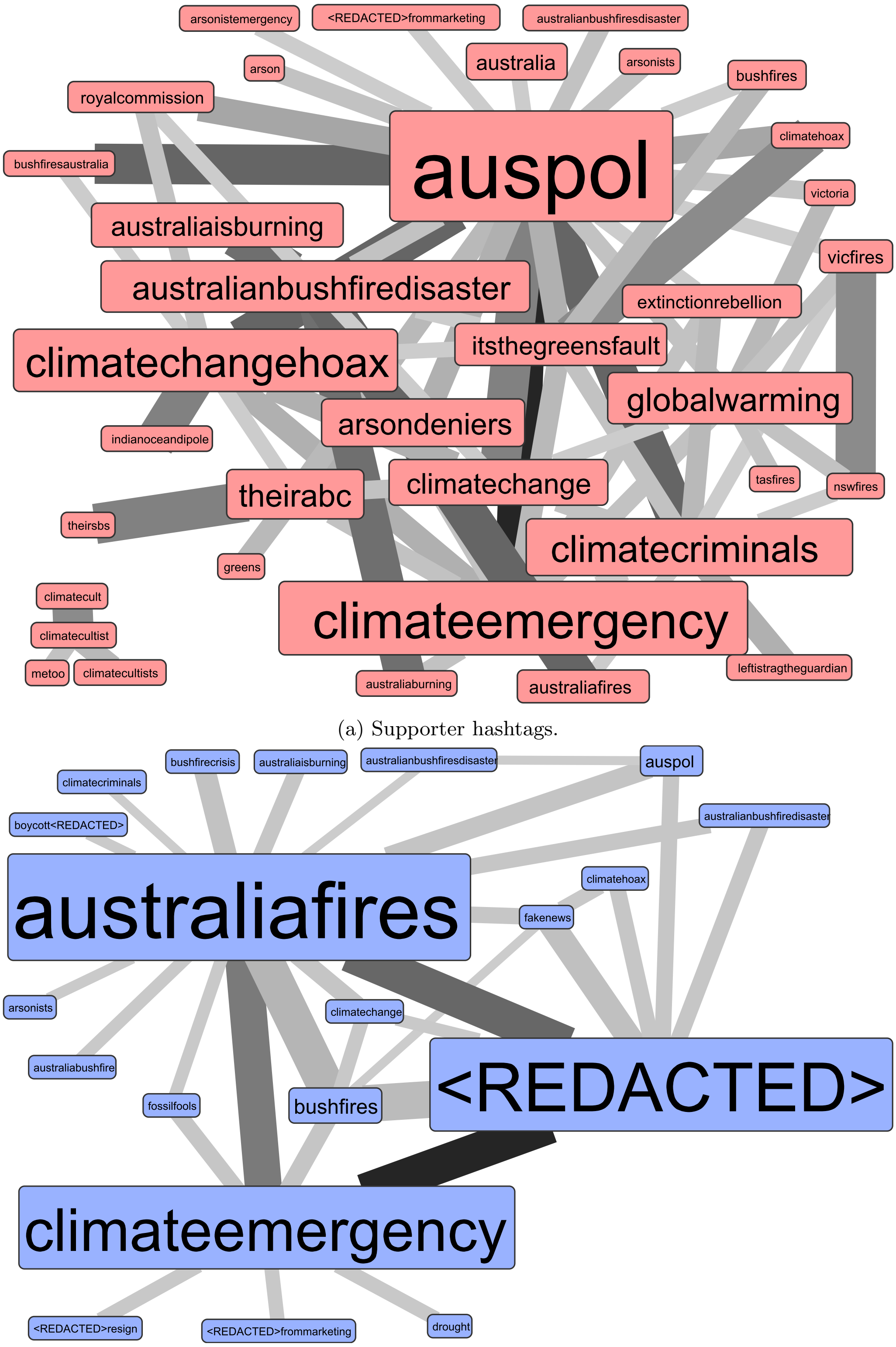
5.3.3 Content dissemination

When contrasting the content of the two affiliated groups, we considered the hashtags and external URLs used. A hashtag can provide a proxy for a tweet's topic, and an external URL can refer a tweet's reader to further information relevant to the tweet, and therefore tweets that use the same URLs and hashtags can be considered related.

5.3.3.1 Hashtags

To discover *how* hashtags were used, rather than simply *which* were used, we developed semantic networks (visualised in Figure 5.11). In these networks: each node is a hashtag in its lower case form, sized by degree centrality; edges represent an account using both hashtags (not necessarily in the same tweet); and the edge weight represents the number of such accounts in the dataset. Nodes are coloured according to the affiliation of the accounts that used them. We removed the #ArsonEmergency hashtag (as nearly each tweet in the dataset contained it) as well as edges having weight less than 5. Opposers used a smaller set of hashtags, predominantly linking #AustraliaFires with #ClimateEmergency[14] and a hashtag referring to a well-known media owner. In contrast, Supporters used many hashtags in a variety of combinations, mostly focusing on terms related to 'fire', but only a few with 'arson' or 'hoax', and linking to #auspol and #ClimateEmergency. Manual inspection of Supporter tweets revealed many containing only a string of hashtags, but these were rare in the Opposer tweets. Notably, the #ClimateChangeHoax node has a similar degree to the #ClimateChangeEmergency node, indicating Supporters' skepticism of climate science, but perhaps also that Supporters were attempting to join or merge the discussion communities defined by those hashtags in order to pollute the predominant hashtag of the #ClimateChangeEmergency community with a counter-narrative (Conover et al., 2011; Woolley, 2016; Nasim et al., 2018). This fits with the evidence found by Graham and Keller (2020), indicating that #ArsonEmergency was deliberately created to challenge climate change-related hashtags.

[14] Capitals are re-introduced to hashtags used in the discussion for readability.



(a) Supporter hashtags.

(b) Opposer hashtags.

Figure 5.11. Semantic network of hashtags of Supporters and Opposers. Hashtag nodes are linked when five or more accounts tweeted both hashtags, and are coloured by the affiliation of the accounts that used them. <REDACTED> hashtags include identifying information. (To aid interpretation, the large redacted node refers to a media owner, while the smaller ones refer to politicians.) Heavy edges (with high weight) are thicker and darker. The hashtag #ArsonEmergency has been removed from each network, as it occurred in every tweet in the dataset.

Even though Supporters used approximately the same number of hashtags per tweet as Opposers (2.92 compared with 2.89), they used 40.9 hashtags per account, including 1.30 unique hashtags per account. In contrast, Opposers only used 17.5 hashtags per account, including 0.36 unique ones. This indicates the pool of hashtags used by the Opposers was much smaller than that of Supporters. The distribution of hashtag uses for the ten most frequently used by each group (which overlap but are not identical), omitting the ever-present `#ArsonEmergency`, is shown in Figure 5.12. It indicates that Opposers focused slightly more strongly on a small set of hashtags, while Supporters spread their use of hashtags over a broader range (and thus their use of even their most frequently used hashtags is less than for Opposers). Unaffiliated accounts used their frequently used hashtags more often than both groups by the 4th hashtag, possibly due to the much greater number of accounts being active but less focused in their hashtag use. A second hashtag appeared in fewer than 20% of each groups' tweets.

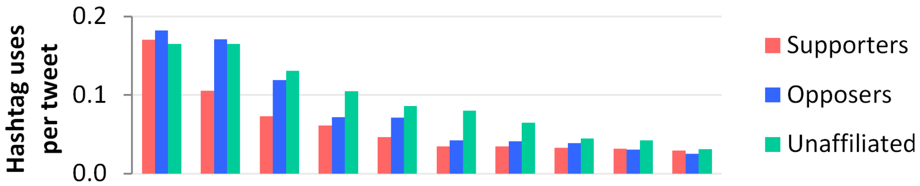
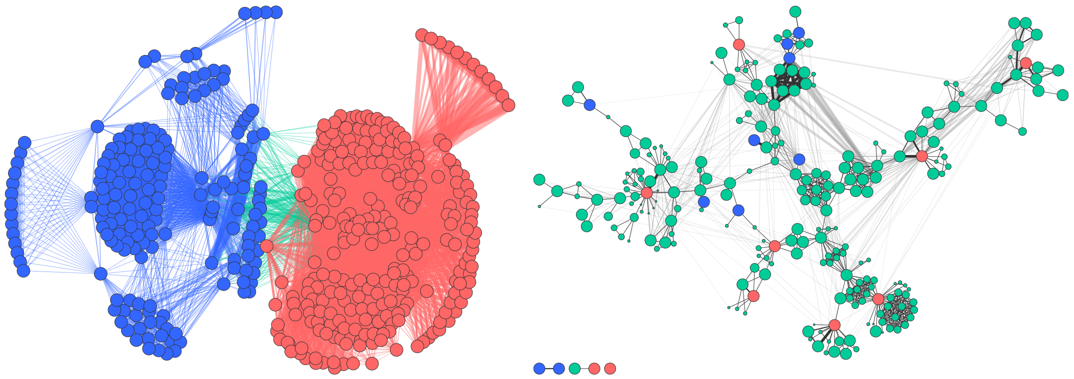


FIGURE 5.12. Hashtag uses per tweet for the ten most used hashtags for Supporters, Opposers and the Unaffiliated, omitting `#ArsonEmergency`. Opposers used hashtags more frequently than Supporters, but after the second hashtag, Unaffiliated accounts used more than either polarised group.

Manual inspection of Supporter tweets revealed that many replies consisted solely of "`#ArsonEmergency`" (e.g., one Supporter replied to an Opposer 26 times in under 9 minutes with a tweet just consisting of the hashtag). This kind of behaviour, in addition to inflammatory language in other Supporter replies, suggests a degree of aggression, though aggressive language was also noted among Opposers. The tweets that included more than 5 hashtags made up only 1.7% of Opposer tweets, but 2.8% of Supporter tweets and 2.1% Unaffiliated tweets. Further analysis of inauthentic behaviour is addressed in Section 5.4.3, and further analysis of the change in hashtags over phases can be found in Appendix A.3.

Polarisation in hashtag use. A statistical examination of how Supporters and Opposers used hashtags also revealed significant levels of homophily when considering only Supporters and Opposers, but less so when the hashtags use of Unaffiliated accounts was included. The statistics were obtained from co-hashtag account networks (networks of accounts associated by hashtag use, described in Section 3.2.2.5). Not all hashtags were used by each group, however. In order to determine to what extent their hashtag use overlapped without the influence of widely used hashtags (which connect the majority of accounts in the network), we created a set of hashtags to focus on beginning with the ten most frequently used hashtags unique to each of



(a) Accounts using partisan and co-occurring hashtags.

(b) Partisan and co-occurring hashtags.

Figure 5.13. Two networks built from the tweets containing 'partisan' hashtags (minus the ten most common hashtags). Left: Figure 5.13a is a network of accounts, linked when they mention the same hashtag. Red nodes are Supporters, while blue ones are Opposers, and edges are coloured according to their endvertices (green edges span the groups). Edge width represents edge weight, and isolates have been removed. Right: Figure 5.13b is a network of hashtags, linked when they are used by the same account. Blue and red nodes represent the Opposer- and Supporter-specific partisan hashtags, respectively. Green nodes are co-occurring hashtags. Nodes are laid out with the backbone algorithm (Serrano et al., 2009; Nocaj et al., 2014), and edges are shaded by backbone strength. The small components were joined only via the removed common hashtags.

the Supporters and Opposers. Using this set of twenty hashtags,[15] we extracted the tweets containing them and created the account network from all the hashtag uses they included (i.e., including the co-occurring hashtags). We then removed uses of the ten most frequently occurring hashtags to produce a final set of 245 hashtags.

Figures 5.13a and 5.13b are visualisations of the account network and the hashtag network, respectively. Though some polarisation should be expected given the partisan hashtags provide a natural axis of polarisation, in the account network it is notable quite how little overlap there is in the use of the co-occurring hashtags. The clusters apparent in the account network (Figure 5.13a) are caused by the fact that partisan hashtags are rarely mentioned by the same account (Figure 5.13b). Instead they are clearly used with a variety of distinct hashtags, implying that although Supporters and Opposers were polarised in their hashtag use, they also had distinct sub-communities within their discussions (using hashtags as a proxy for discussion topic).

The resulting account network consists of 12,867 nodes (including the 493 Supporter and 597 Opposer accounts) and 424,389 edges. The combined modified E-I Index of the network, which only considers edges internal and external to Supporters and Opposers rather than also including edges between Unaffiliated accounts, was 0.250, implying that together the groups expressed a small but solid preference for outside

[15] Supporters' ten most frequently used unique hashtags were `#ItsTheGreensFault` (326), `#Victoria` (123), `#GlobalWarming` (107), `#TheirABC` (78), `#ClimateCultist` (66), `#IndianOceanDipole` (66), `#Greens` (62), `#ecoterrorism` (54), `#Melbourne` (53), and `#NotMyABC` (52), while those of the Opposers were `#BlackSummer` (74), `#FossilFools` (34), `#KoalasNotCoal` (29), `#ArsonHoax` (28), `#ArsonMyArse` (23), `#DontGetDerailed` (23), `#Smoko` (20), `#bots` (19), `#ArseholeEmergency` (17), and `#FossilFuel` (14).

connections (i.e., due to the co-occurring hashtags). When considered separately, the Supporters' E-I Index is 0.147 while the Opposers' is 0.717, suggesting that Supporters, though they used the hashtags of others, did so much less so than Opposers. When we consider only the 114,797 edges between or within the Supporter and Opposer groups (excluding all edges to adjacent Unaffiliated accounts), the modified E-I Index fall to −0.991 and −0.883 for Supporters and Opposers, respectively, which indicates the great majority of such edges were homophilic (i.e., within groups). Given we started with hashtags unique to each group, a degree of homophily is not surprising, however these very strong results imply that not many of the co-occurring hashtags each group used overlapped either. These results are clearly evident in a visualisation of the network (Figure 5.13a).

Quickly returning to the network of hashtags mentioned in the partisan tweets, we can see the clusters in the account network (Figure 5.13a) are caused by the fact that the accounts rarely used multiple partisan hashtags together (otherwise there would be clusters of partisan hashtags); instead, whenever a tweet included a partisan hashtag, they also included one or a few of a variety of non-partisan hashtags, which are represented by clusters of green nodes in Figure 5.13b.

5.3.3.2 External URLs

URLs in tweets can be categorised as *internal* or *external*. Internal URLs refer to other tweets in retweets or quotes, while external URLs are often included to highlight something about their content, e.g., as a source to support a claim. By analysing the URLs, it is possible to gauge the intent of the tweet's author by considering the reputation of the source or the argument offered.

We categorised[16] the ten URLs used most each by the Supporters, Opposers, and Unaffiliated accounts across the three phases, and found a significant difference between the groups. URLs were assigned to one of these four categories:

NARRATIVE Articles used to emphasise the conspiracy narratives by prominently reporting arson figures and fuel load discussions.

CONSPIRACY Articles and web sites that take extreme positions on climate change (typically arguing against predominant scientific opinion).

DEBUNKING News articles providing authoritative information about the bushfires and related misinformation on social media.

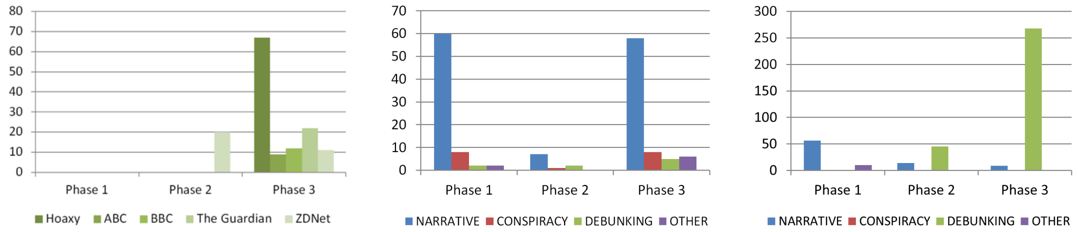
OTHER Other web pages.

URLs posted by Opposers were concentrated in Phase 3 and were all in the DEBUNKING category, with nearly half attributed to Indiana University's Hoaxy service (Shao et al., 2016), and nearly a quarter referring to the original ZDNet article (Stilgherrian, 2020) (Figure 5.14a). In contrast, Supporters used many URLs in Phases 1 and 3,

[16] Categorisation was conducted by two authors of Publication **I** and confirmed by the others.

focusing mostly on articles emphasising the arson narrative, but with references to a number of climate change denial or right wing blogs and news sites (Figure 5.14b).

Figure 5.14c shows that the media coverage changed the content of the Unaffiliated discussion, from articles emphasising the arson narratives in Phase 1 to Opposer-aligned articles in Phase 3. Although the activity of Supporters in Phase 3 increased significantly, the Unaffiliated members appeared to refer to Opposer-aligned external URLs much more often. This suggests that the new Unaffiliated accounts arriving in the final phase (discussed in Section 5.3.1 above) held different opinions on the arson narrative from the Unaffiliated accounts active early in the discussion. In fact, it is possible they acted as bridges bringing in new Opposer accounts – 411 of the 585, or approximately 70% of Opposer accounts active in Phase 3 were were not active in earlier Phases.



(a) Opposer URLs. (b) Supporter URLs. (c) Unaffiliated URLs.

FIGURE 5.14. URLs used by Opposers, Supporters and Unaffiliated accounts.

Supporters used many more URLs than Opposers overall (1,365 to 399) and nearly twice as many external URLs (390 to 212). Supporters seemed to use many different URLs in Phase 3 and overall, but focused much more on particular URLs in Phase 1. Of the total number of unique URLs used in Phase 3 and overall, 263 and 390, respectively, only 77 (29.3%) and 132 (33.8%) appeared in the top ten, implying a wide variety of URLs were used. In contrast, in Phase 1, 72 of 117 appeared in the top ten (61.5%), similar to Opposers' 141 of 212 (66.5%), implying a greater focus on specific sources of information. In brief, it appears Opposers overall and Supporters in Phase 1 were focused in their choice of sources, but by Phase 3, Supporters had expanded their range considerably. Ultimately, Supporters used 195 URLs 390 times (in total), Opposers used 68 URLs 212 times, and the Unaffiliated used 305 URLs 817 times, meaning a mean rate of use of 2.0, 3.1, and 2.7, respectively, meaning Opposers were more focused in their URL use. This is evident in the distributions of URL uses in Figure 5.15, which Supporters use more URLs more often that Opposers, and Opposers focused many of their uses on a small number of URLs.

5.3.4 Coordinated amplification

To investigate whether coordinated dissemination or amplification of content was occurring, we performed co-retweet, co-hashtag and co-URL analyses using the technique we present in detail in Chapter 7. These analyses reveal sub-communities of accounts that retweet the same tweets, and share the same hashtags, URLs, and URL domains

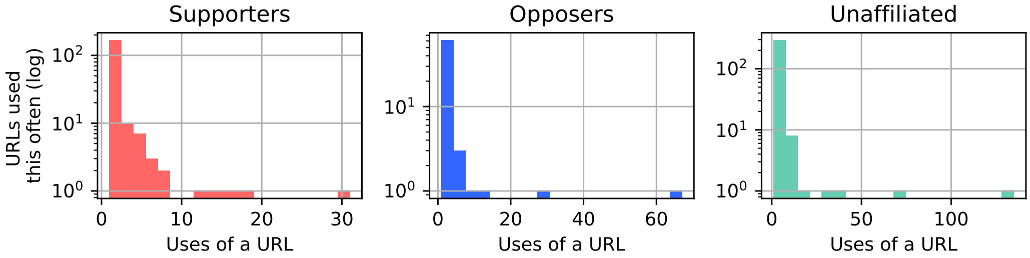


FIGURE 5.15. Distributions of URL use by Supporters, Opposers and Unaffiliated accounts.

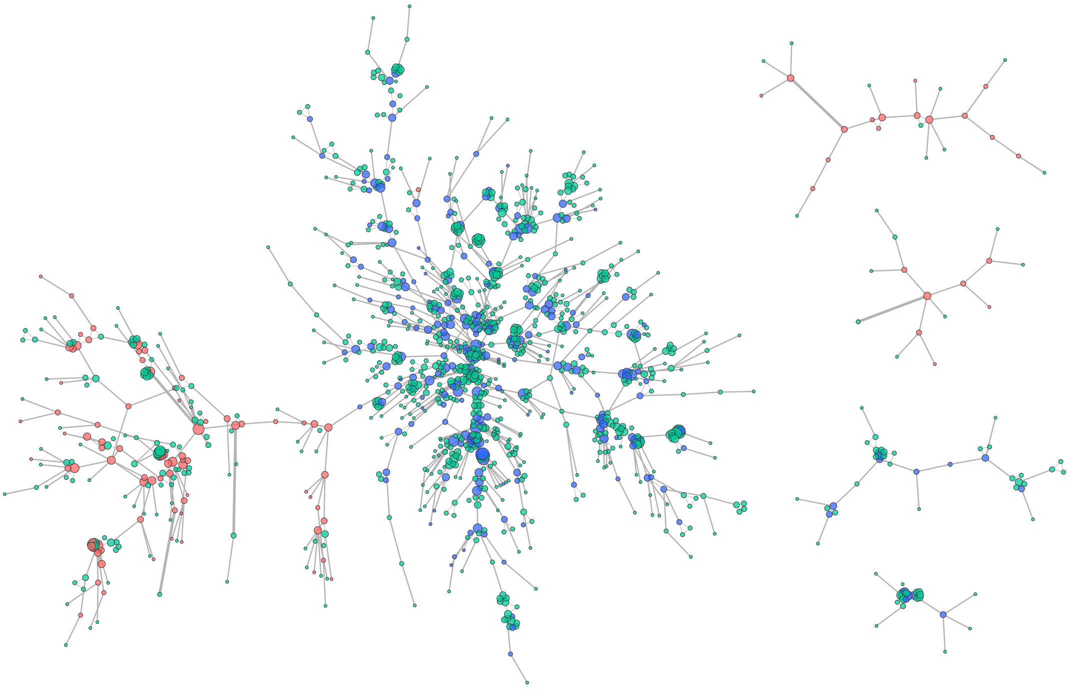


FIGURE 5.16. The five largest connected components of the co-retweet coordination network (γ=1 minute), limited to only Supporter and Opposer accounts, which are sized by indegree. Red nodes are Supporters, blue are Opposers, and edges are sized by frequency of co-retweeting.

within the same timeframes (denoted by γ). Regarding the URLs, Figure 5.14 indicates the nature of article external links referred to, but not the distributions of the URLs or their domains, which is the aim of using these co-activity analyses. The analyses result in weighted networks consisting of the sub-communities as disconnected components of accounts, the edge weights of which indicate the frequency of co-linking or co-mentioning of a hashtag. Further, to examine how the sub-communities relate to one another, we can then re-introduce the URLs and domains as explicit 'reason' nodes into these networks, making them 2-layer networks in which communities are joined according to these 'reason' nodes (see Section 3.2.2.4).

Co-retweet analysis. The largest components of the co-retweet network (γ=1 minute) shown in Figure 5.16 show that the polarisation observed in the retweet network (Figure 5.4) is still evident, as expected, but what is particularly notable is

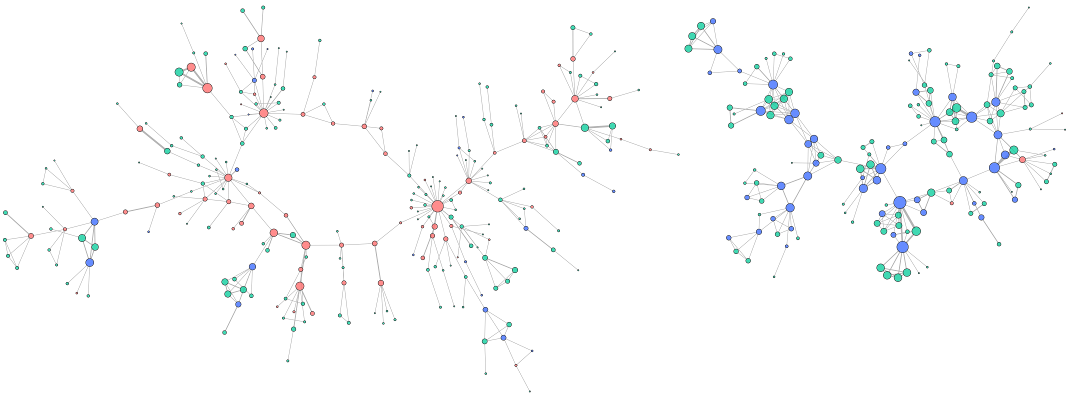


FIGURE 5.17. The two largest connected components of the co-hashtag coordination network (γ=1 minute, excluding `#ArsonEmergency`), with nodes sized by the number of tweets they posted in the discussion. Red nodes are Supporters, blue are Opposers, green are Unaffiliated, and edge widths are sized by the frequency of co-hashtag activity.

the absence of tight cliques amongst the Supporter nodes, which, as promoters of the arson narrative, were originally thought to include a large proportion of bots (Stilgherrian, 2020; Graham and Keller, 2020). Cliques would indicate accounts all retweeting the same tweets within the same timeframe, a signal associated with automation, but also with high popularity (i.e., increasing the number of interested accounts increases the chance that they co-retweet accidentally). Cliques are visible amongst the 103 Opposers and many of the 966 Unaffiliated accounts (and could also be due to simple popularity and coincidence), but rare amongst the 233 Supporters. Instead their connection patterns imply real people seeing and retweeting each others retweets. For example, account A sees a tweet and retweets it, which is then seen by account B (within 1 minute), and then account C sees that and retweets it as well, but longer than 1 minute after A. A 1-minute window is quite large for the purposes of identifying botnets, so this would indicate a lack of evidence of retweeting bots amongst the Supporters.

A further item to note is the degree of support offered by the Unaffiliated accounts, which co-retweet with Opposer accounts far more frequently than Supporter accounts in the coordination networks presented in Figure 5.16. This observation raises the question of whether some of the Unaffiliated accounts may, in fact, be Opposers, but were simply not captured in the application of conductance cutting community detection to the retweet network, and they may have been captured with modification of the detection parameters.

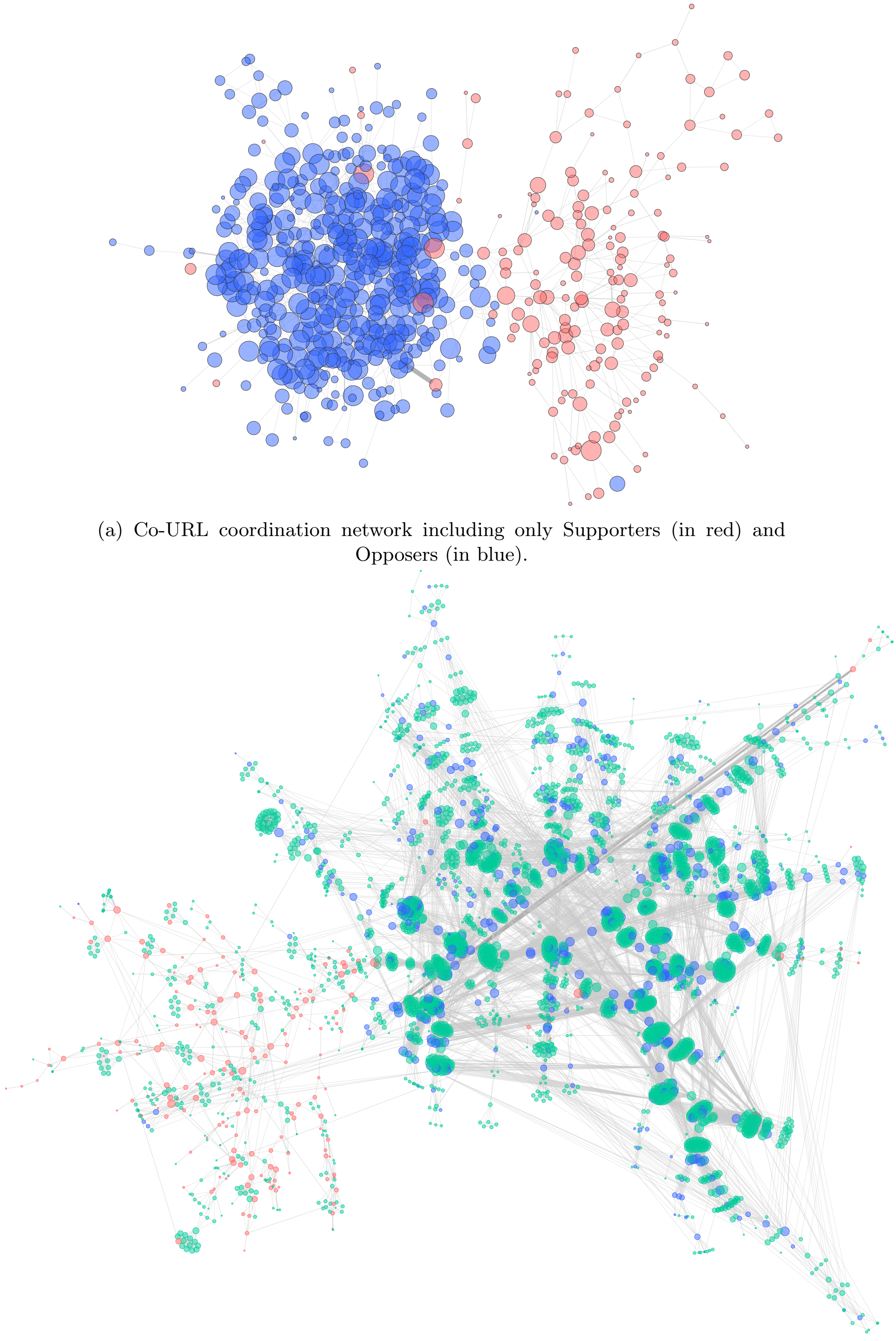
Co-hashtag analysis. As using a hashtag in a tweet can increase its reach to observers of the hashtag as well as one's followers, coordinated promotion of a hashtag is a mechanism to disseminate one's message (Varol et al., 2017b), as well as pollute a discussion space (Woolley, 2016; Nasim et al., 2018). Given how frequently hashtags are used, we chose a tight timeframe of 1 minute and excluded `#ArsonEmergency` from our co-hashtag analysis. The two largest components discovered highlight the polarisation between the Supporter and Opposer communities (Figure 5.17). The ring formation amongst the Supporters and small node sizes indicate less activity including a wider variety of hashtags. Opposers are more active and focused in the hashtags they used. These findings emphasise the findings in Section 5.3.3.1 but also highlight the support of Unaffiliated accounts, the most active of which appear to support the Opposers.

Co-URL and co-domain analysis. For human users, grassroots-style coordinated co-linking should be visible in 'human' timeframes, such as within 10 minutes, allowing time for users to see each others' tweets. The polarisation evident in the retweet network is also evident in the co-linking networks (γ=10 minutes) shown in Figure 5.18, especially considering only the Supporter and Opposer networks (Figure 5.18a). When we examine the co-linking in context in Figure 5.18b, along with the contributions of Unaffiliated accounts, we can see that, again, Unaffiliated accounts co-acted with Opposer accounts far more often than Supporters, which appear relatively isolated, compared with the concentrated co-linking in the Opposer/Unaffiliated clusters on the right. Here, cliques represent groups of accounts sharing the same URLs, but it is unclear whether each clique represents a different URL or simply a different time window. To consider that, we need to introduce 'reason' nodes, representing the shared URLs, to create account/URL 2-layer networks.

Figure 5.19 shows the resulting account/URL 2-layer network, which includes annotations indicating the websites hosting the most shared articles (referred to by the URLs). As expected, there is clear polarisation around the URLs, but it is immediately also clear how focused the Opposer accounts were on a small number of URLs, similar to their use of hashtags. The blue Opposer nodes link mostly to three URLs: the original ZDNet article (Stilgherrian, 2020), the Hoaxy website (Shao et al., 2018b), and an article on The Guardian relating to online misinformation during the bushfires.[17] The Supporter community's use of URLs is more dispersed, and includes MSM sites with the addition of a large cluster of Supporters and Unaffiliated accounts around an article on The Daily Chrenk, the website of an Australian blogger promoting the arson narrative. It is notable that two Australian Broadcasting Corporation (ABC) articles are so centrally located amongst the Supporters, as these were classified as DEBUNKING articles. When we consider the co-domain 2-layer network (Figure 5.20), however, it is clear that the ABC domain binds the polarised Supporter and Opposer communities together, along with, interestingly, The Guardian and the URL shortener `bit.ly`. One `bit.ly` link appeared much more frequently than others, and it resolved to a Spanish news article on online bushfire misinformation.[18] Highlighted in the co-domain 2-layer network are two zones of domains that appear mostly linked to one or the other of the Supporter and Opposer nodes, which are, again, appear polarised in the network. The domains in these zones appear aligned

[17] https://www.theguardian.com/australia-news/2020/jan/08/twitter-bots-trolls-australian-bushfires-social-media-disinformation-campaign-false-claims. Posted 2020-01-08. Accessed 2022-01-10.

[18] https://www.muyinteresante.es/naturaleza/articulo/actualidad-las-fake-news-de-los-incendios-de-australia Posted 2020-01-13. Accessed 2021-01-10.



(a) Co-URL coordination network including only Supporters (in red) and Opposers (in blue).

(b) Co-URL coordination network using the backbone layout.

FIGURE 5.18. The coordination networks resulting from co-URL analysis (γ=10 minutes), with nodes sized by indegree. Red circular nodes are Supporters, blue are Opposers, and the green remainder are Unaffiliated accounts. Edge width shows co-linking frequency.

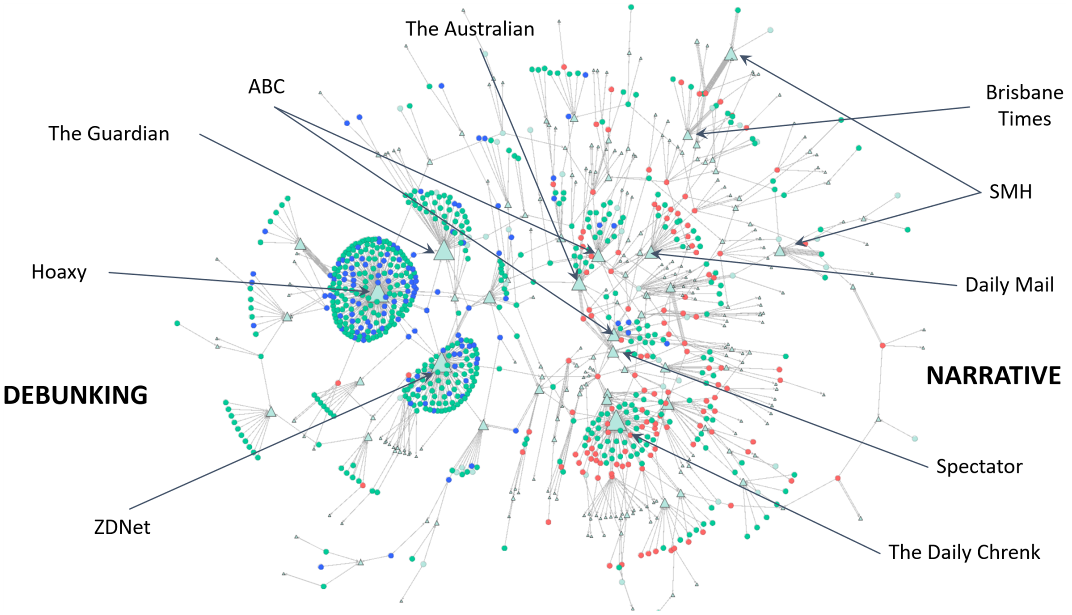


FIGURE 5.19. The account/URL 2-layer network resulting from co-URL analysis (γ=10 seconds), annotated with the websites hosting highly shared articles. Pale green triangular nodes are the URLs, sized by indegree. Red circular nodes are Supporters, blue are Opposers, and the green remainder are Unaffiliated accounts. The most widely shared articles are annotated with the website on which they are hosted (NB, ABC = Australian Broadcasting Corporation, SMH = Sydney Morning Herald). Blue annotated articles are categorised as DEBUNKING, while red ones are categorised as supporting or prominently discussing the 'arson' NARRATIVE.

again with Opposers referring to domains hosting DEBUNKING URLs and Supporters referring to domains hosting NARRATIVE URLs. A few domains are referred to very frequently by individual nodes (visible as dark, large edges), and these are often social media sites, such as YouTube, Instagram, and Facebook.

The analyses of a variety of co-activities here emphasises the polarisation observed in the retweet network permeates the groups' collaborative efforts. Evidence indicates that Opposers, much less so than Supporters, engaged in coordinated action, however, given the significant contribution of Unaffiliated accounts, it is unclear whether this is deliberate or merely a reflection of high popularity (especially given the considerably greater number of Unaffiliated accounts active in the discussion).

5.3.5 Locations

Given the global effect of climate change, any prominent contentious discussion of it is likely to draw in participants from other timezones. Although the activity patterns in Figure 5.1 indicate the majority of activity aligns with Australian timezones, a deeper analysis of the self-reported account 'location' fields in tweets revealed that only 88% of active[19] participants were Australian (Figure 5.21). (Tweets can contain geolocation information but rarely do: only 127 tweets in the 'ArsonEmergency' dataset had any

[19] We considered all Supporters, Opposers, plus all Unaffiliated accounts that tweeted at least three times, and who populated the field.

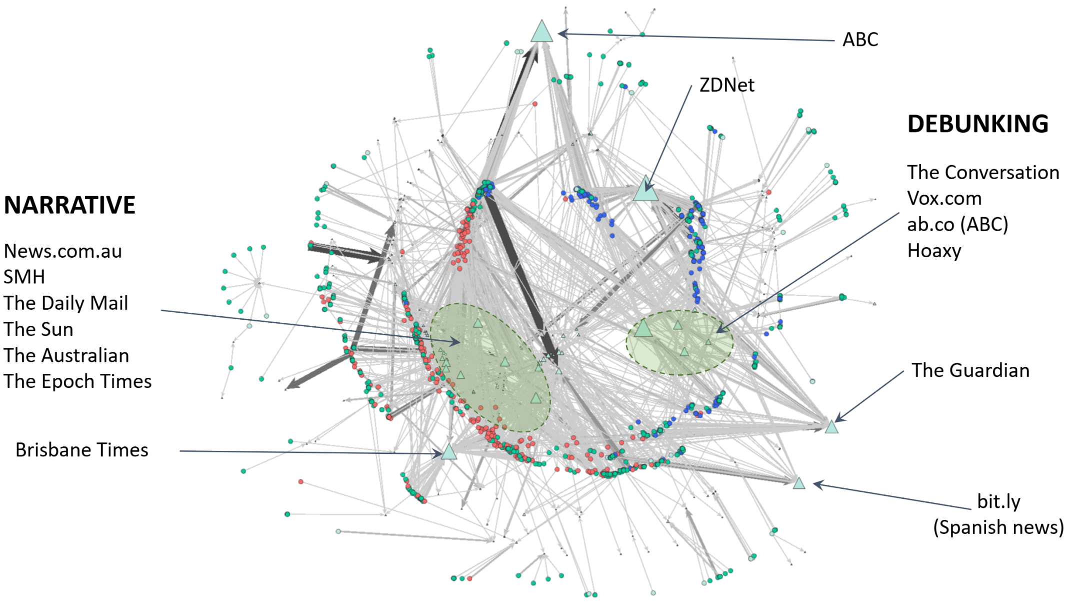


Figure 5.20. The account/domain 2-layer network resulting from co-domain analysis (γ=10 seconds), annotated with the websites hosting highly shared articles. Pale green triangular nodes are the URL domains, sized by indegree. Red circular nodes are Supporters, blue are Opposers, and the green remainder are Unaffiliated accounts. Two zones of contrasting highly linked to domains are highlighted, one primarily used to support the arson narrative, and one used primarily to debunk it (NB, ABC and ab.co = Australian Broadcasting Corporation, SMH = Sydney Morning Herald, News.com.au = News Corporation). The red zone includes a number of DEBUNKING domains and is mostly referred to by Supporters while the blue zone includes academic and centre and left wing domains categorised as DEBUNKING domains, which are referred to predominantly by Opposers.

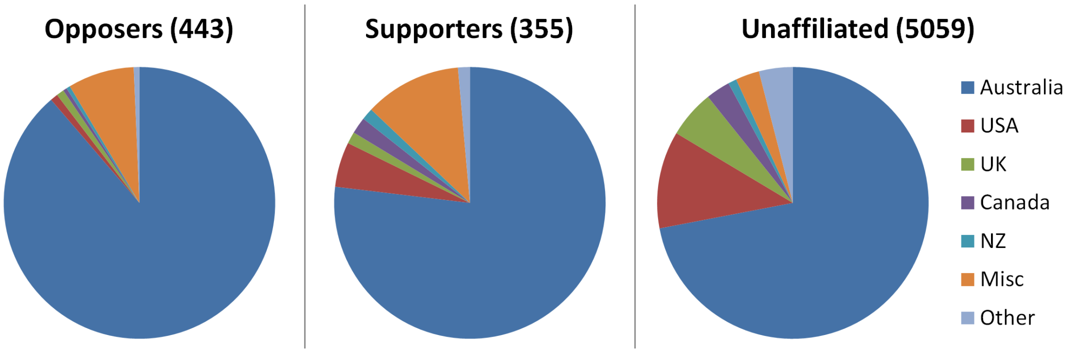


Figure 5.21. The self-reported locations of Supporter, Opposer and Unaffiliated accounts. The number in brackets indicates how many accounts were evaluated. The Miscellaneous category was used for locations which described a physical location but were vague, e.g., Earth, whereas Other was used for whimsical entries, e.g., "Wherever your smartphone is." or "Spot X".

geolocation information, and 114 were posted in Australia.) Based on the self-reported location, more Supporters declared locations outside Australia (23%) than Opposers (11%), but the biggest proportion of non-Australian participants were Unaffiliated, perhaps drawn in by the international news. It is unclear whether the international accounts were drawn in to aid the Supporters or Opposers in Phase 3, but we know the articles the Unaffiliated shared changed to DEBUNKING in that Phase, and

Unaffiliated accounts appeared to coordinate with Opposers.

More detail can be found in Appendix A.1.

5.4 Inauthentic Behaviour Analysis

Inauthentic behaviour has a variety of expressions, as discussed in Section 2.2, including the use of automation by bots, especially social bots that present themselves as human to influence others, but also human-initiated behaviour, such as trolling and hatespeech through sockpuppet accounts. Here we examine the contribution of bots to the #ArsonEmergency discussion, and the frequency of inauthentic patterns of text in tweets. When reaching across the divide between communities, out-group interactions need not always be positive.

5.4.1 Botness analysis

The results reported in ZDNet (Stilgherrian, 2020) indicated widespread bot-like behaviour exposed by analysis with the tweetbotornot[20] R library. Our analysis had two goals: 1) attempt to replicate Graham & Keller's findings in Phase 1 of our dataset; and 2) examine the contribution of bot-like accounts detected in Phase 1 in the other phases. Specifically, we considered the questions:

- Does another bot detection system find similar levels of bot-like behaviour?
- Does the behaviour of any bots from Phase 1 change in Phases 2 and 3?

We evaluated 2,512 or 19.5% of the accounts in the dataset using Botometer (Davis et al., 2016), including all Supporter and Opposer accounts, plus all Unaffiliated accounts that posted at least three tweets either side of Graham & Keller's analysis reaching the MSM (i.e., the start of Phase 3).

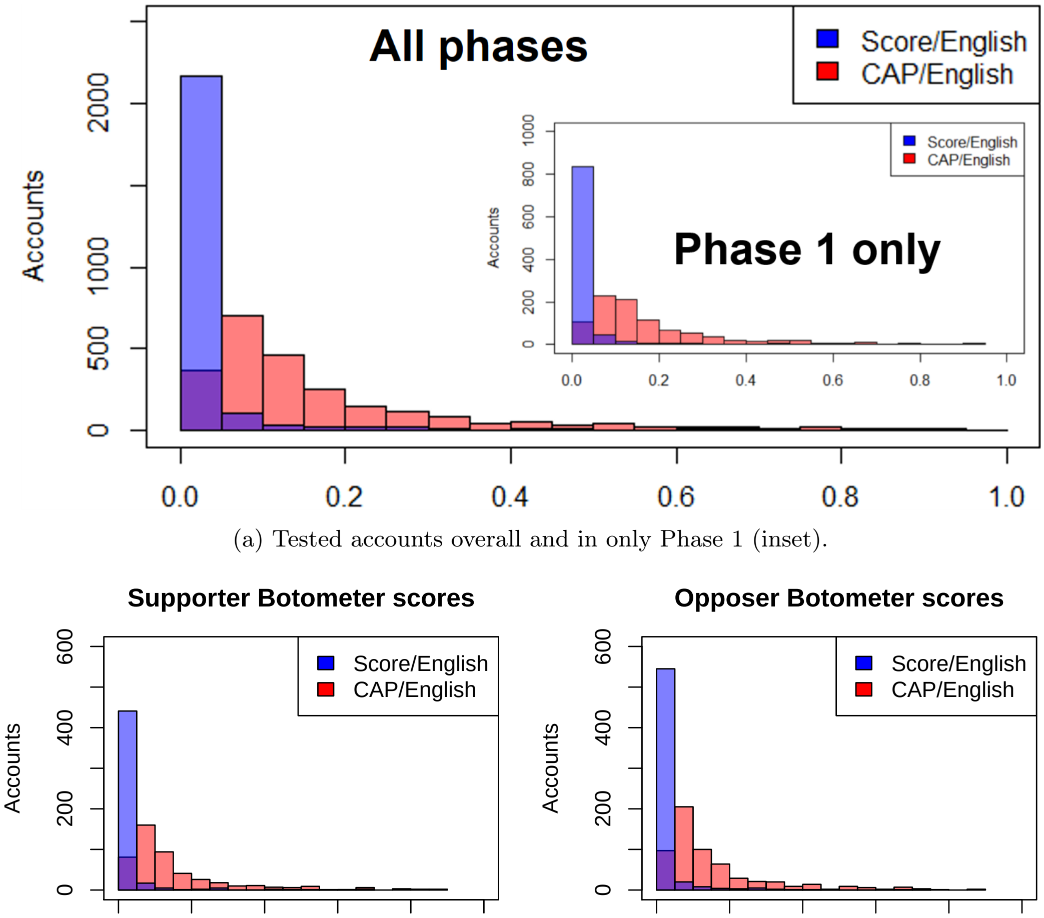
Botometer (Davis et al., 2016) is an ensemble bot classifier for Twitter accounts, relying on over a thousand features drawn from six categories, which provides a structured analysis report of an account, rating various of its features for 'botness'. The report includes a "Complete Automation Probability" (CAP), a Bayesian-informed probability that the account in question is "fully automated", as well as a rating that assumes an account is English-speaking which is different from the language-agnostic rating. This does not accommodate hybrid accounts (Grimme et al., 2018) and only uses English training data (Nasim et al., 2018), leading some researchers to use conservative ranges of CAP scores for high confidence that an account is human (< 0.2) or bot (> 0.6) (e.g., Rizoiu et al., 2018). We adopt that categorisation.

Table 5.7 shows that the majority of accounts were human and contributed more than any automated or potentially automated accounts. The distributions of English and CAP scores for all tested accounts overall and only in Phase 1, when few Opposers

[20] https://github.com/mkearney/tweetbotornot. Accessed 2022-01-10.

Table 5.7. Botness scores and contribution to the discussion across the phases by a subset of the accounts.

			Accounts active			Tweets contributed		
Category	CAP	Total	Phase 1	Phase 2	Phase 3	Phase 1	Phase 2	Phase 3
Human	0.0–0.2	2,426	898	438	1,931	2,213	674	11,700
Undecided	0.2–0.6	66	20	6	56	28	11	304
Bot	0.6–1.0	20	9	4	11	23	6	84



(a) Tested accounts overall and in only Phase 1 (inset).

(b) Supporter accounts.

(c) Opposer accounts.

Figure 5.22. The distribution of Botometer scores. The scores presented are the English score and the Complete Automation Probability, and all are heavily skewed towards low values (i.e., non-automated). The bars are semi-transparent, affecting their colour, to account for their overlap.

were active, and separately for Supporters and Opposers are shown in Figure 5.22. There were no significant differences between the ratings of the tested accounts overall and in Phase 1, nor between Supporters and Opposers, likely due to the dominance of the skew towards human accounts.

5.4.1.1 Discrepancies with the ZDNet results

The analysis in (Stilgherrian, 2020) suggested hundreds of bots were active on #ArsonEmergency, however the results presented here indicate far fewer were present, and they were similarly distributed across the phased and within the polarised groups. The contrast between these results and those reported (Stilgherrian, 2020) is likely to be due to a number of reasons, but the primary one is differences in our datasets. Graham and Keller used the collection tool Twint (which avoids using the Twitter API and instead uses the Twitter web user interface (UI) directly) to focus on results from Twitter's web UI when searching for #ArsonEmergency. Only 812 tweets appeared in both datasets, and even those were restricted to Phase 1. Of the 315 accounts in common, 100 were Supporters and 5 Opposers, implying that those Supporter accounts had already been flagged by misinformation researchers as having previously engaged in questionable behaviour. The size of our dataset and the greater number of accounts we tested is likely to have skewed our Botometer results towards typical users. There are also differences between the bot analysis tools. Botometer's CAP score is focused on non-hybrid, English accounts, whereas tweetbotornot may provide a more general score, taking into account troll-like behaviour. The content and behaviour analysis discussed above certainly indicates Supporters engaged more with replies and quotes, consistent with other observed trolling behaviour (Kumar et al., 2018) or "sincere activists" (Starbird and Wilson, 2020). Follow-up work by Graham and Keller's research group has focused on such "activists" and the contribution of trolls (Graham and Keller, 2020), finding that they appeared to coordinate their activities with prominent public figures and media outlets as part of a broader and longer-running disinformation campaign spanning the months surrounding the period we have focused on (Keller et al., 2020).

As our collection was performed via the Twitter Search API, rather than its Streaming API, and the first of those searches was on the 8th of January, it is possible, if not likely, that Twitter had already stripped some bots and their content from their data holdings. Furthermore, it is unclear whether Twitter results are 'cleaned' before being provided to those requesting them (as discussed in Part I and Section 2.3.2).

Finally, it should be noted that at the time of writing the tweetbotornot library has been replaced with a new version in a completely separate library tweetbotornot2[21] in which the bot rating system has been changed and is now more conservative. In this way, the original findings in January 2020 may be been an artifact of the original implementation, however the polarised communities discovered since are certainly real and worthy of study.

Table 5.8. Supporter and Opposer accounts with a Botometer rating above 0.8. Counts of tweets, friends, and followers, and ages are as of the last tweet captured during the collection period in January, 2020. *This account was found to have been deleted when checked in October, 2020. †This account was found to have been deleted when checked in December, 2020.

	Supporters			**Opposers**	
	Bot 1	Bot 2*	Bot 3	Bot 1	Bot 2†
Contribution	5	9	59	4	4
Retweets	5	9	56	4	4
Age (in days)	1,081	680	1,087	1,424	925
Lifetime tweets	47,402	10,351	349,989	62,201	74
Tweets per day	43.85	15.22	321.98	43.68	0.08
Friends	17,590	13,226	25,457	633	392
Followers	16,507	13,072	24,873	497	55
Reputation	0.484	0.497	0.494	0.440	0.123

5.4.2 The most bot-like accounts

Deeper analysis of the most bot-like accounts (with a CAP rating of 0.8 or more) was conducted, revealing that the kinds of bot-like accounts present in each community differed significantly in a few primary respects (see Table 5.8). For convenience, we will refer to these accounts as "bots", but given all but Opposer 2 clearly present as genuine human users, they all also qualify as "social bots" (Cresci, 2020) and therefore are likely to be tools for influence. The accounts were re-examined in late 2020, finding that two had been suspended. Screenshots were taken of their Twitter profiles (see Figures 5.23[22] and 5.24). Two of the Supporter accounts appear to be American supporters of US President Donald Trump, while the third presents as an Australian indigenous woman from Tasmania who is also an active Trump supporter. The Opposer accounts include one with very little personal detail, mentioning only a hashtag for decentralised finance,[23] in its description, and one that presents as a left-wing individual.

Table 5.9. Changes in bot accounts between January and October 2020. Details for Supporter bot 2 are missing as it had been suspended by October.

Account	Friends	Followers	Tweets	Tweets / day
Supporter bot 1	1.7k ↑	1.1k ↑	36.3k	130
Supporter bot 2	14.5k ↑	13.4k ↑	157.3k	≈ 600
Opposer bot 1	9 ↓	1 ↓	10k	37
Opposer bot 2	581 ↑	1 ↑	25	< 1

[21] https://github.com/mkearney/tweetbotornot2. Accessed 2022-01-10.

[22] Supporter bot 2's account had been removed by this time, and so a mock-up based on the last known tweet in the ArsonEmergency corpus is presented in Figure 5.23b.

[23] *Decentralised finance*: a field of cryptocurrency in which blockchain technology is used to avoid financial institutions in transactions. Source: https://theconversation.com/decentralised-finance-calls-into-question-whether-the-crypto-industry-can-ever-be-regulated-151222. Posted 2020-12-12. Accessed 2022-01-10.

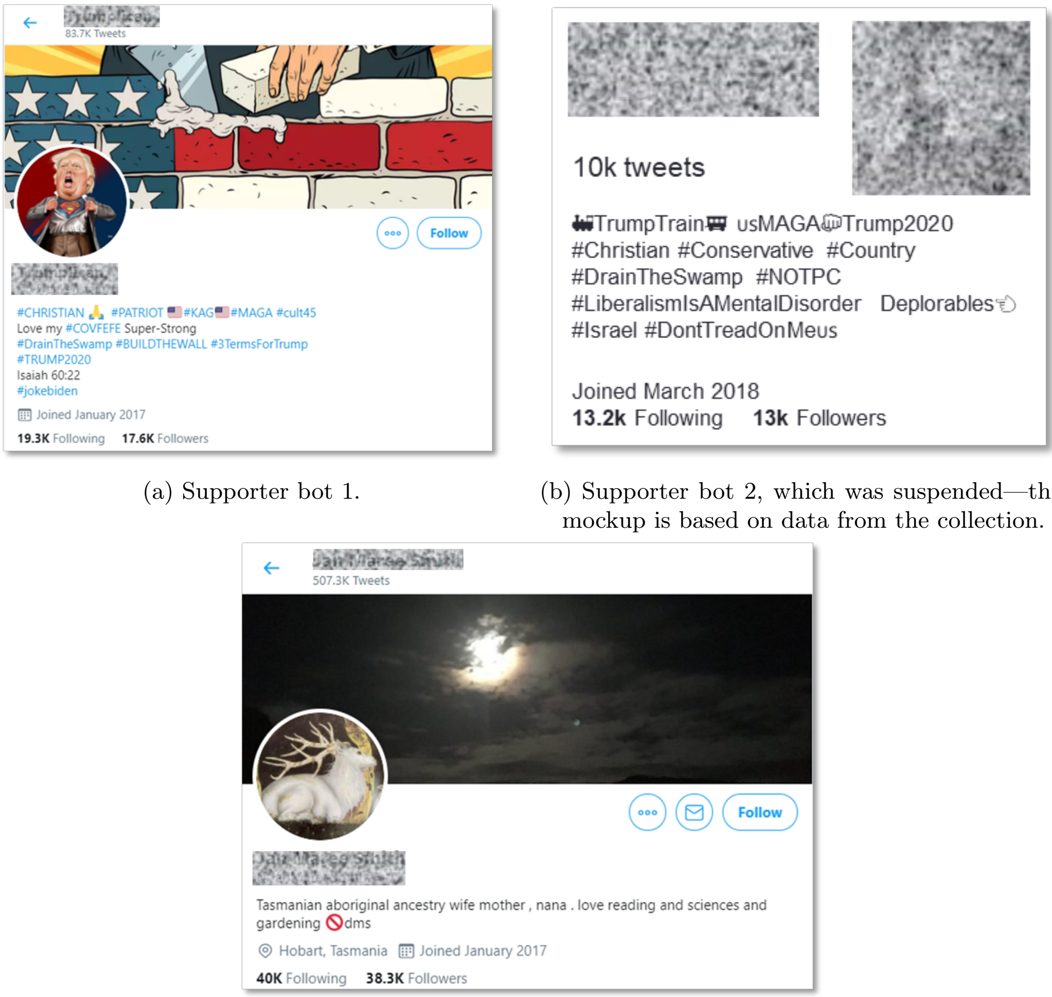
Together, the five accounts contributed 81 tweets over the 18 day collection period, 73 by the Supporters (including 59 from Bot 3) and 4 each from the Opposer bots. This suggests they had very limited opportunity to have an impact on the discussion. All accounts had been active for at least eighteen months, up to a maximum (at the time of the collection) of nearly four years. The variations in posting rates highlight the fact that Botometer's ensemble classifier will catch accounts that do not have high posting rates (e.g., Opposer bot 2 only posted approximately 25 tweets per year, but had been suspended by December, 2020), even though Botometer's performance against newer bots has begun to diminish (Feng et al., 2021). The changes between early and late 2020 offer further evidence of automation (see Table 5.9). The *reputation* score is defined by

$$reputation = \frac{|followers|}{|friends| + |followers|}, \tag{5.1}$$

and is a measure considered desirable enough to be worth manipulating through follower fishing (Dawson and Innes, 2019), yet even the bots' reputation scores are not very different (other than Opposer bot 2, which seems to be a rarely used account). In fact, the primary distinction between the Supporter and Opposer bots is the magnitude of their friend and follower counts.

It is not clear why these accounts are so different. It is possible these accounts are, in fact, merely highly motivated people, who spend a significant amount of time curating their Twitter feeds to include material they prefer and then retweet almost everything they see to simply promote their preferred narrative. This accords with recent observations that Twitter increasingly consists of retweets of official sources and celebrities and tweets with URLs, and rather than being a town square of public discussion, it should be treated as an "attention signal", which highlights the "stories, users and websites resonating" at a given time (Leetaru, 2019). These accounts appear driven to amplify that "attention signal" for ideological reasons, for the most part; Opposer bot 2's tweeting motivations are unclear, but it may have been a bot account left dormant for later commercial use (e.g., for narrative switching, Dawson and Innes, 2019). What also stands out is that the Supporter bots differ distinctly from the rest of the Supporter community who relied much less on retweets than the Opposer community.

Figure 5.25 shows the activity patterns for the Supporter and Opposer bot accounts, and also for the 15 Unaffiliated accounts that had been suspended when the bot analysis was conducted (at the end of January, 2020). The Opposer contribution is small and occurs in Phase 2 and the first day of Phase 3, clearly responding to the MSM news, while the Supporter bots are active in the lead up to Phase 2 and well into Phase 3, engaging in the ongoing discussion, though their activity patterns indicate that if they are bots tweeting frequently, then their tweets mostly avoided using `#ArsonEmergency` (and thus were not captured in our collection). The Unaffiliated accounts are also mostly active only on the day the story reached the MSM and the



[redacted]
83.7K Tweets

[redacted]

#CHRISTIAN 🙏 #PATRIOT 🇺🇸#KAG🇺🇸#MAGA #cult45
Love my #COVFEFE Super-Strong
#DrainTheSwamp #BUILDTHEWALL #3TermsForTrump
#TRUMP2020
Isaiah 60:22
#jokebiden

Joined January 2017

19.3K Following **17.6K** Followers

(a) Supporter bot 1.

[redacted] [redacted]

10k tweets

🚂TrumpTrain🚃 usMAGA🇺🇸Trump2020
#Christian #Conservative #Country
#DrainTheSwamp #NOTPC
#LiberalismIsAMentalDisorder Deplorables👉
#Israel #DontTreadOnMeus

Joined March 2018
13.2k Following **13k** Followers

(b) Supporter bot 2, which was suspended—this mockup is based on data from the collection.

[redacted]
507.3K Tweets

[redacted]

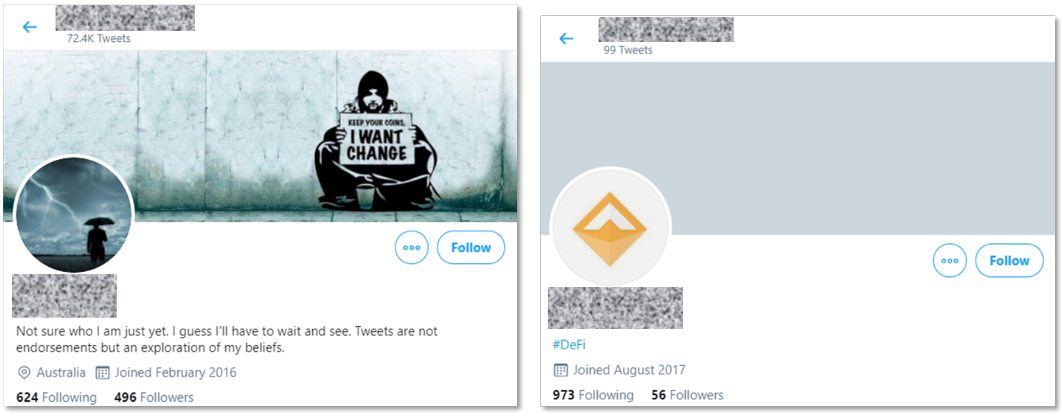
Tasmanian aboriginal ancestry wife mother , nana . love reading and sciences and gardening 🚫dms

Hobart, Tasmania Joined January 2017

40K Following **38.3K** Followers

(c) Supporter bot 3.

Figure 5.23. Supporter accounts with a Botometer rating higher than 0.8, implying a high degree of bot-like traits. Personal details have been obscured. Screenshots of accounts were obtained in mid October, 2020.



[redacted]
72.4K Tweets

[redacted]

Not sure who I am just yet. I guess I'll have to wait and see. Tweets are not endorsements but an exploration of my beliefs.

Australia Joined February 2016

624 Following **496** Followers

(a) Opposer bot 1.

[redacted]
99 Tweets

[redacted]

#DeFi

Joined August 2017

973 Following **56** Followers

(b) Opposer bot 2.

Figure 5.24. Opposer accounts with a Botometer rating higher than 0.8, implying a high degree of bot-like traits. Personal details have been obscured. Screenshots of accounts were obtained in mid October, 2020. Bot 2 was suspended in December, 2020.

following day, and their contribution was limited to only 32 tweets.

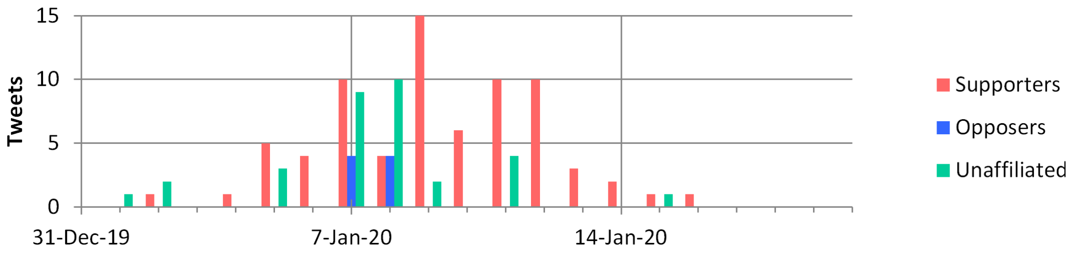


Figure 5.25. Tweets per day by the three Supporter, two Opposer and fifteen bot accounts.

5.4.3 Inauthentic behaviour

Aggressive language was observed in both Supporter and Observer content, but the hashtag and mention use provide the most insight into potential inauthentic behaviour (Gleicher, 2018). Supporters used more hashtags and more mentions in tweets than Opposers in general (Table 5.3), and posted individual tweets with many more of each (the number of tweets with at least 14 hashtags or 5 mentions was 50), though a small proportion of Unaffiliated accounts used even more hashtags in their tweets (a maximum of 27). Supporters posted tweets consisting of only hashtags, mentions and a URL in various combinations (i.e., eschewing actual content) far more frequently than Supporters or Unaffiliated, on a per-account basis, particularly in Phase 3 (see Table 5.10). Using hashtags and mentions in these numbers is a way to increase the reach of your message (though, ironically, it often leaves little space for the message itself), but can also be used to attack others or pollute hashtag-based discussion communities (Conover et al., 2011; Woolley, 2016; Nasim et al., 2018).

Table 5.10. Frequency of inauthentic text patterns in the ArsonEmergency tweets (includes retweeted text).

		Overall		**Phase 1**		**Phase 2**		**Phase 3**	
		Count	% of All	Count	% of All	Count	% of All	Count	% of All
Supporters	*All tweets*	6,972	100.0%	1,573	100.0%	121	100.0%	5,278	100.0%
	Hashtag(s)	20	0.3%	1	0.1%	0	0.0%	19	0.4%
	Hashtag(s) + URL	669	9.6%	160	10.2%	7	5.8%	502	9.5%
	Mention(s) + Hashtag(s)	340	4.9%	60	3.8%	3	2.5%	277	5.2%
	Mention(s) + Hashtag(s) + URL	73	1.0%	12	0.8%	2	1.7%	59	1.1%
Opposers	*All tweets*	3,587	100.0%	33	100.0%	327	100.0%	3,227	100.0%
	Hashtag(s)	0	0.0%	0	0.0%	0	0.0%	0	0.0%
	Hashtag(s) + URL	47	1.3%	1	3.0%	3	0.9%	43	1.3%
	Mention(s) + Hashtag(s)	0	0.0%	0	0.0%	0	0.0%	0	0.0%
	Mention(s) + Hashtag(s) + URL	5	0.1%	0	0.0%	0	0.0%	5	0.2%
Unaffiliated	*All tweets*	16,987	100.0%	1,961	100.0%	759	100.0%	14,267	100.0%
	Hashtag(s)	34	0.2%	2	0.1%	0	0.0%	32	0.2%
	Hashtag(s) + URL	629	3.7%	181	9.2%	14	1.8%	434	3.0%
	Mention(s) + Hashtag(s)	180	1.1%	35	1.8%	8	1.1%	137	1.0%
	Mention(s) + Hashtag(s) + URL	102	0.6%	18	0.9%	1	0.1%	83	0.6%

In one notable instance, a Supporter account posted 26 highly repetitive tweets to an Opposer account within 9 minutes, including only the #ArsonEmergency hashtag in the majority of them. In six tweets, other accounts were mentioned, including prominent Opposer and Unaffiliated accounts, perhaps in the hope that they would engage by retweeting and thus draw in their own followers.

As name switching had been observed in other discussions (Mariconti et al., 2017; Ferrara, 2017), we examined the accounts for such behaviour finding only 13 examples, including one Opposer and five Supporters. Some of the changes appeared to reflect a new 'personality' (*cf.*, Dawson and Innes, 2019), but not in a particularly deceptive way – instead, the changes of name seemed whimsical. Further results of name switching analysis can be found in Appendix A.2.1.

These findings make it clear that although Supporters were directly engaging with other accounts, their interactions did not always necessarily appear positive or genuine. Supporters consistently tweeted just hashtags and URLs in around 10% of their tweets in the larger phases, and use of the other inauthentic text patterns grew between the first and last phases, possibly in response to the ZDNet article. Unaffiliated accounts in Phase 1 did the same, but use of that pattern dropped away in the later phases. Opposers rarely engaged in any of these text patterns.

Network analyses can reveal the existence of interactions between accounts, but not their nature. The sheer numbers of interactions prevent manual inspection, however searching for text patterns such as those above, based on manual inspection of samples, can provide an indication of the authenticity or inauthenticity of the interactions, and are easy to calculate. Further relatively simple analysis of use of specific hashtag sequences, e.g., "`#ArsonEmergency #EcoTerrorism #ClimateChangeHoax`", in that order (*cf.*, Pacheco et al., 2021, case study 3), is another potentially simple yet informative analysis relying on sequence mining (Mooney and Roddick, 2013).

Further detail on inauthentic behaviour can be found in Appendix A.2.

5.5 Discussion

Our discussion addresses the research questions we posed in Section 5.1.4.

RQ1 *Discerning misinformation-sharing campaigns.* Analysis revealed two distinct polarised communities, each of which amplified particular narratives. The content posted by the most influential accounts in each of these communities shows Supporters were responsible for the majority of arson-related content, while Opposers countered the arson narrative, debunking the errors and false statements with official information from community authorities and fact-check articles. Prior to the release of the ZDNet article, the discussion on the `#ArsonEmergency` hashtag was dominated by arson-related content. In that sense, the misinformation campaign was most effective in Phase 1, but only because its audience was small. Once the audience grew, as the hashtag received broader attention, the conversation became dominated by the Opposers' narrative and related official information.

RQ2 *Differences in the spread of information across phases and other discussions.* We regarded URL and hashtags as proxies for narrative and studied their dissemination, finding distinct differences between the groups and the their activity in different phases. In Phase 1, only Supporters and Unaffiliated shared URLs, the most popular of which were in the NARRATIVE category, but by the third Phase, the most popular URLs shared were DEBUNKING in nature by a ratio of 9 to 1, and NARRATIVE URLs were share only by Supporter accounts. Although it is unclear whether this change in sharing behaviour was due to changes in opinions or the influx of new accounts, there was certainly a changing of the guard. Of the 2,061 accounts active in Phase 1, less than 40% (787) remained active in Phase 3. While most Phase 1 Supporters (339 of 360) posted in Phase 3, many fewer Unaffiliated accounts did (427 of 1,680) indicating that the Supporters lost the support of most of the Phase 1 Unaffiliated accounts.

The diversity of URL and hashtag use also changed from Phase 1 to Phase 3: while the number of active Supporters grew modestly from 360 to 474, the number of unique external URLs they used grew more, proportionately, from 193 to 321. Opposers and Unaffiliated used more unique URLs in Phase 3 (492 and 4,368, respectively), but Figures 5.15 and 5.19 shows they focused on a small set of URLs more than Supporters did.

The number of hashtags Supporters used increased from 191 hashtags used 5,382 times to 543 hashtags used 14,472 times. This implies Supporters attempted to connect `#ArsonEmergency` with other hashtag-based communities, which could have been to in order to promote their message widely, to co-opt existing discussion spaces, or due to non-Australian contributors being unfamiliar with which hashtags would be relevant to the mostly Australian audience. From Phase 1 to Phase 3, Opposer activity increased from 34 hashtags used 150 times to 200 hashtags used 9,549 times, and Figure 5.11b shows Opposers focused the majority of their discussion on a comparatively small number of hashtags.

The `#ArsonEmergency` discussion's growth rate was similar to another contemporary discussion (`#AustraliaFire`), inasmuch as they both experienced events causing significant changes in their participants, but it was clearly different from that of a well-established discussion (`#Brexit`).

RQ3 *Behavioural differences over time and the impact of media coverage.* Supporters were more active in Phase 1 and 3 and used more types of interaction than Opposers, especially replies and quotes, implying a significant degree of engagement, whether as trolls or as "sincere activists" (Starbird and Wilson, 2020). Unaffiliated accounts were consistently drawn in to the discussion in Phase 1, but most of these accounts left in the later phases and were replaced with many more Unaffiliated accounts who presumably joined based on reports in the MSM. Opposers and Supporters made up the majority of retweeted accounts overall, and made up 22 of the top 25 accounts retweeted by Unaffiliated accounts in Phase 3. Supporters' use of interaction types remained steady from Phase 1 to 3. While behaviour remained relatively similar, activity grew for both groups after the story reached the MSM. The vast majority of accounts shared articles debunking the false narratives. The publication of the ZDNet article (Stilgherrian, 2020) also affected activity, spurring Opposers and others to share the analysis it reported.

RQ4 *Position of communities in the discussion network.* Supporter efforts to engage with others in the discussion resulted in them being deeply embedded in the discussion's reply, mention and quote networks and having correspondingly high centrality values. Our k-core analysis showed they were evenly distributed throughout the networks, from the periphery to the cores. Despite Opposers staying more on the periphery of the networks, they maintained high closeness and eigenvector centrality scores, meaning they stayed connected to more of the network than Supporters and certainly to more important nodes in the network. Correspondingly, this may imply that Supporters, though being highly connected, were not connecting as efficiently as Opposers, in order to spread their narrative. Both Opposer and Supporter groups were highly insular with respect to each other, across a variety of network analyses, but they connected strongly to the broader community according to E-I indices.

RQ5 *Content dissemination and coordinated amplification.* Analyses of hashtag and URL use revealed further evidence of the gap between Supporters and Opposers, not just in terms of connectivity, as discussed above, but also in terms of narrative. Supporters used a variety of hashtags to reach greater audiences, to disrupt existing communication channels, or to otherwise harass. In doing so, they exhibited less evidence of coordination than Opposers, who were focused in both the hashtags and URLs they used, supported by, or in concert with, the much greater number of Unaffiliated accounts. Analysis of co-activities (namely co-retweeting, and co-URL and co-hashtag instances) suggested a lack of botnets in the discussion and that some Unaffiliated and Opposers were coordinating their URL sharing, appearing together in cliques that are often attributed to automation (e.g., Pacheco et al., 2020, and the case study presented in Section 7.4.4). The apparent coordination could, however, be attributed to high levels of popularity driven by increased activity in Phase 3 (i.e., coincidence due to high numbers of discussion participants), and the co-activities of Supporters indicated the presence of genuine human users more than any automated coordination. Further analysis using account/URL 2-layer networks showed that Opposers and Unaffiliated were focused on sharing a small set of URLs, compared with Supporters' greater variety. These findings imply the Supporter community members, for all they attempted to engage with others via replies, mentions and hashtags, becoming deeply embedded in the interaction networks, remained relatively isolated from a narrative perspective.

RQ6 *Support from non-Australian accounts.* Based on manual inspection of accounts' free text 'location' fields (and assuming the majority were honest), the Supporter group included more non-Australian than Opposers, with the greatest number of non-Australian accounts Unaffiliated with either, but the vast majority of all groups indicated they were located in Australia ($> 70\%$). Despite the large number of Unaffiliated accounts present in Phase 1 (1,680), the majority joined the discussion in Phase 3, likely bringing in the majority of non-Australian accounts. Investigations of content dissemination also revealed that Opposers received the majority of Unaffiliated support, resulting in a majority of debunking article shares in Phase 3 from a majority of narrative-aligned article shares in Phase 1, so it is possible that this also included non-Australian support. Given most accounts do not report their location, and locations have not been verified, this conclusion remains speculative.

RQ7 *Support from bots and trolls.* We found very few bots and their impact was limited: only 0.8% (20 of 2,512) had a Botometer (Davis et al., 2016) CAP score above 0.6 while 96.6% (2,426) were highly likely to be human (CAP < 0.2). In contrast, Graham & Keller had found many more bots (46%) and fewer humans ($< 20\%$) in their smaller sample (Stilgherrian, 2020; Graham and Keller, 2020). The affiliated 'bot' accounts, on closer examination, may not all have been automated, but the ones with bot-like posting rates could certainly be classed as 'social bots' (Cresci, 2020) given their appearance as genuine human users. In fact, following the ZDNet article, Graham & Keller argued that (non-automated) trolls are the more insidious element of this campaign, providing evidence that `#ArsonEmergency` was created specifically to counter `#ClimateEmergency` (Graham and Keller, 2020) and may even have been part of a broader disinformation campaign involving elements of the political and media elite (Keller et al., 2020). Aggressive language was observed in both affiliated groups, but troll-like tweet text patterns including only hashtags, mentions and URLs (i.e., without content terms) were employed far more often by Supporters. Distinguishing deliberate baiting from honest enthusiasm (even with swearing) is non-trivial (Starbird et al., 2019; Starbird and Wilson, 2020), but identifying targeted tweets lacking content is a more tractable approach to detect inauthentic and potentially malicious behaviour.

A number of further issues raised in this chapter are worth commenting on.

5.5.1 A disinformation campaign?

There is good reason to believe that #ArsonEmergency was deliberately created (Graham and Keller, 2020), forming a 'data deficit' (Smith et al., 2020) for the sharing of misinformation regarding the arson narrative. This could form an isolated echo chamber for recruiting a new user base and radicalising it. Then, once established, it could link into the broader discussions by using a variety of hashtags in their tweets, which is what we observed. Radicalisation may not have been the ultimate goal of this particular community, but the technique could be used by other Twitter users. Large isolated communities of accounts have been discovered by researchers before (e.g., the Star Wars botnet found by Echeverria and Zhou, 2017), and though humans would need to be more careful to avoid adding hashtags (thus linking in other communities), moderate activity could be obscured. #ArsonEmergency was discovered because participating accounts were known to Graham and Keller. This study provides confirmation of the presence of trolling, but no direct evidence of disinformation (*cf.*, Graham and Keller, 2020; Keller et al., 2020).

5.5.2 A successful intervention?

If the publication of the ZDNet article was intended as an intervention to counter the misinformation campaign on #ArsonEmergency, was it successful? Supporter numbers and activity rose dramatically after the story reached the MSM, drawing in many overseas contributors and shifting towards more inauthentic behaviour patterns. In contrast, however, the Opposer response was swift and simple, focusing on retweeting links to the ZDNet article and other fact-checks and official information, as it became available. Opposer activity was highest in Phase 2, but may have helped provide content for the incoming Unaffiliated accounts to share. In this way, the Unaffiliated accounts eventually shared DEBUNKING articles much more frequently than NARRATIVE aligned ones in the third phase. This occurred despite great increases in activity by Supporters, including Supporters using more hashtags, mentions, replies, retweets and quotes than in Phase 1.

At an high level, this situation involved a number of factors:

1. researchers noticed a hashtag, which itself was misinformation (that there was an #ArsonEmergency);
2. then noticing that the slowly growing surrounding discussion had a high proportion of bot accounts;
3. discussing these findings with a technology journalist who wrote an article on the findings.

The bot analyses were preliminary and did not stand further scrutiny (shown here and elsewhere, Graham and Keller, 2020),[24] but only days later the researchers clarified that much of the behaviour may have been due to (human) trolls (Graham and Keller, 2020), and later presented evidence to suggest that the activity may have been coordinated with a broader disinformation campaign (Keller et al., 2020). The initial article, however, was enough to draw public attention to it, initially through ZDNet's audience and their followers, spurring the Opposer community to form. By mid-morning after the news reached the breakfast MSM, it had had time to spread around the world as well as become known amongst the broader community, attracting attention in the Australian Twittersphere. By the end of Phase 2, official announcements refuting the information were reported on, and these became the focal points of the Opposer and Unaffiliated URL sharing.

5.5.3 On the role of academics and researchers

Researchers typically rely on peer-reviewed channels to disseminate their findings and insights. Those researching online misinformation now find that, for their insights to have a real-world impact, they need to augment these channels with non-peer-reviewed ones. Many social media researchers observe social media on a daily basis and offer informal contemporary comments via social media and when interviewed for media reports. Organisations and institutions established with specific anti-misinformation agendas produce non-peer-reviewed technical reports on social media studies to provide more comprehensive analysis at the cost of weeks or months of delay. A part of the role of peer-reviewed literature is to evaluate the effectiveness of these 'interventions' of commentary or technical reports, particularly by those conducting the interventions, because they have the best knowledge of their original aims.

5.5.4 On labelling communities

The Supporter and Opposer communities were relatively easy to label, based on manual inspection of their most retweeted accounts. In studies of larger datasets, other methods may need to be considered, relying on automated analyses or other cues. Textual analyses such as topic modelling of profile descriptions, as discussed in Section 5.4.2, could reveal a community's major interests, but such descriptions are often very terse and sometimes do not align with account behaviour. In their study of the 2017 German election, Morstatter et al. (2018) identified several very large clusters of accounts using Louvain (Blondel et al., 2008), but then determined the content of each major cluster using a hierarchical topic modelling technique applied to the hashtags they used. As discussed in Section 5.3.3, hashtags can be considered proxies for discussion themes. Furthermore, the predominant language used in the clusters also helped reveal distinct German-speaking alt-right and English-speaking alt-right clusters.

[24] In fact, the ZDNet article *could* be argued to be misinformation itself, albeit intended to expose the arson-related misinformation.

5.5.5 Nuance of Supporter interactions

Our analysis indicated that Supporters used original content and interactions, such as mentions and replies, more than Opposers, and that their increased activity implied they contributed to the discussion more. Opposers retweeted more, suggesting they did not interact as much, and their primary contribution was in the short Phase 2. The network analyses in this work also suggested that Supporters were more deeply embedded and positioned throughout the networks of interactions in the discussion, but the analysis of inauthentic behaviour revealed that significant portions of these interactions with the broader community we not constructive, and some could be considered clear harassment.

These findings emphasise the need to incorporate mixed-method analyses, and to use the methods to complement each other and take advantage of their relative strengths.

5.5.6 Methodological contributions

Methodologically, the approach taken in this paper has taken advantage of recent advances in network science, bolstered them with established network, statistical and bot analyses, and proposed platform feature use patterns as a simple approach to identifying inauthentic behaviour. This final element helps illuminate the tone of interactions between the Supporter community and those outside it, which could have been assumed cordial based on network analyses of their frequency. In fact, Supporters' inauthentic behaviour seemed to mostly increase after the ZDNet article was published, particularly their use of targeting tweets with @mentions. Determining whether this contributed to the Unaffiliated accounts' shift away from the arson narrative remains an open question.

The co-activity analyses used in this study further validate the utility of the approach (which we present in Chapter 7 and review in Section 2.6). Similar recent work has applied co-URL and co-domain analysis to expose information polluters on the basis of the news they disseminate (and the sources from which it comes) (Truong et al., 2022). The inclusion of a temporal constraint aids in identifying concerted coordination over grassroots coordination, but improvements could be introduced to account for the scale of the discussion (and therefore further reduce the impact of coincidental co-activities).

Further research is required to examine the dynamic aspects of the social and interaction structures formed by groups involved in spreading misinformation to learn more about how to better address the challenge they pose to society. Future work will draw more on social network analysis based on interaction patterns and content (Bagrow et al., 2019) as well as developing a richer, more nuanced understanding of the Supporter community itself, including revisiting the polarised accounts over a longer time period and consideration of linguistic differences. A particular challenge is determining a social media user's intent when they post or repost content, which could help distinguish between disinformation intended to deceive, and merely biased presentation of data or misinformation that aligns with the user's worldview.

5.6 Conclusion

The study of polarised groups, their structure and their behaviour, during times of crisis can provide insight into how misinformation can enter and be maintained in online discussions, as well as provide clues as to how it can be removed. The #ArsonEmergency activity on Twitter in early 2020 provides a unique microcosm to study the growth of a misinformation campaign before and after it was widely known. Here we have shown that polarised groups can communicate over social media in very different ways while discussing the same issue. In effect, these behaviours can be considered communication strategies, given they are used to promote a narrative and represent attempts to convince others to accept their ideas. Supporters of the arson narrative used direct engagement with mentions and replies to reach individuals and hashtags to reach groups with a wide range of URLs to promote their message, while Opposers focused on using retweets and a select set of URLs to counter their message. Supporter activities resulted in them being deeply embedded and distributed in the interaction networks, yet Opposers maintained high centrality and were supported by and appeared to coordinate with active Unaffiliated accounts. The counteraction appears to have been successful, with the predominant class of articles shared shift from narrative-aligned in Phase 1 to debunking articles in Phase 3. Graham & Keller's efforts to draw attention to the #ArsonEmergency discussion (Stilgherrian, 2020), and the subsequent associated MSM attention, is likely to have contributed to this effect, given the significant increase in discussion participants in Phase 3. This highlights the value in publicising research into misinformation promotion activities.

We speculate that the communication patterns documented in this study could be communication strategies discoverable in other misinformation-related discussions, such as those relating to vaccine conspiracies (Broniatowski et al., 2018), COVID-19 anti-lockdown regulations (Loucaides et al., 2021), challenging election results (Scott, 2021) or QAnon (The Soufan Center, 2021b), and could help inform the design and development of counter-strategies. An approach similar to Graham and Keller's could form a model for future similar interventions, if the conditions are suitable.

Chapter 6

Persistent Polarisation

Contrary to expectations that the increased connectivity offered by the internet and particularly Online Social Networks (OSNs) would result in broad consensus on contentious issues, we instead frequently observe the formation of echo chambers, in which only one side of an argument is entertained, particularly in contentious two-sided discussions. These can progress to filter bubbles, actively filtering contrasting opinions, resulting in vulnerability to misinformation and increased polarisation on social and political issues, with real-world effects when they spread offline, such as vaccine hesitation and violence. This work seeks to develop a better understanding of how echo chambers manifest in different discussions dealing with different issues over an extended period of time. We explore the activities of two groups of polarised accounts across three Twitter discussions in the Australian context spanning almost a year. We find that the groups form echo chambers across different interaction types. More specifically, we found accounts arguing against marriage equality in 2017 were more likely to support the notion that arsonists were the primary cause of the 2019-2020 Australian bushfires, and those supporting marriage equality argued against that arson narrative. We also found strong evidence that the stance people took on marriage equality in 2017 did not predict their political stance in discussions around the federal election two years later. The findings suggest that, specifically, 1) people chose to support marriage equality or not based on factors other than political leaning, 2) there was alignment between opinions on marriage equality and bushfire causes, and more generally, 3) the echo chamber and polarisation effects observed in the two sets of groups in the datasets explored may be at least partially due to social circles as well as stances on the issues being discussed, and 4) although mostly isolated from each other, the fact that the polarised groups frequently interact with the broader community offers hope that the echo chambers may be reduced with concerted outreach to members.

*The content for this chapter is based on Publication **IX**.*

6.1 Introduction

The increased connectivity and relative anonymity offered by the internet and especially by social media platforms (aka online social networks, or OSNs) was once hoped to provide a mechanism for a more inclusive society, especially with regard to political involvement, "promot[ing] more civic engagement and participation in elections" (p.40, Hwang et al., 2012). OSNs in particular allow people to connect with friends, family and like-minded individuals to form and maintain communities with shared beliefs, values and interests. Observers of modern social media will note, however, that, like with any complex system, there are unintended consequences of making reaching out to others so easy, including the broad spread of conspiracies (e.g., QAnon and the Flat Earth Society, The Soufan Center, 2021b; Brazil, 2020), increased polarisation (Garimella and Weber, 2017), especially in political discussions (Garimella et al., 2018a), providing environments for radicalization (Badawy and Ferrara, 2018) and extremism (Baumann et al., 2021), and coordinated aggression (Bot Sentinel, 2021; Mariconti et al., 2019). The general consensus on contentious issues expected by classical opinion modelling theory (DeGroot, 1974; Baronchelli, 2018) has instead been replaced by communities focused around competing stances on those issues, *echo chambers* in which only one opinion is entertained (Pariser, 2012; Barberá et al., 2015), entrenched by recommender systems preventing contrary voices from entering, thus forming *filter bubbles* (Pariser, 2012), which leaves us vulnerable to misinformation (Nikolov et al., 2021) and disinformation (Starbird, 2019). When this misinformed aggression moves beyond the online sphere it has real-world effects such as vaccine hesitancy and anti-lockdown movements in a time of pandemics (Broniatowski et al., 2018; Loucaides et al., 2021; Loomba et al., 2021), and violence (Samuels, 2020), some of which is politically motivated (Scott, 2021; Mackintosh, 2021).

The dynamics of these echo chambers is of particular interest, because their entrenchment of particular viewpoints drives the in-group/out-group mentality behind polarisation, which, left unchecked, can lead to fundamental difficulties in cooperation, with particular implications for democratic political systems (Bail et al., 2018). Not all are convinced of their danger, however (Bruns, 2019b), because individuals are known to be members of many social circles, each with their own common attributes and interests (e.g., family, friends, work, or sports, referred to ask *foci* by Feld, 1981), and each of these circles will provide new and potentially contrasting viewpoints on a various of overlapping issues. Questions remain over how these social circles and echo chambers influence social behaviour, both online and offline (Bruns, 2019b; Nasim, 2019), but it is known that there is alignment between some sets of opinions (Baumann et al., 2021), particularly with regard to political viewpoint (Jost et al., 2003; Jost, 2017).

Given the relative youth of OSNs, longitudinal studies of online polarisation are only just beginning to appear, but often seek to follow polarisation on specific contentious issues over time (Garimella and Weber, 2017; Garimella et al., 2017). Our focus, instead, is on investigating communities that remain polarised over time across a variety of issues. Furthermore, it is important to study their activities in the context of the broader discussion to determine not just to what degree the groups isolate themselves from each other, but also how isolated the groups remain from the surrounding community. For these reasons, we require datasets in which known polarised groups are known to be active that are collected over a reasonable period of time *and* relate to a variety of discussion topics. The issue of political alignment is also relevant, due to vulnerability to misinformation introduced by increased partisanship (Nikolov et al., 2021) and the fact that political alignment has been observed to correlate with different personal values (Jost et al., 2003), for example, right-aligned people value tradition more than left-aligned people while left-aligned people value egalitarianism more (Jost, 2017).

Although OSNs share many features (see Section 2.3) the openness of micro-blog platforms, such as Twitter, Parler and Gab, where one account can directly connect to any other (via, e.g., mentions, replies and retweets and their equivalents), provides the best opportunity for accounts in polarised communities to bridge the gaps. Doing so enables new and different information to flow between the communities, enabling the potential to grow consensus. In contrast, participants in Facebook, Instagram, Reddit and WhatsApp discussions can usually only refer to others in the same discussion thread or channel. We use Twitter data in this study, as it is the longest established of the three microblogs mentioned, and has the largest and most representative user base. It also provides a freely available rich data model, which includes information on the directed interactions between accounts, resulting in an up-to-date window into the direction and degree of information and influence flow between Twitter accounts.

In this chapter, we examine the roles played online by members of two previously identified polarised communities in the context of three separate online discussions, each focused on different topics and themes, over the period of almost a year. The polarised groups had been identified in discussions of contentious issues:

- Those using `#VoteYes` and those using `#VoteNo` (mutually exclusively) during the same sex marriage (SSM) debate during the Australian postal survey on the matter in late 2017 (Nasim et al., 2019), dubbed the *YES* and *NO* communities, respectively; and

- Those debating the role of arson and climate change during the 2019-2020 Australian bushfires (see Chapter 5), in which *Supporters* of the arson theory were countered by an *Opposer* community.

Notably, we have found these groups (which were clearly found to be polarised in previous analyses, based on the content they shared and their clustering) to intermix, but, at times, remain aligned, in the three datasets inspected. Our aim is to study the activities of these groups over time in different contexts to determine whether they remain polarised, and to characterise the nature of that polarisation using network and content analysis. Our network analysis relies upon accounts' interactions (i.e., retweets, replies, mentions and quotes) and the associations between topics they discuss as represented by partisan hashtags as proxies for clear stances on the issues at hand.

6.1.1 Research questions

We will guide our investigation of these groups' behaviour with the following research questions:

RQ1 Do polarised accounts continue to be active in the Australian Twittersphere over a period of years?

RQ2 Is their polarisation reflected in a range of their interactions (on Twitter) and discussion topics, or is it limited to just a particular type of interaction?

RQ3 Are accounts found to be polarised in one dataset still polarised in later datasets, including ones discussing different topics? In particular is there any alignment between partisan communities and those that were found to be polarised over other issues (e.g., SSM, bushfires)?

Our expectation is that the Australian Twittersphere is sufficiently well established to support persistent communities of accounts over long periods of time, ones which discuss related issues, and though they may be polarised on some issues, that polarisation may not be so pronounced on others and the communities may, at times, overlap. If this is found to be true, we can conjecture that the filter bubble effect is not as strong as it was thought to be, and the echo chambers constantly reconfigure and reorganise, allowing interaction between the members of different communities. Such an observation will also be in line with previous social interaction theories that established that people are a part of various overlapping social circles (Feld, 1981).

We also expect that the degree of polarisation will vary across interaction types because different interaction types are used for different purposes. Interactions between accounts may be direct, requiring that one account be aware of the other's identity (e.g., with an @mention, a reply or retweet), while others are indirect, requiring only knowledge of intermediary data and perhaps an associated common stance (e.g., common use of a partisan hashtag or URL). For direct interactions, there is the possibility that the connection is made because of a personal connection (e.g., a friendship or indication of personal respect) in addition to an agreement on stance. Furthermore, different direct interactions have different audiences: while a reply or a mention may be directed at the replied to or mentioned account, a retweet or quote tweet is aimed at the poster's followers despite the reference back to the originator of the retweeted or quoted tweet. For this reason, networks built from different interactions can be expected to exhibit different degrees of polarisation.

6.1.2 Contribution

This work provides the following contributions to the literature:

1. Two original datasets on the 2017 SSM debate in Australia, and the 2019 Australian federal election;
2. A methodology for the analysis of online polarisation between two non-overlapping groups based on their behaviour and discussion content; and
3. A longitudinal study of two sets of such polarised communities and their degree of alignment over a series of three Twitter datasets.

The remainder of this paper is structured as follows. We next explore related work in polarisation, particularly in relation to opinion modelling, and then discuss the four datasets used in our study. At that point, we clarify the labelling we use for our polarised groups and provide more specific hypotheses regarding their behaviour in our datasets. We next discuss the methods we use, which primarily rely on social network analysis (SNA) and homophily metrics, and then present our results. We conclude our results by directly addressing our research questions and hypotheses, and discuss implications and possible improvements before concluding.

6.2 Related Work

Although classical opinion formation analysis predicts gradual consensus through exchange of opinions and finding mutual similarities (DeGroot, 1974; Baronchelli, 2018), instead the size of the internet and community-focused features of social media platforms have resulted in increasing polarisation (Garimella and Weber, 2017). Add to this deliberate fostering of populist and introspective attitudes, and echo chambers (Barberá et al., 2015) and filter bubbles (Pariser, 2012) begin to form drifting towards polarisation on a range of societal issues.

A recent work of particular note is Baumann et al. (2021)'s study of opinion formation in multi-dimensional topic spaces using multi-agent simulations of social media interactions. Particular attention was paid to the dynamic nature of the social media interactions and how social networks emerge from them over time, rather than assuming they are static. We adopt a similar approach in this chapter. The researchers found that the combination of topology and agent homophily influenced the progression of opinion formation through stages, first developing consensus, but then veering to complete isolation, especially as topics overlap more (i.e., individual interactions pertain to multiple topics, as a tweet might include many hashtags) and as opinions increase in strength. This may help explain alignment amongst sets of contentious issues, a particular risk of online polarisation.

More detailed discussion can be found in Sections 2.1 and 2.2, where we have addressed the literature regarding information disorders and their exploitation as part of organised malicious online behaviour, such as computational propaganda. Echo chambers and polarisation are examined in Section 2.5.

This chapter seeks to provide empirical evidence from an Australian perspective, providing not just a longitudinal study of opinion polarisation over a number of distinct contentious and non-contentious topics, but also considering whether polarisation extends through different methods of online interaction. Baumann et al. (2021) considered homophily and heterophily based on political opinions from election-related surveys, whereas we infer opinion based on users' interactions and their use of partisan hashtags, and Garimella and Weber (2017) studied polarisation on Twitter in a longitudinal setting, but did so by focusing on particular issues rather than the communities around them.

6.3 Datasets

We analysed four datasets of tweets collected between 2017 and 2020. Two of those datasets were compiled on contentious social issues (marriage equality in Australia and climate change), one on the Australian Federal election, and a final dataset related to sports. Details are presented in Table 6.1.

The remainder of this section describes each of these datasets, clarifies the nomenclature we use to refer to the datasets and the polarised groups within them, and ends with our hypotheses regarding how these groups behave in three of the datasets.

Table 6.1. Dataset details.

Dataset	Tool	Twitter API	Duration	Tweets	Accounts	Method of Collection
SSM	GNIP	10% academic API	2017-09-01 to 2017-11-20	79,725	54,855	Keywords: `#MarriageEquality`, `#SSM`, `#auspol`, `#VoteYes`, `#VoteNo`
Election	TWINT	Web UI	2019-05-01 to 2019-05-21	398,352	4,429	Timeline scraping of seed accounts
ArsonEmergency	Twarc	Search API	2019-12-31 to 2020-01-17	27,546	12,872	Keyword: ArsonEmergency
AFL	RAPID	Streaming API	2019-03-22 to 2019-03-25	21,799	11,573	Keyword: afl

6.3.1 The marriage law postal survey (late 2017)

In late 2017, the Australian federal government conducted an optional national postal survey asking Australian voters "Should the law be changed to allow same-sex couples to marry?"[1] On the basis of a majority affirmative result, the government would commit to passing legislation to change the Marriage Act accordingly. From August, when the survey was announced, through to the final acceptance date of the ballots in November and beyond, discussions and debate raged on social media with strong opinions both for and against marriage equality. Ultimately, over 60% of the nearly 13 million responses voted 'yes' and the Australian Parliament changed the law to permit marriage between any two individuals.

[1] https://www.abs.gov.au/ausstats/abs@.nsf/mf/1800.0. Posted 2017-11-15. Accessed 2022-01-10.

During three months of the campaign, we collected tweets from Twitter's 10% academic sample stream[2] based on the keywords #MarriageEquality,[3] #SSM, #auspol, #VoteYes, and #VoteNo, capturing close to 80k tweets (and associated metadata) by almost 55k unique accounts.

The hashtags used as keyword filters belonged to two categories: general marriage equality-related terms (#MarriageEquality, #SSM, and #auspol), and ones clearly reflecting an opinion (#VoteYes and #VoteNo). We focused on the 17.3k accounts which used the opinion-linked hashtags, which we hypothesised would have relatively high structural cohesion around users of the same hashtag, and low structural cohesion among users of different hashtags. YES accounts were those that used only #VoteYes, NO accounts used only #VoteNo, and BOTH accounts used both hashtags. Of these, there were slightly more YES accounts than NO accounts (8.6k to 7.9k), and those using both made up just under 5% of the accounts using opinion hashtags (778). YES accounts contributed more tweets (18,621) than NO accounts (11,261) and BOTH accounts (7,246).

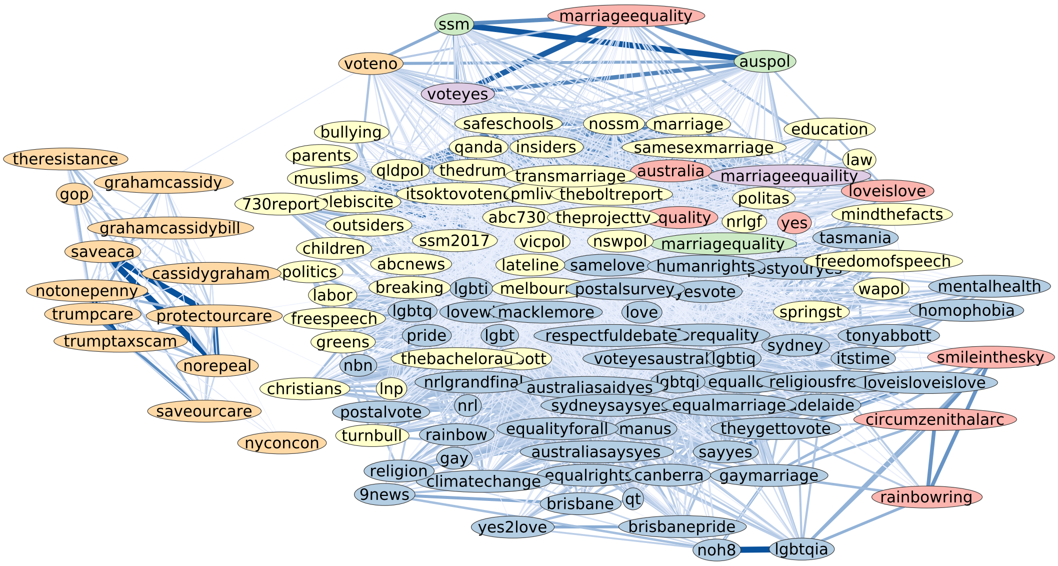


FIGURE 6.1. The hashtag network from the SSM dataset. Two hashtag nodes are linked if they were tweeted by the same user (though not necessarily in the same tweet), and the size and colour of the edge represents the frequency of co-mentioning (wider and darker = more frequent). Nodes are coloured according to Louvain cluster. Names of prominent public figures have not been anonymised in order to provide context. The orange cluster on the left clearly refers to US politics rather than the Australian SSM postal survey.

Some cleaning of the data was required due to international overlap with #VoteNo, which was also used in American discussions surrounding a medical insurance-related bill before the US Congress at the time. These tweets were identified through the use of a hashtag network. The network is visualised in Figure 6.1 with a force-directed layout clearly showing a minimally linked cluster of hashtags on the left that relate to

[2] The Decahose: https://developer.twitter.com/en/docs/twitter-api/enterprise/decahose-api/overview/decahose. Accessed 2022-01-10.

[3] All hashtag analysis was performed ignoring case, but capitals are included here for readability.

the foreign discussion. 6,295 tweets posted by 5,366 accounts mentioning the hashtags in the orange-coloured Louvain cluster (Blondel et al., 2008) to the left (other than `#VoteNo`) were identified as pollution and removed.

This dataset is referred to herein as the *SSM* dataset, and the YES and NO accounts as the *SSM* accounts (BOTH accounts are not included in the analysis as their position on the matter is just as obscure as OTHER accounts).

6.3.2 The Australian federal election (May 2019)

A total of 4,429 of the YES, NO and BOTH accounts (3,390, 631 and 408, respectively) were active during the election period surrounding the Australian federal election held on the 18th of May, 2019. Their activity was tracked, resulting in a dataset of nearly 400k tweets spanning three weeks. These activities were obtained, post-election, by retrieving their timelines via Twint[4] (a tool that obtains Twitter data directly from its web UI avoiding any recommender influence or constraint present in the APIs). Nearly 3.4k YES accounts were active during the campaign, compared to only 631 NO accounts. The data includes a variety of politically-relevant hashtags, and in particular we have identified 44 partisan hashtags.

6.3.3 Australia's "Black Summer" (2019-2020)

During the 2019-2020 southern summer, referred to as Australia's 'Black Summer', bushfires burnt over 16 million hectares of the Australian mainland, destroyed over 3,500 homes, and caused at least 33 human and a billion animal fatalities (NSW Bushfire Inquiry, 2020). While scientists attributed these bushfires to natural causes such as lightning, an alternative theory labelled arson as the cause of bushfires. At the peak of the bushfires season the hashtag `#ArsonEmergency` started trending on Twitter and was observed to include a high proportion of bot and troll activity (Stilgherrian, 2020; Graham and Keller, 2020). As discussed in Chapter 5, we collected a dataset of tweets during that period, both before and after news of bots and trolls reached the mainstream media. The dataset consisted of 27.5k tweets containing the term 'ArsonEmergency' posted by 12.9k unique accounts over 18 days in early January, 2020. The Tweets were obtained with Twitter's Standard Search API using Twarc.[5] We found two polarised communities in the retweet network, referred to here as the Arson groups. One community strongly supported the arson narrative (*Supporters*), claiming arson was the cause of the bushfires, posting 6,972 tweets, while the other community opposed that narrative with fact-check articles and official announcements in 3,587 tweets (*Opposers*). A second study on this hashtag and contemporary news media reports found evidence of a disinformation campaign conducted by trolls, which appeared coordinated with the help of prominent public figures (Keller et al., 2020).

This dataset provides our second set of polarised accounts.

[4] https://github.com/twintproject/twint. Accessed 2022-01-10.

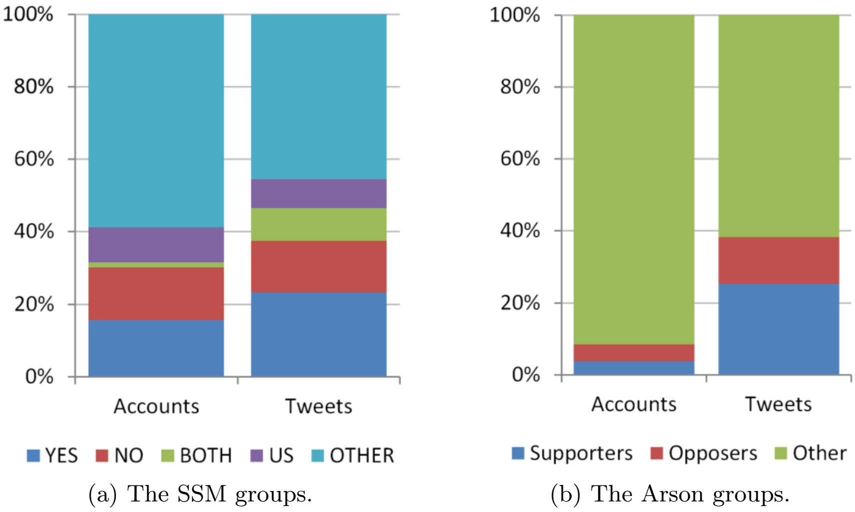
[5] https://github.com/DocNow/twarc. Accessed 2022-01-10.

6.3.4 AFL (March 2019)

A further, non-political dataset that could also exhibit patterns of polarisation was sought as a contrast. Australian Rules Football is a national pastime in Australia, particularly following the national competition run by and synonymous with the Australian Football League (AFL). Although fandom does not equate to polarisation, there are some combinations of clubs that might exhibit heightened aggressive or other extreme behaviour (e.g., traditional foes, such as the Capulets and Montagues from Shakespeare's "Romeo and Juliet"). The AFL1-en dataset from Chapter 4 is used for this purpose. It consisted of three days of AFL discussions collected over a weekend in March, 2019, just as the annual season began. Although a federal election was expected around this time, it was not called for another two weeks, and so little political content was expected to be captured. The collection tool RAPID (Lim et al., 2019) was used to stream all tweets from the Standard Twitter v1.0 Streaming API (up to rate limits) using the keyword 'afl' and a language filter for English and undefined (i.e., a 'lang' value of 'und', which captures text too short to inform Twitter's language detection).

Table 6.2. Sizes of the labelled polarised communities.

SSM			ArsonEmergency	
YES	NO	BOTH	Supporters	Opposers
8,623	7,880	778	497	593



(a) The SSM groups.

(b) The Arson groups.

Figure 6.2. The proportional contributions of the polarised SSM and Arson groups in the datasets in which they were found.

6.3.5 Polarisation labelling

Different methods were used to identify polarised communities in the SSM and Bushfires datasets due to the different ways in which they were collected. The sizes of the groups discovered are shown in Table 6.2 and their relative contributions in Figure 6.2.

Generalising our terminology, we refer to YES and Supporter groups as Category 1 accounts and NO and Opposer groups as Category 2 accounts later in this work. Any unaffiliated accounts appearing in networks are given the label OTHER.

TABLE 6.3. Relevant statistics of the datasets analysed.

Dataset	Retweets	Mentions	Replies	Quotes	Accounts	YES	NO	Supporters	Opposers
Election	331,682	51,673	12,397	26,025	4,429	3,390	631	72	156
ArsonEmergency	21,526	7,523	3,031	1,542	12,872	698	148	493	592
AFL	7,047	19,222	6,060	1,670	11,573	376	53	42	73

A summary of the content of the datasets and the extent of the polarised group presence in them in shown in Table 6.3.

6.3.6 Specific Hypotheses

We are now in a position to guide our investigation with specific hypotheses regarding these labelled groups.

As mentioned above, direct and indirect interactions can be expected to exhibit polarisation differently. Content-based connections made through hashtag use are based on what the hashtag expresses rather than who else is using it. For direct interactions, where the other account is known (at least by name), that other identity may influence a user's decision to interact or not. For these reasons, we might expect that the polarisation evident in the Arson groups might spread across other interactions (e.g., from retweets to mentions, replies and quotes) because the accounts know each other, whereas polarisation across hashtags (as themes) might be more diffuse, because they relate to opinions and are not directly associated with individuals. People's opinions (which guide their hashtag use) may have more variety and overlap differently from the individuals they interact with regularly. Thus, we may expect polarisation in one type of interaction to persist into others, but less polarisation in content as the discussion changes to different topics.

Now knowing our labelled groups, our hypotheses moving forward are that:

1. Because the SSM groups are so tightly tied to the use of `#VoteYes` and `#VoteNo` and the previously mentioned strong association between political outlook at progressive issues (such as marriage equality), we expect their interactions to be moderately homophilic and their discussion topics to be strongly homophilic, as they disagree strongly on SSM and have no evidence of other socialisation in the original SSM dataset.

2. For the Arson groups, their retweet network strongly defines their communities based on shared opinions, so we expect strong homophily to be visible in their interactions, however it may be only moderate for mention and quote networks, which can be used to refer to non-community members without much risk of engagement or confrontation (compared with a more direct reply interaction).

Furthermore, given the political and, to some degree, ideological nature of the ArsonEmergency discussion, we expect the Arson groups to also remain strongly polarised in the hashtags they use.

6.4 Methods

We used a variety of measures to uncover polarised groups in social networks, identify their extent and characterise their connectivity and their content. We did this by building networks of accounts linked by interactions (retweets, mentions, replies, and quotes) and the common use of partisan hashtags, and then systematically considering a variety of measures of homophily of the polarised communities within those networks.

6.4.1 Constraints of OSN data

Despite the appeal of social media as a rich data source for sociological research, a number of questions and challenges remain, as we discussed in Chapter 4 and Section 2.3. For example,

- Restrictions on access to OSNs' data via their APIs, such as rate limits and limited data models, mean that social networks built from such data are necessarily limited (Nasim et al., 2016);
- Evidence of inconsistencies in data retrieved using different tools indicate that any given dataset may be incomplete with potentially significant effects on the results of subsequent analyses (Chapter 4);
- A lack of tools to measure confidence in prediction models and other social media analytics may affect the interpretation of their results (Assenmacher et al., 2021, and Section I.1); and
- The extent traditional sociological and psychological theories of human communication are applicable to social networks built from social media data is still an active area of research (e.g., Schroeder, 2018).

That said, interactions on social media, limited in data model though they may be, provide the best portal we have to relevant data and therefore the best opportunity to understand the degree and nature of activity between particular actors at a particular time on a given topic of discussion.

6.4.2 Social networks

Using methods described in Section 3.2.2, a number of social networks are constructed from the datasets to examine how accounts interact, including retweet, mention, reply and quote networks. To then examine how they discuss contentious issues in the datasets, co-hashtag account networks are constructed from tweets including partisan hashtags (described in Section 3.2.2.5). Partisan hashtags typically indicate a position on the concept in the hashtag (e.g. `#<partyname>liars`), which creates an axis of polarisation and is often associated with one of the polarised groups. Politically partisan hashtags are used for political datasets, while for non-political datasets, *faux* partisan hashtags are sought by examining which hashtags are most used that are also exclusive to the polarised groups. Faux hashtags are expected to co-occur with hashtags used by multiple groups, but will still form a strong basis to judge whether the polarisation found elsewhere also appears in the co-hashtag network.

Once the networks are created, we use the backbone layout introduced in Section 3.2.1.7 to visualise and analyse their structure. By darkening edges with high backbone strength and sizing nodes according to account activity or weighted degree, we can highlight not only edges that form the 'backbone' of the network, but also the most strongly connected communities, and they inter-relate. Beyond visually-guided structural analysis, homophily measures (see Section 3.2.1.6) are used to examine how the members of polarised groups interact, and binomial tests with a range of p-value thresholds clarify the measures' statistical significance.

6.5 Results

We address the research questions posed in the Introduction through the lens of the polarised groups identified in the SSM and Bushfires datasets. Initially, we confirm the presence of polarisation between the SSM groups having already established the presence of polarisation in the ArsonEmergency dataset (see Chapter 5). We then consider the activity of the polarised accounts over an extended period of ten months, whether the polarisation spans interaction types and discussion topics, and then whether the polarisation remains regardless of the topic of the discussion.

6.5.1 Polarisation in the SSM discussion

In the SSM tweets, as mentioned above, one set of hashtag filter terms were general in nature, referring to the marriage equality voting activity and politics, namely **`#MarriageEquality`**, **`#SSM`**, and **`#auspol`**, while the second set reflected users' opinions about marriage equality, namely `#VoteYes` and `#VoteNo`. These last two are the defining feature of YES, NO and BOTH accounts. We hypothesised a lot of repulsion between these YES and NO accounts, in particular; specifically, we anticipated relatively high structural cohesion within the groups of accounts who used the same hashtag, and relatively low cohesion among accounts who used opposite hashtags.

To consider this, we retrieved as many followers of YES, NO, and BOTH accounts as possible[6] and constructed a network of their follower relations, ignoring accounts outside the YES, NO and BOTH groups. The resulting network consisted of 2,973 YES nodes, 3,417 NO nodes, and 473 BOTH nodes, and 22,139 directed follower

[6] An account's followers may be unavailable for a variety of reasons, such as the account being protected, suspended or deleted.

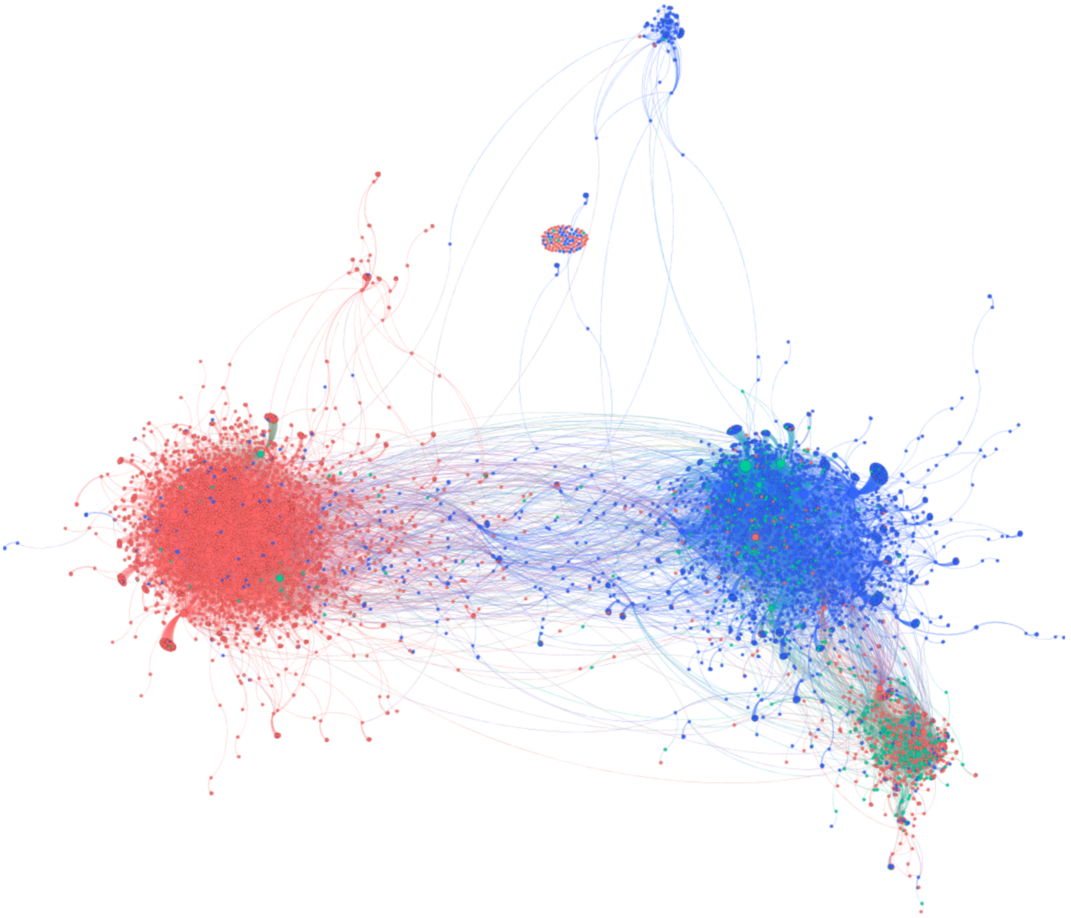
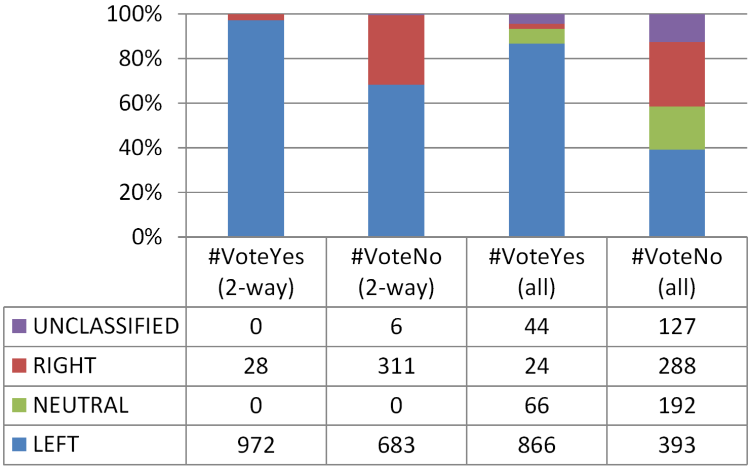


Figure 6.3. The largest component of the network of follow relations of the YES (blue), NO (red) and BOTH (green) accounts in the SSM dataset. The directed edges are coloured according to the following (i.e., source) node. Although BOTH accounts are primarily embedded in the YES community, the YES and NO communities are clearly polarised. (Visualised with Gephi.)

edges, where $\{v_i, v_j\}$ indicates that account v_i follows account v_j. Considering only edges adjacent to a YES or NO node, we find a E-I index of -0.84, implying a high degree of homophily, as expected. Further confirmation of this polarisation is evident in a visualisation of the largest component of the follower network, which includes BOTH nodes (in green) for completeness, shown in Figure 6.3. On this basis, we can confirm the YES and NO groups are polarised, as not only do they use disjoint sets of hashtags but they mostly only follow fellow community members.

Previous works have revealed alignment between people's political leaning and their support for egalitarianism and inclusivity (e.g., Jost et al., 2003; Jost, 2017; Albada et al., 2021), so it is a reasonable to expect a similar pattern on a progressive issue, such as marriage equality. To examine whether the SSM groups corresponded with political alignment, a manual review of 1,000 random samples from YES and NO Election tweets was conducted. Tweets were labelled at two resolutions, one aiming for a simple two-way left-wing or liberal (LEFT), or right-wing or conservative (RIGHT) alignment label, and the other also permitting NEUTRAL and UNCLASSIFIED labels. Tweets



	#VoteYes (2-way)	#VoteNo (2-way)	#VoteYes (all)	#VoteNo (all)
🟪 UNCLASSIFIED	0	6	44	127
🟥 RIGHT	28	311	24	288
🟩 NEUTRAL	0	0	66	192
🟦 LEFT	972	683	866	393

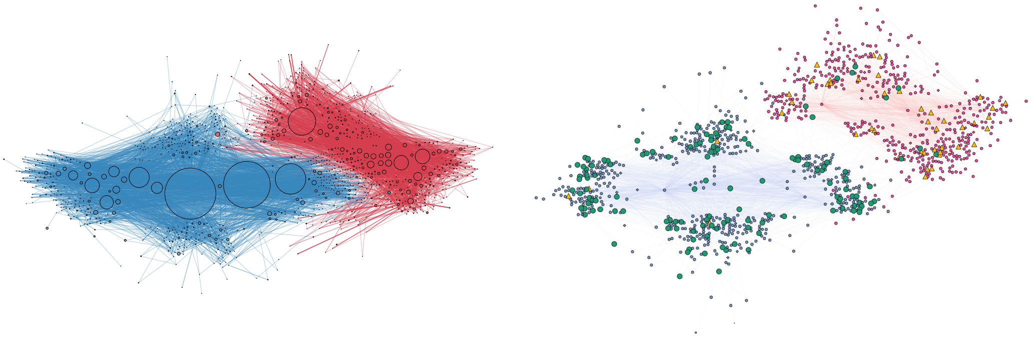
FIGURE 6.4. Results of manually labelling a random subset of tweets by YES and NO members in the Election dataset (1,000 each). The first two columns represent an attempt to assign only LEFT and RIGHT (political alignment) labels to the tweets, while in the second two columns, a further category of NEUTRAL was included. A political alignment for UNCLASSIFIED accounts could not be assigned due to a lack of suitable content.

were judged on their content, and if that were not sufficiently clear, the profile of the tweet's author would be inspected (such content was preserved in the metadata of collected tweets). The results presented in Figure 6.4 indicate that YES members were almost exclusively LEFT-aligned, while the alignment of NO members was more diverse. On deeper inspection, many tweets could be labelled only as NEUTRAL, which is not unexpected in a wide-ranging political discussion, as they often include simple statements of fact. Furthermore, a significant number could not be reasonably classified due to a lack of content. The implications are that the NO members are much more politically diverse than the YES members, who are mostly LEFT-aligned and that the polarisation observed in their use of `#VoteYes` and `#VoteNo` may not be sustained in other political discussions.

Based on this identification of polarised YES and NO accounts, and the analysis of their political stance, we then observed those accounts' behaviour in the lead up to the 2019 Australian federal election, with the aim of testing whether their polarisation on SSM also led to polarisation over the political issues being discussed. Prior to presenting those results, however, we discuss a significant overlap between the SSM groups and the Arson groups.

6.5.2 A chance finding

It was observed that 1,015 SSM accounts from YES, NO and BOTH groups were active in the ArsonEmergency discussion, and that they appeared to still be polarised. Furthermore, of those 1,015 accounts, a full 995 of them appeared in the retweet network, in which the Supporters and Opposers appeared. We highlighted the SSM accounts in a reproduction of the original retweet network visualisation (Figure 6.5) and observed that the groups appeared to have remained polarised. To examine



(a) Arson groups.

(b) SSM groups.

FIGURE 6.5. The retweet network in the ArsonEmergency dataset including just the Arson groups on the left and just the SSM groups on the right. Directed edges are coloured according to their source nodes, and are semi-transparent to manage occlusion.

this statistically, one-tail probability tests for each group were used to confirm that Supporters ↦ NO accounts and Opposers ↦ YES accounts by rejecting the null hypothesis that the polarised groups were independent, at $\alpha = 0.01$.

With this encouragement, we used the ArsonEmergency dataset and earlier Australia-focused datasets, to determine if the SSM and Arson accounts were present and active, and whether the polarisation observed elsewhere was maintained across interaction types and in the content they discussed.

6.5.3 Enduring polarisation

There is some evidence to suggest that if people have strong moral convictions, then they are likely to continue engaging politically (Skitka and Bauman, 2008), and so it is possible, if not likely, that those who participated in the SSM discussion and the ArsonEmergency discussion (both topics with a strong political element) would also have been active in the Australian Twittersphere in the intervening period.

We therefore now turn to examine the presence and polarisation of the SSM and Arson groups in the Election, ArsonEmergency and AFL datasets. To do this, from these datasets we construct and examine retweet, reply, mention and quote interaction networks, as well as content-related networks based on hashtag use.

6.5.3.1 Continued presence

Table 6.4 shows the number and proportion of SSM and Arson accounts in the four datasets. Although some of the proportions drop considerably from the original groups, there are still sufficient absolute numbers to draw conclusions regarding their behaviour (the smallest presence still has 42 members, nearly 10% of the original community). The considerable drop in SSM accounts, especially in the NO group, does raise the question of how these accounts have been used, despite the time between the SSM collection (late 2017) and the earliest of the other datasets (the AFL, in early 2019). Given so many accounts did not participate in these discussions, was it because they were still active but discussing other topics, or is it that they were used only or mostly for the SSM discussion and then left inactive. The great number of YES accounts active in the Election dataset indicates that perhaps NO accounts were used in this single-purpose manner.

Table 6.4. Sizes and proportions of the presence of the polarised groups in the datasets. Bolded figures belong to the original datasets.

	SSM		AFL		Election		ArsonEmergency	
	Late 2017		*March 2019*		*May 2019*		*January 2020*	
YES	**8,623**	**(100.0%)**	376	(4.4%)	3,390	(39.3%)	698	(8.1%)
NO	**7,880**	**(100.0%)**	53	(0.7%)	631	(8.0%)	148	(1.8%)
Supporter	93	(18.7%)	42	(8.5%)	72	(14.5%)	**497**	**(100%)**
Opposer	240	(40.5%)	73	(12.3%)	156	(26.3%)	**593**	**(100%)**

There is a possibility that these accounts have been created for use only in the SSM discussion by those wishing to exacerbate conflict in society around what was already a sensitive topic for many. This tactic has been used in the past, especially at times of political significance (e.g., Hegelich and Janetzko, 2016; CREST, 2017; Graham et al., 2020b; Bot Sentinel, 2021). The presence of many recently created accounts in a discussion is a flag of potential coordinated inauthentic behaviour (Gleicher, 2018). An alternative tactic that makes use of established accounts, avoiding the 'fresh account' indicator, is *narrative switching*, where an account is used to promote a particular narrative for a period, then its tweets are deleted and it goes dormant for a period before being resurrected and changing its presentation (screen handle and profile information) to promote a new narrative (Dawson and Innes, 2019). The fact that the Supporter proportions are much lower than Opposer proportions also suggests that perhaps some Supporter accounts were used in a similar way. The AFL figures are less easy to explain in terms of political engagement, given the discussion's clearly different topic.

6.5.3.2 Networks based on interactions

The results of a systematic examination of the presence and interactivity between the YES and NO and Supporter and Opposer accounts in the AFL, Election and ArsonEmergency datasets are presented in Table 6.5. For each group and interaction in each dataset, we considered how often they interacted amongst themselves (i.e., homophilic connections) and with each other (i.e., heterophilic connections, ignoring the broader network). For both circumstances, we present the sum of the edge weights rather than just the edge count. We used binomial tests to examine the null hypothesis that they had no connection preference, and where p-values are presented, this is the confidence in rejecting the null hypothesis. This provides us with evidence to address all three research questions.

Polarisation on `#ArsonEmergency`. Given polarisation between Supporters and Opposers was first observed in the ArsonEmergency dataset, we can now see that that

Table 6.5. Summary details of the inter- and intra-group interactions by the SSM and Arson polarised groups in networks built from three datasets. Significance p-values are based on using binomial tests with the null hypothesis that the groups had no connection preference.

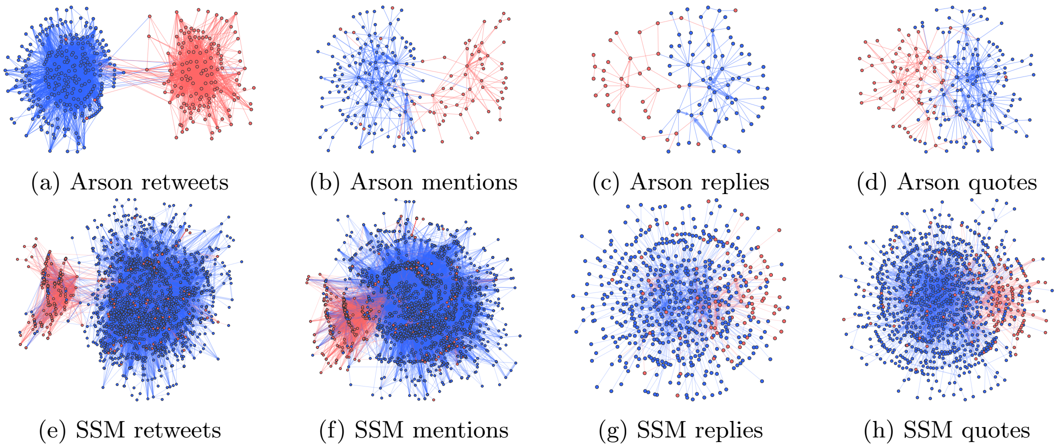
			Election			ArsonEmergency			AFL		
			Target			*Target*			*Target*		
	Network	*Source*	YES	NO	Sig. ($p <$)	YES	NO	Sig. ($p <$)	YES	NO	Sig. ($p <$)
SSM	Retweets	YES	55,792	2,359	*0.0001*	349	27	*0.0001*	45	–	*0.0001*
		NO	2,091	4,677	*0.0001*	17	96	*0.0001*	–	6	*0.05*
	Mentions	YES	5,337	209	*0.0001*	9	2	–	23	1	*0.0001*
		NO	749	261	*0.0001*	24	21	–	–	1	–
	Replies	YES	2,231	106	*0.0001*	5	–	–	21	1	*0.0001*
		NO	133	175	*0.05*	10	10	–	–	–	–
	Quotes	YES	3,303	183	*0.0001*	10	3	–	10	–	*0.01*
		NO	250	335	*0.001*	1	9	*0.05*	1	–	–
	Hashtags	YES	42,683,122	79,068,551	*0.0001*	381	1,258	*0.0001*	652	1,577	*0.0001*
		NO	79,068,551	258,579	*0.0001*	1,258	611	*0.0001*	1,577	105	*0.0001*
			Target			*Target*			*Target*		
	Network	*Source*	Supporters	Opposers	Sig. ($p <$)	Supporters	Opposers	Sig. ($p <$)	Supporters	Opposers	Sig. ($p <$)
Arson	Retweets	Supporters	5,725	23	*0.0001*	3,603	23	*0.0001*	4	–	–
		Opposers	12	11,878	*0.0001*	13	3,006	*0.0001*	–	6	*0.05*
	Mentions	Supporters	509	14	*0.0001*	567	343	*0.0001*	4	–	–
		Opposers	12	562	*0.0001*	28	30	–	–	1	–
	Replies	Supporters	81	5	*0.0001*	288	144	*0.0001*	6	–	*0.05*
		Opposers	6	250	*0.0001*	11	33	*0.01*	–	1	–
	Quotes	Supporters	183	27	*0.0001*	212	50	*0.0001*	–	–	–
		Opposers	9	568	*0.0001*	6	62	*0.0001*	–	–	–
	Hashtags	Supporters	173,488	977,889	*0.0001*	106,433	106,922	–	60	553	*0.0001*
		Opposers	977,889	5,941,413	*0.0001*	106,922	7,875	*0.0001*	553	37	*0.0001*

polarisation extends across the other interactions significantly in all cases except for Opposers' use of mentions. Our findings in Chapter 5 that Opposers focused almost exclusively on retweets, while Supporters also made use of many other interactions is corroborated – in fact, it is clear that Supporters' use of non-retweet interactions was less polarised that it might have been, as they interacted with Opposers between 20% and 35% of the time.

SSM accounts were fewer in number and less active in the ArsonEmergency dataset, and only statistically significantly maintained their polarisation in the retweet network and only amongst NO accounts in the quotes network. Considering the raw numbers, we see NO accounts also used mentions and replies more often than YES accounts, but were relatively balanced in how they used them. In contrast YES accounts connected almost exclusively to other YES accounts with the same interactions.

Polarisation leading up to the federal election. The polarisation in the Election dataset is statistically significant across all groups and interactions, and it is homophilic in all but one condition: NO accounts mention YES accounts more frequently than other NO accounts. This is not necessarily surprising given there are more than five times more YES accounts than NO accounts active in the dataset.

Amongst the Arson groups, in particular, the echo chamber effect (with relation to each other, at least) is stark, with both groups preferring internal to external connections by several orders of magnitude. The smallest ratio of homophilic to heterophilic connections was Supporters' use of quotes (at approximately 6.8), but most were



(a) Arson retweets (b) Arson mentions (c) Arson replies (d) Arson quotes

(e) SSM retweets (f) SSM mentions (g) SSM replies (h) SSM quotes

FIGURE 6.6. The largest components in interaction networks of Supporter (red) and Opposer (blue) accounts (top row) and YES (blue) and NO (red) accounts (bottom row) active in the Election dataset. Edge width describes the backbone strength.

much greater than that. This marked polarisation is also immediately apparent in visualisations of the interaction networks (Figures 6.6a to 6.6d).

The results amongst SSM groups were also all statistically significant, and in all but one case were homophilic, as mentioned above, but the pattern of polarisation differs due to the relative sizes of the groups present (Figures 6.6e to 6.6h). The imbalance in use of interactions is immediately apparent, with the greater number of active YES accounts (presented in Table 6.4) contributing more proportionally than NO accounts across all interaction types. YES accounts outnumber NO accounts five to one, but posted 8.2 times as many retweets, 10.5 times as many mentions, 8.4 times as many replies and 6.9 times as many quotes, so more YES accounts were present in the Election dataset but they were also more active. Furthermore, their echo chamber effect was more pronounced, retweeting, mentioning, replying to and quoting each other over 95% of the time, while NO accounts interacted with each other slightly less than half the time. These findings indicate that: 1) polarisation detected amongst one type of interaction can be present across other types of interaction; and 2) polarisation detected in one issue-related discussion can be found in other issue-related discussions, including across a variety of interactions.

The issues discussed in the ArsonEmergency and Election datasets can be regarded as at least partially political in nature, so the question remains whether the above phenomena persist in non-political discussions. We use the AFL dataset for this contrast, assuming that, whatever their political opinions, any alignment with people's political opinions is likely to be coincidental. Political discussion in the AFL dataset is minimal, and even the most prominent political hashtag is the non-partisan `#auspol`.

Polarisation discussing the AFL. Very few Arson accounts interacted with other accounts in the AFL dataset, but where they did it was strongly homophilic relative to each group. The majority of their connections were to the broader network as sources of interactions (i.e., they reached out to others).

SSM accounts also interacted rarely in the AFL dataset, but the much greater number of YES accounts were strongly homophilic in the connections they made, with respect to the two groups. Again, both groups interacted strongly with the broader network, with some accounts frequently the recipient of interactions rather than just the instigator, as was the case with the Arson group members.

Summary of interaction network findings. In almost all circumstances, the echo chamber effect appears to be maintained to some degree, with internal connections preferred over external ones, especially between Supporters and Opposers. The only circumstance where that effect is reduced is in the NO group's use of replies and mentions and Opposers' use of mentions in the ArsonEmergency dataset, where they more even in their connections. It is possible that some of these mentions were used for aggressive, rather than collegiate, interactions, but analysis of their content is required for this judgement and there were relatively few of these interactions, so any such judgement is unlikely to be indicative of a broader pattern of behaviour.

6.5.3.3 Networks based on content

Results so far indicate the echo chamber effect is strongly maintained across most interactions in most datasets, especially where there is reasonable amount of activity. Here we consider whether the topics also under discussion also exhibit similar patterns of polarisation, and we use hashtags as an indicator of those topics.

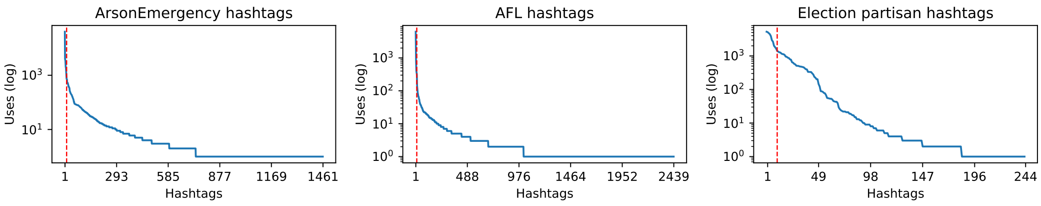


Figure 6.7. Hashtag use distributions for the ArsonEmergency and AFL datasets, and the distribution of the use of partisan and co-occurring hashtags for the Election dataset. The red vertical line indicates the 10^{th} most frequently used hashtag; this and the more frequently used hashtags were removed before building the hashtag networks.

First, however, we must cull the hashtags under consideration, as the high frequency of popular hashtags can hamper the discovery of the structures underlying their use. Instead, as discussed above, we explicitly filter the most frequent hashtags and we additionally make use of partisan and faux partisan hashtags. Examining the distributions of hashtag use in each of the dataset revealed that removing the ten most frequent hashtags in each would be sufficient to avoid the majority of their binding effects (shown as the dashed red vertical lines in Figure 6.7).

Second, we developed the (faux) partisan hashtag sets. In the Election dataset, we identified 44 hashtags of the 200 most frequent as clearly partisan (e.g., `#corrupt<party>` or `#<party>liars`). For the AFL and ArsonEmergency datasets, we identified the ten most frequently used hashtags unique to each group. We considered the tweets containing these hashtags and created semantic networks using all the hashtags that appeared in them (save for the most frequently occurring hashtags, as mentioned above). The number of hashtags considered for each group and dataset is shown in parentheses next to the "Hashtags" label in Table 6.6, which also shows the number of SSM and Arson group accounts present in the resulting networks, and their respective connectivity.

Above, Table 6.5 shows that although the connectivity between the polarised groups was often statistically significant, it was often within groups rather than across groups, meaning the groups most often used the same hashtags. That said, there were large imbalances between the homophilic connections of the groups: YES accounts used YES-specific hashtags far more frequently than NO accounts used NO-specific hashtags in the Election dataset, while the same applied for Opposer accounts with respect to Supporter accounts. In the other datasets, only Opposers' use of Opposer-specific hashtags in the ArsonEmergency dataset stand out, and that is because there are so few connections, relatively (there were 7,875 Opposer–Opposer connections, compared with 106,433 Supporter–Supporter connections and 106,922 Supporter–Opposer connections). Opposers strongly shared hashtags with Supporters, while Supporters also connected internally strongly to a similar degree. In all other cases, heterophilic connections dominated. This suggests that although the groups tended to interact amongst themselves, they often discussed similar topics, even with similar partisan leanings. A deeper exploration of which particular hashtags accounted for these heterophilic connections could reveal further insights regarding the axes of polarisation and agreement between the groups.

Table 6.6. Summary details of SSM and Arson polarised groups in networks built from three datasets. The ‘Hashtag’ networks are the co-hashtag account networks, and the number in parentheses is the total count of the partisan and co-occurring hashtags.

Group	Dataset	Network	Category 1/2 Nodes		Category 1/2 Edge Weights			Homophily		Broader Network		
			Category 1	Category 2	Homophilic	Heterophilic	Sum	E-I index	Assortativity	Nodes	Edge Weights	E-I Index
	Election	Retweet	3,045	592	82.5%	17.5%	64,919	-0.650	0.459	13,585	331,682	0.530
		Mention	1,611	217	61.0%	39.0%	6,556	-0.221	0.247	8,490	51,673	0.701
		Reply	1,005	142	76.1%	23.9%	2,645	-0.523	0.331	4,316	12,397	0.484
		Quote	1,757	234	76.0%	24.0%	4,071	-0.520	0.459	6,315	26,025	0.641
SSM		Hashtags (244)	3,390	631	53.3%	46.7%	47,038,810	-0.066	0.029	4,429	79,327,130	-0.143
Category 1:	ArsonEmergency	Retweet	688	144	88.9%	11.1%	489	-0.778	0.765	12,076	21,526	0.618
YES		Mention	140	67	64.2%	35.8%	56	-0.285	0.160	3,206	7,523	0.881
Category 2:		Reply	99	50	75.0%	25.0%	25	-0.500	0.323	2,041	3,031	0.853
NO		Quote	57	35	83.5%	16.5%	23	-0.669	0.588	1,268	1,542	0.779
		Hashtags (171)	698	148	98.7%	1.3%	998	-0.975	0.970	12,867	63,819	0.870
	AFL	Retweet	276	41	100.0%	0.0%	51	-1.000	1.000	5,735	7,047	0.761
		Mention	177	19	97.9%	2.1%	25	-0.958	0.648	7,740	19,222	0.913
		Reply	118	9	95.5%	4.5%	22	-0.909	0.000	4,815	6,060	0.745
		Quote	78	7	50.0%	50.0%	11	0.000	0.000	1,821	1,670	0.765
		Hashtags (347)	376	53	100.0%	0.0%	757	-1.000	1.000	11,573	76,700	0.871
	Election	Retweet	158	314	99.7%	0.3%	17,638	-0.995	0.983	13,585	331,682	0.643
		Mention	120	243	97.6%	2.4%	1,097	-0.952	0.857	8,490	51,673	0.745
		Reply	86	163	95.9%	4.1%	342	-0.918	0.868	4,316	12,397	0.668
		Quote	109	199	92.8%	7.2%	787	-0.856	0.818	6,315	26,025	0.682
Arson		Hashtags (244)	72	156	60.3%	39.7%	6,908,584	-0.207	0.055	4,429	79,327,130	0.695
Category 1:	ArsonEmergency	Retweet	493	592	99.5%	0.5%	6,645	-0.989	0.988	12,076	21,526	-0.787
Supporters		Mention	288	149	57.0%	43.0%	968	-0.140	0.107	3,206	7,523	0.702
Category 2:		Reply	247	105	70.8%	29.2%	476	-0.417	0.196	2,041	3,031	0.606
Opposers		Quote	190	104	86.0%	14.0%	330	-0.721	0.613	1,268	1,542	0.402
		Hashtags (245)	493	592	98.4%	1.6%	114,797	-0.968	0.965	12,867	424,389	0.250
	AFL	Retweet	30	69	100.0%	0.0%	10	-1.000	1.000	5,735	7,047	0.874
		Mention	19	22	100.0%	0.0%	5	-1.000	1.000	7,740	19,222	0.928
		Reply	16	11	100.0%	0.0%	7	-1.000	1.000	4,815	6,060	0.682
		Quote	7	4	–	–	–	0.000	0.000	1,821	1,670	1.000
		Hashtags (338)	42	73	100.0%	0.0%	97	-1.000	1.000	11,573	81,457	0.949

Visualisation reveals deeper community structures. Using the backbone layout to visualise the co-hashtag account networks (Figure 6.8) makes clear the extent of the isolation of the groups despite their heterophilic connections, as well as the implications of the homophily measures. Nodes are sized according to weighted degree, using the backbone strength for edge weights. Edges are also coloured and sized according to backbone strength.

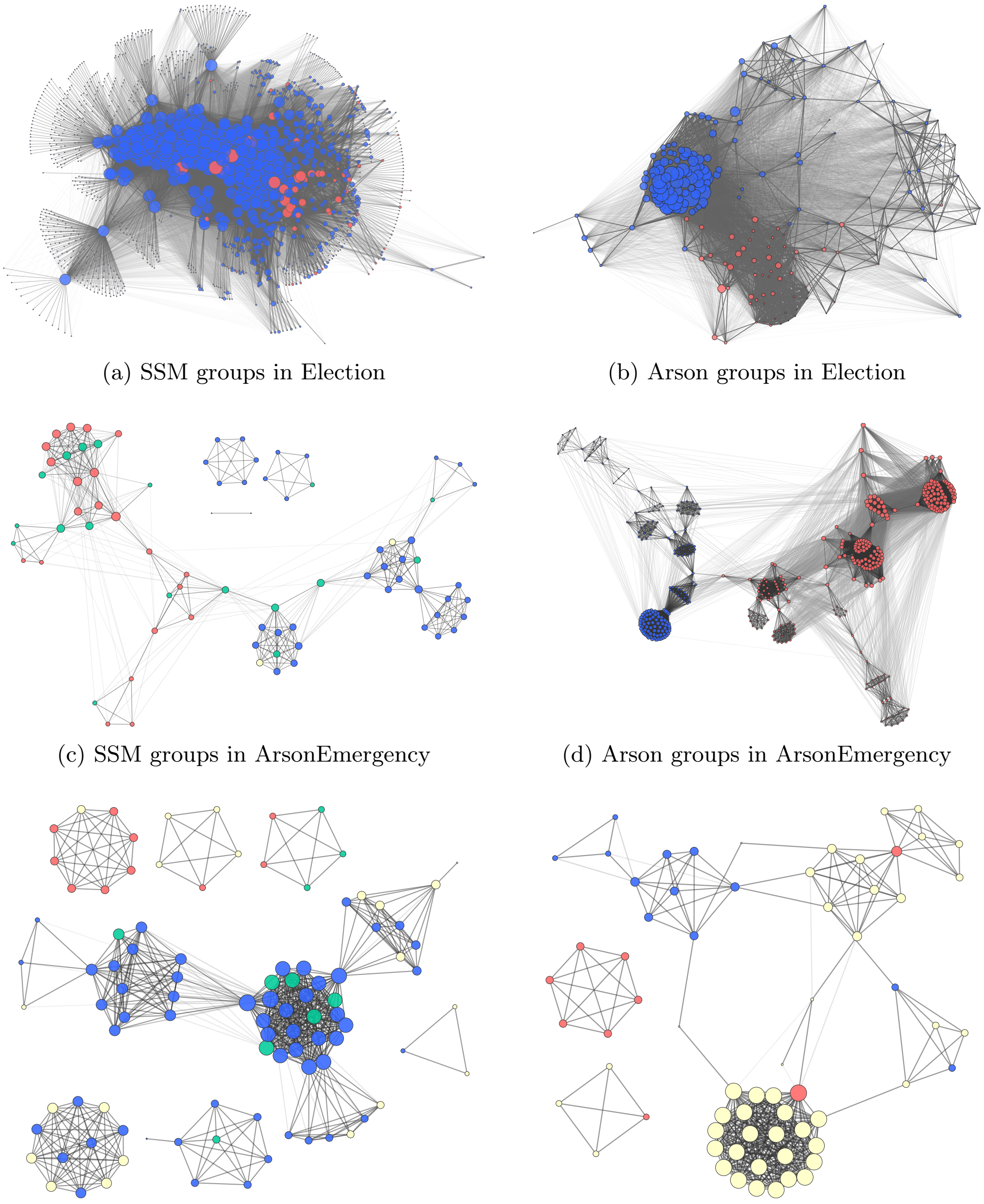
The relatively low homophily of the YES and NO groups during the Election (Figure 6.8a) is primarily due to the relatively small number of NO-only connections (see Table 6.5), which is evident from the NO nodes' dispersed placement throughout the network. Despite the placement, their size indicates they have high centrality and are therefore deeply embedded in the network. In contrast, the Supporter nodes active in the Election in Figure 6.8b are not deeply embedded in the network (according to their sizes) but they clearly form a cluster of their own (to the bottom of the figure). The majority of Opposer nodes reside in a large cluster (top left) and *are* deeply embedded. The relatively moderate E-I Index and assortativity scores in Table 6.6 (-0.207 and 0.055, respectively) indicate that the Supporter nodes are highly connected to the Opposer nodes, which outnumber them, one to two (72 to 156).

Both SSM and Arson groups formed mostly homophilic tight clusters in the Arson-Emergency dataset (Figures 6.8c and 6.8d, respectively), but NO accounts were more often associated with BOTH accounts, which suggests they shared views on the arson narrative, given the alignment between NO and Supporter groups mentioned previously. Opposers and Supporters formed multiple separate clusters, but the most deeply embedded Opposers are clearly strongly concentrated in a single cluster (bottom left), while the deeply embedded Supporters form several groups. Deeper analysis is needed to examine which hashtags linked which clusters.

Similar patterns of hashtag co-use are present in the AFL dataset (Figures 6.8e and 6.8f), but unaffiliated accounts contributed more structure. The nature of the AFL discussion is relatively clustered in general, however, as sports fans discuss specific games, each of which has their own hashtag, which they use along with the `#AFL` hashtag – the top five used hashtags after `#AFL` were `#AflPiesCats`, `#AflDogsSwans`, `#AflLionsEagles`, `#AflFreoNorth` and `#AflDeesPower`, all of which refer to the AFL and two teams that played each other in that round of the competition. It is therefore unsurprising to see some significant degree of clustering in these networks, but the fact that accounts from different groups do not seem to mix in each is notable, and may suggest a strong degree of influence from social circles.

6.5.3.4 Homophily measures and the broader network

The homophily measures in Table 6.6 provide a more nuanced view of the groups' homophily or heterophily in different circumstances than the statistics in Table 6.5. The SSM groups remained moderately to strongly polarised among all interactions



(a) SSM groups in Election

(b) Arson groups in Election

(c) SSM groups in ArsonEmergency

(d) Arson groups in ArsonEmergency

(e) SSM groups in AFL

(f) Arson groups in AFL

Figure 6.8. The largest components in co-hashtag account networks of YES (blue) and NO (red) accounts and Supporter (red) and Opposer (blue) accounts active in the Election, ArsonEmergency and AFL dataset. Green nodes represent accounts that used both `#VoteYes` and `#VoteNo`, and yellow nodes represent OTHER nodes co-mentioning hashtags with affiliated accounts. Node size is determined by the sum of the backbone strength values on incident edges, i.e., degree weighted by backbone strength, indicating each node's embeddedness. Edge width describes the backbone strength.

except for mentions in the Election and ArsonEmergency datasets and the few quotes they posted in the AFL dataset. The Arson groups were mostly highly polarised in all cases except in replies (moderately) and mentions (mildly) in the ArsonEmergency dataset. Regarding their content, polarisation remained in the ArsonEmergency and AFL datasets but was only mild to moderate (E-I Indexes of -0.066 to -0.207 for SSM and Arson groups, respectively) during the Election.

Considering the broader network (as defined above), it is clear that all groups interacted and shared discussions with adjacent non-polarised accounts in all but one circumstance. Interestingly, the homophily in topics discussed increased modestly for the SSM groups in the Election, suggesting some further divide in the extra 4k OTHER accounts.

6.5.4 Addressing the research questions

We now directly consider the research questions posed in the Introduction.

RQ1 *Do Twitter accounts remain involved in Australian discussions for extended periods?*

We have shown that a significant number of accounts have remained active in the Australian Twittersphere over a number of years, with nearly 40% of SSM YES accounts active nearly two years later in the lead up to the federal election and some hundreds still active in early 2020 during the Australian "Black Summer" bushfires. Conversely, some hundreds of ArsonEmergency group members had also been active during the SSM discussion in 2017, exhibiting a high degree of alignment with 65 of 93 Supporters using `#VoteNo` (24 of them also used `#VoteYes`) and 152 of 240 Opposers using `#VoteYes` (33 of them also used `#VoteNo`).

RQ2 *Is polarisation observed in one interaction type present across other interaction types?*

To the greater extent, the echo chambers observed in the Arson and SSM groups persisted through most interaction types according to E-I Index scores, at least moderately. E-I Index scores rose dramatically (towards heterophily) when the broader network was considered, indicating that the polarised groups were primarily polarised with regard to one another, and did, in fact, interact strongly with those outside their groups.

RQ3 *Do accounts found to be polarised in some discussions maintain their polarisation in different discussions, and does the theme of the discussion impact this polarisation?*

In contrast to the interaction networks, analysis of the common use of partisan hashtags revealed more heterophily in the Election dataset (in which a great variety of political issues were discussed), leading to the conclusion that although the groups mostly interacted amongst themselves, they discussed similar partisan topics and so probably also held similar positions on those topics (as described by the stance of the hashtags). The mix of political leanings exhibited by the NO accounts in the Election may have contributed to the overall greater heterophily in the Election dataset.

In contrast to the Election, homophily remained very high in the topics discussed in the other datasets. For the ArsonEmergency discussion, this is likely due to the high alignment between the SSM and Arson groups, and for the AFL discussion, it is likely due to the match-specific nature of parts of the discussion.

When considering the broader network, heterophily in discussions topics was mostly very strong across all datasets, except for when it was moderately homophilic and heterophilic amongst the SSM groups in the Election and the Arson groups in the ArsonEmergency datasets, respectively. This implies the SSM groups had their own distinct discussion topics during the election, which they shared amongst themselves but not with the broader community, and which might also provide an avenue for further integration.

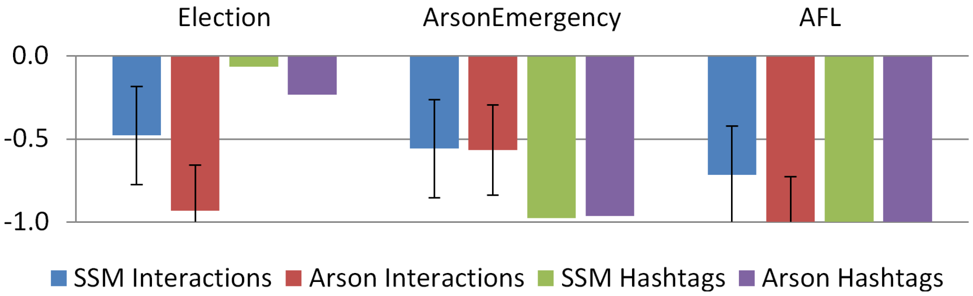


Figure 6.9. E-I Index scores for the SSM and Arson groups in each of the datasets. The interaction scores have been averaged, and the error bars indicate the standard deviation across all scores for that group. The hashtag scores refer to the co-hashtag account networks. Scores approaching -1.0 indicate greater homophily, where 1.0 indicates entirely heterophilic and 0 indicates balance between homophilic and heterophilic edges (i.e., by the sum of their weights). As the hashtag bars represent single values, no error bars are required.

The interaction- and content-based E-I Index scores of the polarised groups revealed the groups interacted differently to how they discussed topics, raising the important question why. We summarise the E-I Index scores for each group and dataset in Figure 6.9, averaging the interaction network scores and contrasting them with the scores from the corresponding partisan co-hashtag account networks. First, as mentioned above, polarisation varies from moderate to high across all interaction types for both SSM and Arson groups in all datasets, but is particularly pronounced between the Arson groups in the Election and AFL datasets. Second, the use of partisan hashtags during the Election was remarkably even compared to the other datasets. In fact, partisan hashtag use was almost entirely homophilic in the ArsonEmergency and AFL datasets. This could be explained in at least two ways. The first is that, although the partisan hashtags clearly align with political camps, the hashtags that co-occur with them in tweets might overlap significantly between the groups. Given the large number of them in the Election dataset (200), it is possible that there are many opportunities for accounts in different groups to use the same one. We might expect that, if this were the case, then the number of co-occurring hashtags in the other datasets should be low, however this is not what we find. Both had hundreds of co-occurring hashtags (see the counts next to the 'Hashtags' labels in Table 6.6). The second possibility is that the polarisation between the SSM and Arson groups is less to do with political opinions and more to do with social circles. People may have used `#VoteNo` in the SSM dataset for a variety of non-political reasons and factors, including religion, culture or general conservatism, and therefore may share the political opinions of many in the YES group. This orthogonality is perhaps less likely in the Arson group, given the motivation for being a Supporter or Opposer is easier to attribute to political outlook, and we can see that the partisan hashtag use E-I Index score of the Arson groups in the Election dataset reflects this, being slightly more homophilic than that of the SSM groups.

6.5.5 Addressing the hypotheses

The statistical support shown in Table 6.6 for the hypotheses presented at the end of the Datasets section is mixed.

SSM groups were very polarised in discussion topics in the ArsonEmergency and AFL datasets, but much less so in the Election dataset, and their interactions varied from moderately to highly homophilic (especially amongst the few present in the AFL dataset). In this way, the interactions observed indicate socially connected groups, while the content connections suggest they shared discussion topics strongly, even when they may have been partisan in nature.

Similarly, Arson groups interacted in strongly polarised ways in the Election and AFL datasets, but were only strongly homophilic in retweets and quotes in the ArsonEmergency dataset, where they were initially identified. Their connections were only weakly to moderately homophilic in their mentions and replies, respectively, in that dataset. Their use of content, however, was strongly polarised in all but the Election, where again they seemed to often share partisan discussion topics, but only within established social groups. The lower homophily in the reply and networks seems to suggest that Supporters and Opposers were willing to bridge the gap between the groups, but this may be a reflection of direct conflict, rather than genuine debate, based on the degree of aggressive behaviour observed in the tweets in Chapter 5.

6.6 Discussion

This work touches on a variety of research questions, including how people decide their position in a social space when presented with conflicting opinions about contentious topics, how political ideology drives people's stance on issues, and what could make an echo chamber transient or persistent. How behavior is affected by the social relations is described as one of the classic questions of social theory (Granovetter, 1985). A listener who is not an active part of the conversation experiences the occurrences of the others' actions as "events occurring in outer time and space" (Garfinkel, 2005). This view on shared events is a motivating factor for studying interactions which do not share physical presence, such as those in the online space. Studies have shown that people are influenced by online interactions, for instance, when it comes to making decisions about vaccination, opinions about vaccination on Twitter can act as a precursor to making a practical decision (Dunn et al., 2015).

Our analysis of the structural properties of a variety of networks based on their follower relations, interactions and hashtag use suggest that accounts expressing positive opinions about marriage equality (in the SSM dataset) or the arson narrative (in the ArsonEmergency dataset) were more closely connected in some parts of these networks leading to greater statistical homophily. Similar patterns held for those arguing against marriage equality and the arson narrative. A number of factors could be involved in causing this connection preference, some of which have been previously identified in the literature (Rogers and Bhowmik, 1970). These include that communication is more effective amongst those who share common meanings, attitudes and beliefs, and that many people prefer communicating with others who are similar in social status, education and beliefs. Use of common information sources leads to a perception of greater trustworthiness and credibility within a community, while heterophilic interaction risks distortion of the message and potential for cognitive dissonance inasmuch as new messages can conflict with current beliefs. Such interactions can be valuable, however, helping to break the filter bubbles, exposing people to new ideas and points of view and challenging them to critically evaluate their own.

Based on our observations, the primary cause for persistent polarisation may be the existence of social groups more than any difference of opinion. As discussed above, a reason for homophilic connections is similarity between conversants, but that similarity may be due to being friends or acquaintances, rather than on less personal attributes, such as education or social status (as noted elsewhere in the literature, e.g., Rogers and Bhowmik, 1970). This is reinforced by the fact that the use of partisan hashtags in the Election dataset was so evenly distributed, suggesting that although the accounts interacted in what might be called echo chambers, they often discussed similar topics and held similar partisan views. In that sense, they may be more accurately described as social circles. That said, in other discussions, not only did they not interact, but they did not share content either, particularly in the ArsonEmergency dataset, so concerns that people are cutting themselves off from alternative viewpoints remain. Evidence from heterophilic connections in the ArsonEmergency dataset also aligns with observations of a high degree of antagonism (recounted in Section 5.4.3).

More broadly, the criticism offered by Bruns (2019b) regarding the lack of clear definitions for the terms ‘echo chamber’ and ‘filter bubble’ is well-founded, but these labels still hold value for communicating high level concepts. We offer a conceptual definition of an echo chamber as a community formed around a shared opinion on a particular issue or discussion topic, within which that same opinion is reinforced as part of the community’s interactions and discussion. This is consistent with the literature (Garimella et al., 2018b). The members still interact with those outside the echo chamber, but may do so by also discussing other issues, which is in line with Simmel’s theory of intersection ‘social circles’ (Simmel, 1908). Echo chambers can be identified as communities whose content, when analysed, is highly focused and of a similar opinion (e.g., through the use of partisan hashtags, which declare a stance on an issue), but whose members still interact frequently with those outside the community. The members of a filter bubble, in contrast, lack significant interaction with those outside the community (i.e., instigated from within). This situation is often blamed on OSN recommendation algorithms in pursuit of personalised information offerings (Pariser, 2012; Massanari, 2016; Bruns, 2019b). This kind of connectivity can be observed with network analysis and the discussion topics and stances can be identified with content analysis, but hard and fast rules such as ‘filter bubble members never interact with new content’ are too strict to be of use in the highly varied world of social media. Of course, these definitions are limited to the OSNs (and other communication environments) available for analysis. A person might only ever tweet about arson, but will still interact with family, friends and workmates outside of Twitter, so a filter bubble is only likely to occur in the most extreme of circumstances (e.g., isolated cults).

6.6.1 Critique and opportunities

There are a number of ways to improve the approaches we have taken in this study, including the following considerations.

The weights in the co-hashtag account networks are calculated as the sum of the products of each pair of accounts’ uses of a common hashtag, which may potentially inflate weights and not reflect imbalanced use between the members of the pair. Others (e.g., Magelinski et al., 2021) have used the minimum instead, or a more sophisticated calculation may be warranted.

In fact, there may be benefit in additionally scaling edge weights by user activity: if an account is very active, it might co-use a hashtag more often just by chance, which will connect it to other accounts using the hashtag with very heavy edge weights. Taking relative activity into account may lighten these edges.

The manner in which partisan hashtags are chosen is also, to some degree, a subjective activity. Furthermore, the faux partisan hashtags are highly likely to generate polarised groups, after all that is how they are chosen. We have, however, revealed interesting findings in the hashtags that co-occur with them, so there is merit in the approach but deeper investigation is required. For example, these commonalities could be studied separately by ignoring the specifically partisan hashtags as the co-hashtag account network is created from the filtered tweets.

The assumption that homophilic and heterophilic interactions are all equally representative of civil communication is lacking, and deeper examination is required to determine to what degree the interactions are positive or negative. In Chapter 5, we found high degrees of aggression between the Arson groups, so it is possible that that aggression exists in the heterophilic connections in the other datasets. Methods exist for examining online group conflict that could be applied for this purpose (e.g., Kumar et al., 2018; Datta and Adar, 2019). An analysis of URL-sharing behaviour in these datasets may also reveal shared or divided stances on issues, as defined by the content referred to by the URLs.

6.7 Conclusion

Echo chambers on OSNs provide fertile ground for misinformation and polarisation on social and political issues, which can influence offline behaviour with real-world effects such as vaccine hesitation and even violence. This study began by identifying the SSM groups, a pair of polarised groups in the Twitter discussion surrounding the 2017 Australian postal survey on marriage equality. The activities of the SSM and Arson groups were tracked over several Twitter datasets spanning a ten month period and a variety of discussion topics. The aim of the study has been to characterise the nature of their polarisation in terms of the interactions used and the topics discussed to determine if such communities are persistently polarised, or whether they mix over time and as the issues at hand change.

Our findings reveal that persistent communities of Australian Twitter users exist and remain polarised in the social groups they form over periods of several years, but that the topics they discuss are often similar, even in the context of partisan topics. Furthermore, these polarised groups interact strongly with those outside their groups even while they avoid each other, which offers hope that the echo chambers they form between themselves can be pierced and infiltrated through further encouraging and facilitating engagement with the broader online community.

6.8 Part Summary

In this Part, Chapter 5 addressed **TRQ2**, presenting methods to identify and characterise the behaviour of polarised communities in a contentious online discussion, and demonstrated them in a phased case study. We found that communication strategies differed between communities and observed how the dominant narrative changed over the discussion from supporting the arson theory to refuting it, despite the efforts to directly engage by Supporters of the theory. Following up on the Arson groups, a longitudinal study addressing **TRQ3** confirmed that many of their members remained active for years in other contentious discussions and that they overlapped with groups polarised over marriage equality. Further, the different groups largely aligned in their polarisation, but their partisan hashtags indicated they may agree on more issues than they disagree on; if their social circles also overlapped, it could show they have much in common.

Our focus now moves, in Part III, to finding groups of accounts coordinating their behaviour, as they are capable of consolidating and exacerbating the polarisation we observed in this Part, intensifying information disorders with propaganda.

Part III

The Hunt: Detecting Amplified Influence

The presence of polarisation observed in online communities in Part II mirrors the divisions in opinion communities that have formed during debates on social, political and ideological issues, observed in the public sphere since well before the internet was created. These debates, where the divisions edge towards being irreconcilable, form cracks and fissures in society that can be exploited by nefarious state, non-state and other malicious actors. The connectivity, anonymity and opportunity for automation mean that small groups of individuals can have an out-sized voice in the online debate. It is the discovery of these groups to which we now turn.

A shift towards group-focused inauthentic behaviour detection methods began approximately a decade ago (Cresci, 2020), as evidence emerged that malicious campaigns consisted increasingly of groups of accounts engaging in coordinated inauthentic behaviour (e.g., starting with the analysis of the 2010 special election in Massachussetts, Metaxas and Mustafaraj, 2012). Yu et al. (2015) suggested that "Traditional anomaly detection on social media mostly focuses on individual point anomalies while anomalous phenomena usually occur in groups" (p.1, Yu et al., 2015). Similarly, Şen et al. (2016)'s community detection algorithm was designed to identify *focal structures* of "influential sets of individuals" (p.1, Şen et al., 2016), rejecting the notion that centrality measures of individuals alone were sufficient to explain influence in real-world social networks. Specifically with regard to inauthentic behaviour, emphasis has shifted from identifying automated accounts such as bots to systems that prioritise finding groups of accounts (automated, hybrid or human-driven) that engage in "orchestrated activities" (Grimme et al., 2018). Cresci (2020) explicitly encouraged "embracing the complexity of deception, manipulation and automation by devising unsupervised techniques for spotting suspicious coordination" (p.82, Cresci, 2020). Cresci (2020) also noted that, although group-detection approaches had begun to emerge around 2012, well before the public and OSNs had acknowledged or understood the danger of coordinated inauthentic behaviour, few approaches directly attempted *general* detection of coordination.

In this Part, in addressing **TRQ4**, we present our own general approach to detecting groups engaging in suspicious coordinated behaviour. Our method identifies anomalies in common interactions as they are used to disseminate content and narratives and amplify their influence. Given inauthentic coordination is ill-defined (Douek, 2020; Cresci, 2020), a significant portion of our contribution, beyond the generalised detection pipeline proposal and a community extraction method, is a wide variety of validation techniques aimed at building trust in our results. These also serve to further characterise the behaviour and attributes of coordinating groups of online accounts.

Chapter 7

Coordinated Amplification

Political misinformation, astroturfing and organised trolling are online malicious behaviours with significant real-world effects that rely on making the voices of the few sounds like the roar of the many. These are especially dangerous when they influence democratic systems and government policy. Many previous approaches examining these phenomena have focused on identifying campaigns rather than the small groups responsible for instigating or sustaining them. To reveal latent (i.e., hidden) networks of cooperating accounts, we propose a novel temporal window approach that can rely on account interactions and metadata alone. It detects groups of accounts engaging in various behaviours that, in concert, come to execute different goal-based amplification strategies, a number of which we describe, alongside other inauthentic strategies from the literature. The approach relies upon a pipeline that extracts relevant elements from social media posts common to the major platforms, infers connections between accounts based on criteria matching the coordination strategies to build an undirected weighted network of accounts, which is then mined for communities exhibiting high levels of evidence of coordination using a novel community extraction method. We address the temporal aspect of the data by using a windowing mechanism, which may be suitable for near real-time application. We further highlight consistent coordination with a sliding frame across multiple windows and application of a decay factor. Our approach is compared with other recent similar processing approaches and community detection methods and is validated against two politically relevant Twitter datasets with ground truth data, using content, temporal, and network analyses, as well as with the design, training and application of three one-class classifiers built using the ground truth; its utility is furthermore demonstrated in a case study of contentious online discussions.

*The content of this chapter was originally published in Publication **II** and expanded in Publication **VII**.*

7.1 Introduction

Online social networks (OSNs) have established themselves as flexible and accessible systems for activity coordination and information dissemination. This benefit was illustrated during the Arab Spring (Carvin, 2012) but inherent dangers are increasingly apparent in ongoing political interference and disinformation (Howard and Kollanyi, 2016; Ferrara, 2017; Keller et al., 2017; Neudert, 2018; Singer and Brooking, 2019; Nimmo et al., 2020). Modern Strategic Information Operations (SIOs) are participatory activities, which aim to use their audiences to amplify their desired narratives, not just receive it (Starbird et al., 2019). The widespread use of social media for political communication and its identity-obscuring nature have made it a prime target for politically-driven influence, both legitimate and illegitimate. Through cyclical reporting (i.e., social media feeding stories and narratives to traditional news media, which then sparks more social media activity), social media users can unknowingly become "unwitting agents" as "sincere activists" of concerted operations (Benkler et al., 2018; Starbird and Wilson, 2020). The use of *political* bots and trolls to influence the framing and discussion of issues in the mainstream media (MSM) remains prevalent (Bessi and Ferrara, 2016; Woolley, 2016; Woolley and Guilbeault, 2018; Rizoiu et al., 2018; Cresci, 2020). The use of bots and sockpuppet accounts to amplify individual voices above the crowd, sometimes referred to as the *megaphone effect*, requires coordinated action and a degree of regularity that may leave traces in the digital record.

Relevant research has focused on high level analyses of campaign detection and classification (Lee et al., 2013; Varol et al., 2017b; Alizadeh et al., 2020), the identification of botnets and other dissemination groups (Vo et al., 2017; Woolley and Guilbeault, 2018), and coordination at the community level (Kumar et al., 2018; Hine et al., 2017; Cresci, 2020). Some have considered generalised approaches to social media analytics (e.g., Graham et al., 2020c; Fazil and Abulaish, 2020; Nizzoli et al., 2021; Pacheco et al., 2021), but unanswered questions regarding the clarification of coordination strategies and their detection remain. Forensic studies of SIOs and other influence campaigns using these strategies (e.g., Benkler et al., 2018; Jamieson, 2020; Nimmo et al., 2020) currently require significant human input to reveal the covert ties underpinning them, and could benefit greatly from enhanced automation.

In this chapter, we present a novel approach to detect groups engaging in potentially coordinated amplification activities, revealed through anomalously high levels of coincidental behaviour. Links between the group members are inferred from behaviours that, when used intentionally, are used to execute a number of identifiable coordination strategies. We use a range of techniques to validate our new approach on two relevant datasets, as well as comparison with ground truth and a synthesized dataset, and show it successfully identifies coordinating communities.

Our approach infers ties between accounts to construct *latent coordination networks* (LCNs) of accounts, using criteria specific to different coordination strategies, which are based on features common to major OSNs. The accounts may not be directly connected, thus we use the term 'latent' to mean 'hidden' when describing these connections. The inference of connections is performed solely on the accounts' interactions, i.e., not their content or friending/following behaviour, only metadata and temporal information, though it could incorporate them.

Highly coordinating communities (HCCs) are then detected and extracted from the LCN. We propose a variant of *focal structures analysis* (FSA, Şen et al., 2016) to do this, in order to take advantage of FSA's focus on finding influential sets of nodes in a network while also reducing the computational complexity of the algorithm. A window-based approach is used to enforce temporal constraints.

The following research questions guided our evaluation:

RQ1 How can HCCs be found in an LCN?

RQ2 How do the discovered communities differ?

RQ3 Are the HCCs internally or externally focused?

RQ4 How consistent is the HCC messaging?

RQ5 What evidence is there of consistent coordination?

RQ6 How well can HCCs in one dataset inform the discovery of HCCs in another?

This chapter provides significant methodological detail and experimental validation, and a case study in which the technique is applied to new real-world Twitter datasets relating to contentious political issues, as well as consideration of algorithmic complexity and comparison with several similar techniques. Prominent among the validation provided is the use of machine learning classifiers to show that our datasets contain similar coordination to our ground truth, and the application of a sliding frame across the time windows as a way to search for consistent coordination.

We first provide a focused overview of relevant literature, followed by a discussion of online coordination strategies and their execution. Our approach is then explained, and its experimental validation is presented. Following the validation, the algorithmic complexity and performance of the technique is presented, a case study is explored, demonstrating the utility of the approach with real-world politically-relevant datasets, and we compare our technique to those of Pacheco et al. (2021), Graham et al. (2020c), Nizzoli et al. (2021) and Giglietto et al. (2020b).

7.1.1 Online information campaigns and related work

As we discussed in Chapter 2, targeted marketing and automation coupled with anonymity provide the tools required for potentially significant influence in the online sphere, perhaps enough to swing an election, but certainly enough to be associated with real-world violence (The Soufan Center, 2021b; Karell et al., 2021). Here, we recap a snapshot of literature relevant to this chapter, but a more detailed exploration of literature on coordination can be found in Section 2.6, while a discussion of inauthentic online behaviour is offered in Section 2.2.

Effective influence campaigns relying on these capabilities will somehow coordinate the actions of their participants. Early work on the concept of coordination by Malone and Crowston (1994) described it as the dependencies between the tasks and resources required to achieve a goal. One task may require the output of another task to complete. Two tasks may share, and require exclusive access to, a resource or they may both need to use the resource simultaneously.

At the other end of the spectrum, sociological studies of influence campaigns can reveal their intent and how they are conducted, but they consider coordination at a much higher level. Starbird et al. (2019) highlight three kinds of campaigns: *orchestrated*, centrally controlled campaigns that are run from the top down (e.g., paid teams, Chen, 2015; King et al., 2017); *cultivated* campaigns that infiltrate existing issue-based movements to drive them to particular extreme positions (e.g., encouraging political violence during elections, Nimmo et al., 2020; Jamieson, 2020; The Soufan Center, 2021b); and *emergent* campaigns arising from enthusiastic communities centred around a shared ideology (e.g., conspiracy groups and other fringe movements). Though their strategies differ, they use the same online interactions as normal users (e.g., posts, shares, mentions, hashtags, URLs), but their patterns differ. Fundamentally, however, they rely on influencing others by spreading an agenda-driven message or narrative.

At the scale of nation states, multiple disinformation campaigns may be run as part of an operation, each with different targets and different intended outcomes. The 2016 US presidential election has received significant academic (as well as political and diplomatic) attention, and deep analysis of the interference by Russia has revealed a variety of such campaigns were employed to promote Donald Trump, detract from Hilary Clinton, sow doubt in the country's democratic system and generally exacerbate divisions in society (Benkler et al., 2018; Mueller, 2018; Jamieson, 2020). Furthermore, much of the social media activity in particular was conducted by accounts made to look like average Americans, including "personable swing-voters" (p.134, Jamieson, 2020) and comparatively simple analyses of individual accounts over long periods has revealed how they were used to build audiences susceptible to their narratives (Dawson and Innes, 2019). America is clearly not the only target—campaigns have been directed across any national border as well as within (Woolley and Howard, 2018; Singer and Brooking, 2019; Nimmo et al., 2020), with influence operations observed in as many as 81 countries in 2020 (Bradshaw et al., 2021). Many of the analyses mentioned in these works rely on direct connections between entities (e.g., Benkler et al.'s mentions of articles and YouTube videos and Nimmo et al.'s follower networks, and studies of retweet and mention networks in chapters of Woolley and Howard's book), but Jamieson makes it clear that covert or at least indirect behaviour-related

TABLE 7.1. Detecting inauthentic behaviour in the computer science literature.

Application	Relevant research
Automation	Ferrara et al. (2016), Grimme et al. (2017), Cresci (2020), and Latah (2020)
Campaigns	
- by content	Lee et al. (2013), Assenmacher et al. (2020), Alizadeh et al. (2020), and Graham et al. (2020c)
- by URL	Ratkiewicz et al. (2011), Cao et al. (2015), Giglietto et al. (2020b), and Broniatowski (2021) Yu (2021), Ng et al. (2021), Graham et al. (2021), and Bruns et al. (2021)
- by hashtag	Ratkiewicz et al. (2011) and Burgess and Matamoros-Fernández (2016) Varol et al. (2017b) and Weber et al. (2020a), and Chapter 5
Synchronicity	Chavoshi et al. (2017), Hine et al. (2017), Nasim et al. (2018), and Mazza et al. (2019) Dawson and Innes (2019), Pacheco et al. (2020), and Magelinski et al. (2021)
Communities	Keller et al. (2017), Vo et al. (2017), Morstatter et al. (2018), and Gupta et al. (2019)
Political bots	Bessi and Ferrara (2016), Woolley (2016), Hegelich and Janetzko (2016), and Ferrara (2017) Rizoiu et al. (2018) and Woolley and Guilbeault (2018)

connections were a key part of the Russian operation during the 2016 US presidential election.

Disinformation campaigns effectively trigger human cognitive heuristics, such as individual and social biases to believe what we hear first (*anchoring*) and what we hear frequently and can remember easily (*availability* cascades, Tversky and Kahneman, 1973; Kuran and Sunstein, 1999); thus the damage is already done by the time lies are exposed. This is especially true if they are promoted under the guise of authority, such as from accounts purporting to be media outlets, like `@TodayPittsburgh` or `@KansasDailyNews` (p.188, Miller, 2018). Persuasive messaging also relies on emotion, especially fear, and appeals to religion (Jamieson, 2020), and have been effective even when such claims border on the ridiculous and conspiratorial (The Soufan Center, 2021b; Brazil, 2020). Recent experiences of false information moving beyond social media during Australia's 2019-2020 bushfires highlight that identifying these campaigns as they occur can aid OSN monitors and the media to better inform the public (Graham and Keller, 2020; Keller et al., 2020, and Chapter 5).

In between task level coordination and entire SIOs, at the level of social media interactions, as demonstrated by Graham and Keller (2020) and Keller et al. (2020), we can directly observe the online actions and effects of such activities, and infer links between accounts based on pre-determined criteria. Relevant efforts in computer science have focused on a variety of methods and domains (see Table 7.1 for a summary and Sections 2.2 and 2.6 for detail). These efforts have uncovered a new field of research: the computer science study of the "orchestrated activities" of accounts in general, as Grimme et al. (2018) put it, regardless of their degree of automation (Cresci et al., 2017b; Alizadeh et al., 2020; Nizzoli et al., 2021; Vargas et al., 2020). It must be noted that bot activity, even coordinated activity, may be entirely benign and even useful (Ferrara et al., 2016; Graham and Ackland, 2017).

Though some studies have observed the existence of strategic behaviour in and between online groups (e.g., Keller et al., 2017; Kumar et al., 2018; Hine et al., 2017;

Keller et al., 2019; Giglietto et al., 2020b; Broniatowski, 2021), the challenge of identifying a broad range of their interaction strategies and their underpinning execution methods remains to be fully explored, especially as new strategies are constantly be devised (Nimmo et al., 2020). Not all strategies will be used equally, however, and detecting campaigns based on coordinated amplification will remain a valid concern, due to its effectiveness (Paul and Matthews, 2016).

Inferring social networks from OSN data requires attendance to the temporal aspect to understand information (and influence) flow and degrees of activity (Holme and Saramäki, 2012). Real-time processing of OSN posts can enable tracking narratives via text clusters (Carnein et al., 2017; Assenmacher et al., 2020), but to process networks requires graph streams (McGregor, 2014) or window-based pipelines (e.g., Graham et al., 2020a; Pacheco et al., 2021; Magelinski et al., 2021), otherwise processing is limited to post-collection activities (Graham et al., 2020c; Alizadeh et al., 2020; Vargas et al., 2020; Pacheco et al., 2021; Ng et al., 2021).

This chapter contributes to the identification of interaction-based strategic coordination behaviours observable over relatively short time frames, along with a general technique to enable detection of groups using them, couple with a variety of validation methods. As such, this enhances the toolbox of techniques available to higher level explorations of information campaigns and operations (e.g., Benkler et al., 2018; Jamieson, 2020; Nimmo et al., 2020; The Soufan Center, 2021b).

7.2 Coordinated Amplification Strategies

Influencing others online, especially on political and social issues, relies on two primary mechanisms to maximise the reach of a given narrative thus amplifying its effect: *mass dissemination* and *engagement*. For example, an investigation of social media activity following terrorist attacks in the UK in 2017 identified accounts promulgating contradictory narratives, inflaming racial tensions and simultaneously promoting tolerance to sow division (CREST, 2017). By engaging aggressively, the accounts drew in participants who then spread the message.

Mass dissemination aims to maximise audience, to convince through repeated exposure and, in the case of malicious use, to cause outrage, polarisation and confusion, or at least attract attention to distract from other content.

Engagement is a form of dissemination that solicits a response. It relies on targeting individuals or communities through mentions, replies and the use of hashtags as well as rhetorical approaches that invite responses (e.g., inflammatory comments or, as present in the UK terrorist example above and observed by Nimmo et al. (2020), pleas to highly popular accounts).

A number of online coordination strategies have been observed in the literature making use of both dissemination and engagement to amplify their effect, including specifically

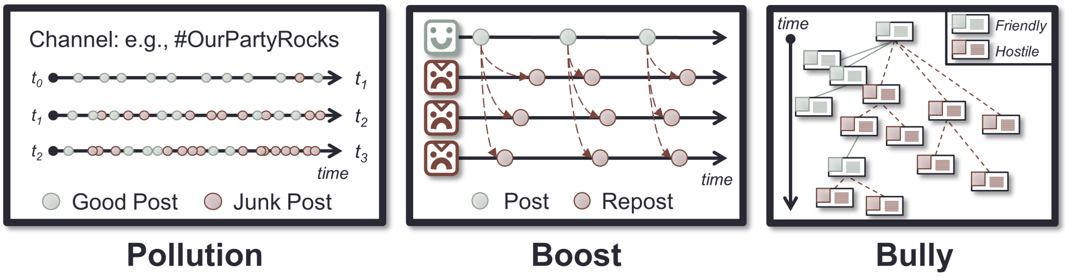


FIGURE 7.1. Patterns matching the mentioned coordinated amplification strategies. Green posts and avatars are benign, whereas red or maroon ones are malign.

TABLE 7.2. Coordinated amplification strategies

Pollution	Flooding a community with repeated or objectionable content, causing the OSN to shut it down.
Observed by	Ratkiewicz et al. (2011), Woolley (2016), Hegelich and Janetzko (2016), and Hine et al. (2017) Nasim et al. (2018), Fisher (2018), Mariconti et al. (2019), and Graham et al. (2020b)
Boost	Heavily reposting or duplicating content to make it appear popular.
Observed by	Ratkiewicz et al. (2011), Cao et al. (2015), Varol et al. (2017b), and Vo et al. (2017) Gupta et al. (2019), Keller et al. (2019), Mazza et al. (2019), and Graham et al. (2020c) Assenmacher et al. (2020), Yu (2021), and Ng et al. (2021)
Bully	Groups engaging in organised harassment of an individual or community.
Observed by	Ratkiewicz et al. (2011), Burgess and Matamoros-Fernández (2016), and Hine et al. (2017) Kumar et al. (2018), Datta and Adar (2019), and Mariconti et al. (2019)

those identified in Table 7.2. These in particular are all potentially observable in short periods of online activity, e.g., a political debate (Rizoiu et al., 2018). Other coordinated behaviour observed in the literature require some ability to identify accounts of interest and track them over extended periods of time. *Metadata shuffling* involves groups of accounts hiding through changing and even swapping their names and other metadata (Mariconti et al., 2017; Ferrara, 2017). Related to this is *narrative switching*, in which an account suddenly deletes all their posts and then, potentially after a significant period of time, starts posting about different themes and issues (perhaps also having changed their account's appearance, Dawson and Innes, 2019). Dawson and Innes (2019) also observed changes in accounts' follower counts to identify the purchase of *fake followers* and *follower fishing* (used to boost reputation metrics), both of which require records of potentially lengthy periods of activity. Dawson and Innes (2019) also use *synchronicity* to identify groups temporally correlated through activity, but neglect to describe their specific method.

Different behaviour primitives, such as those originally introduced in Table 2.1, can be used to execute the amplification strategies mentioned. Many of these behaviour primitives have analogies on multiple OSNs, so techniques devised to detect them on one could be employed effectively on others. Dissemination can be carried out by reposting, using hashtags, or mentioning highly connected individuals in the hope they spread a message further. Accounts doing this covertly will avoid direct connections, and thus inference is required for identification. Giglietto et al. (2020b) propose detecting anomalous levels of coincidental URL use as a way to do this; we expand this approach to other interactions.

Some strategies require more sophisticated detection: detecting bullying through *dog-piling* (e.g., as happened during the `#GamerGate` incident, studied by Burgess and Matamoros-Fernández (2016), or to those posing questions to public figures at political campaign rallies[1]) requires collection of (mostly) entire conversation trees, which, while trivial to obtain on forum-based sites (e.g., Facebook and Reddit), are difficult on stream-of-post sites (e.g., Twitter,[2] Parler and Gab). As mentioned, detecting metadata shuffling requires long term collection on broad issues to detect the same accounts being active in different contexts, and other follower and narrative analyses can also require extended collection periods.

Figure 7.1 shows representations of the strategies highlighted above, offering clues about how they might be identified. To detect *Pollution*, we match the authors of posts mentioning the same (hash)tag. This way we can reveal not just those who are using the same hashtags with significantly greater frequency than the average but also those who use more hashtags than is typical. To detect a variant of *Boost*, we match authors reposting the same original post, and can explore which sets of users not only repost more often than the average, but those who repost content from a relatively small pool of accounts. Alternatively, we can match authors who post identical, or near identical text, as seen in our motivating example (Chapter 1); Graham et al. (2020c) have recently developed open sourced methods for this kind of matching, which have previously been used for campaign analysis (Lee et al., 2013). Considering reposts like retweets, however, it is unclear whether platforms deprioritise them when responding to stream filtering and search requests, so special consideration may be required when designing data collection plans. Finally, to detect *Bully*, we match authors whose replies are transitively rooted in the same original post, thus they are in the *same conversation*. This requires collection strategies that result in complete conversation trees, and also stipulates a somewhat strict definition of 'conversation'. On forum-based OSNs, the edges of a 'conversation' may be relatively clear: by commenting on a post, one is 'joining' the 'conversation'. Delineating smaller sets of interactions within all the comments on a post to find smaller conversations may be achieved by regarding each top-level comment and its replies as a conversation, but this may not be sufficient. Similarly, on stream-based OSNs, a conversation may be engaged in by a set of users if they all mention each other in their posts, as it is not possible to *reply* to more than one post at a time.

7.2.1 Problem statement

A clarification of our challenge at this point is:

> *To identify groups of accounts whose behaviour, though typical in nature, is anomalous in degree.*

[1] https://www.bbc.co.uk/bbcthree/article/72686b6d-abd2-471b-ae1d-8426522b1a97. Posted 2020-07-13. Accessed 2022-01-11.

[2] Changes introduced with Twitter's API version 2.0 aim to make this easier: https://developer.twitter.com/en/docs/twitter-api/conversation-id. Accessed 2022-01-11.

There are two elements to this. The first is *discovery*. How can we identify not just behaviour that appears more than coincidental, but also the accounts responsible for it? That is the topic of the next section. The second element is *validation*. Once we identify a group of accounts via our method, what guarantee do we have that the group is a real, coordinating set of users? This is especially difficult given inauthentic behaviour is hard for humans to judge by eye (Cresci et al., 2017b; Benkler et al., 2018; Jamieson, 2020).

7.3 Methodology

The major OSNs share a number of features, primarily in how they permit users to interact with each other, digital media and the platforms (e.g., Table 2.1); hashtags, URLs, and mentions work much the same way across many OSNs. By focusing on these commonalities, we can develop approaches that generalise across OSNs.

Traditional social network analysis relies on long-standing relationships between actors (Wasserman and Faust, 1994; Borgatti et al., 2009). On OSNs this requirement is typically fulfilled by friend/follower relations. These are expensive to collect and quickly degrade in meaning if not followed with frequent activity. By focusing on active interactions, however, it is possible to understand not just who is interacting with whom, but to what degree. This provides a basis for constructing (or inferring) social networks, acknowledging they may be transitory.

LCNs are built from inferred links between accounts. Supporting criteria relying on interactions alone, as observed in the literature (Ratkiewicz et al., 2011; Keller et al., 2019), include retweeting the same tweet (*co-retweet*), using the same hashtags (*co-hashtag*) or URLs (*co-URL*), or mentioning the same accounts (*co-mention*). To these we add joining the same 'conversation' (a tree of *reply* chains with a common root tweet) (*co-conv*). As mentioned earlier, other ways to link accounts rely on similar or identical content, metadata and temporal patterns (see Section 7.2). The criteria underpinning LCN links may be a combination of these and other interaction types.

7.3.1 The LCN / HCC pipeline

The key steps to extract HCCs from raw social media data are shown in Figure 7.2 and documented in Algorithm 1. The example in Figure 7.2 is explained after the algorithm has been explained, in Section 7.3.1.2.

Step 1. Convert social media posts P to common interaction primitives, I_{all}. This step removes extraneous data and provides an opportunity for the fusion of sources by standardising all interactions (thus including only the elements required for the coordination being sought).

Step 2. From I_{all}, filter the interactions, I_C, relevant to the set $C = \{c_1, c_2, ..., c_q\}$ of criteria (e.g., co-mentions and co-hashtags).

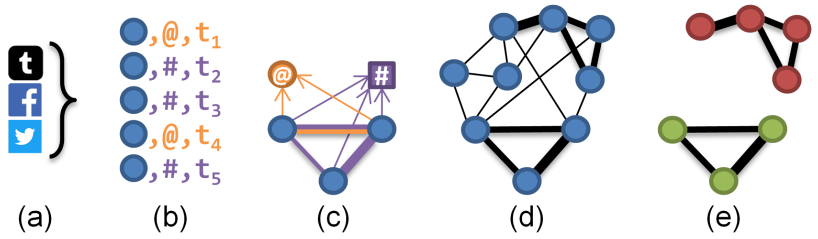


FIGURE 7.2. Conceptual LCN construction and HCC discovery process.

Algorithm 1 FindHCCs

Input: P: Social media posts, C: Coordination criteria, θ: Extraction parameter
Output: H: A list of HCCs

```
1: $I_{all} \leftarrow \text{ParseInteractionsFrom}(P)$
2: $I_C \leftarrow \text{FilterInteractions}(I_{all}, C)$
3: $M \leftarrow \text{FindCoordination}(I_C, C)$
4: $L \leftarrow \text{ConstructLCN}(M)$
5: $H \leftarrow \text{ExtractHCCs}(L, \theta)$
```

Step 3. Infer links between accounts given C, ensuring links are typed by criterion. The result, M, is a collection of inferred pairings. The count of inferred links between accounts v_i and v_j due to criterion $c \in C$ is $\beta^c_{\{v_i,v_j\}}$.

Step 4. Construct an LCN, L, from the pairings in M. This network $L = (V, E)$ is a set of vertices V representing accounts connected by undirected weighted edges E of inferred links. These edges represent evidence of different criteria linking the adjacent vertices. The weight of each edge $e \in E$ between vertices representing accounts v_i and v_j for each criterion c is $\omega^c(e)$, and is equal to $\beta^c_{\{v_i,v_j\}}$.

Most community detection algorithms will require the multi-edges be collapsed to single edges. The edge weights are incomparable (e.g., retweeting the same tweet is not equivalent to using the same hashtag), however, for practical purposes, the inferred links can be collapsed and the weights combined for cluster detection using a simple summation, e.g., Equation (7.1), or a more complex process like varied criteria weighting.

$$\omega(e) = \sum_{c=1}^{q} \omega^c(e) \tag{7.1}$$

Some criteria may result in highly connected LCNs, even if its members never interact directly. Not all types of coordination will be meaningful – people will co-use the same hashtag repeatedly if that hashtag defines the topic of the discussion (e.g., `#auspol` for Australian politics), in which case it is those accounts who co-use it significantly more often than others which are of interest. If required, the final step filters out these coincidental connections.

Step 5. Identify the highest coordinating communities H in L (Figure 7.2e) using a suitable community detection algorithm, such as Blondel et al. (2008)'s Louvain algorithm (used by Morstatter et al., 2018; Nasim et al., 2018; Vosoughi et al., 2018; Nizzoli et al., 2021), *k nearest neighbour* (kNN, used by Cao et al., 2015), markov clustering (used by Fazil and Abulaish, 2020), edge weight thresholding (used by Lee et al., 2013; Graham et al., 2020a; Pacheco et al., 2021), or FSA (Şen et al., 2016), an algorithm from the Social Network Analysis community that focuses on extracting sets of highly influential nodes from a network. Depending on the size of the dataset under consideration, algorithms suitable for very large networks may need to be considered (Fang et al., 2019). Some algorithms may not require the LCN's multi-edges to be merged (e.g., Bacco et al., 2017). We present a variant of FSA (Şen et al., 2016), FSA_V (Algorithm 2), designed to take advantage of FSA's benefits while addressing some of its costs. FSA does not just divide a network into communities (so that every node belongs to a community), but extracts only subsets of adjacent nodes that form influential communities within the overall network. FSA_V reduces the computational complexity introduced by FSA, which recursively applies Louvain to divide the network into smaller components and then, under certain circumstances, stitches them back together. The reason for this is to make FSA_V more suitable for application to a streaming scenario, in which execution speed is a priority.

Similar to FSA, FSA_V initially divides L into communities using the Louvain algorithm but then builds candidate HCCs within each, starting with the 'heaviest' (i.e., highest weight) edge (representing the most evidence of coordination). It then attaches the next heaviest edge until the candidate's mean edge weight (MEW) is no less than θ ($0 < \theta \leq 1$) of the previous candidate's MEW, or is less than L's overall MEW. In testing, edge weights appeared to follow a power law, so θ was introduced to identify the point at which the edge weight drops significantly; θ requires tuning. A final filter ensures no HCC with a MEW less than L's is returned. Unlike in FSA, recursion is not used, nor stitching of candidates, resulting in a simpler algorithm.

This algorithm prioritises edge weights while maintaining an awareness of the network topology by examining adjacent edges, something ignored by simple edge weight filtering. Our goal is to find sets of strongly coordinating users, so we prioritise strongly tied communities while still acknowledging coordination can also be achieved with weak ties (e.g., 100 accounts paid to retweet one tweet).

The complexity of the entire pipeline is low order polynomial due primarily to the pairwise comparison of accounts to infer links in Step 3, which can be constrained by window size when addressing the temporal aspect. For large networks (i.e., networks with many accounts), that may be too costly to be of practical use; the solution for this relies on the application domain inasmuch as it either requires a tighter temporal constraint (i.e., a smaller time window) or tighter stream filter criteria, causing a reduction in the number of accounts, potentially along with a reduction in posts. Algorithmic complexity is discussed in Section 7.3.3.

Algorithm 2 ExtractHCCs (FSA_V)

```
Input: $L$=$(V, E)$: An LCN, $\theta$: HCC threshold
Output: $H$: Highly coordinating communities
$E' \leftarrow$ MergeMultiEdges($E$)
$g\_mean \leftarrow$ MeanWeight($E'$)
$louvain\_communities \leftarrow$ ApplyLouvain($L$)
Create new list, $H$
for $l \in louvain\_communities$ do
    Create new community candidate, $h = (V_h, E_h)$
    Add heaviest edge $e \in l$ to $h$
    $growing \leftarrow$ true
    while $growing$ do
        Find heaviest edge $\vec{e} \in l$ connected to $h$ not in $h$
        $old\_mean \leftarrow$ MeanWeight($E_h$)
        $new\_mean \leftarrow$ MeanWeight(Concatenate($E_h, \vec{e}$))
        if $new\_mean < g\_mean$ or
        $new\_mean < (old\_mean \times \theta)$ then
            $growing \leftarrow$ false
        else
            Add $\vec{e}$ to $h$
    if MeanWeight($E_h$) $> g\_mean$ then
        Add $h$ to $H$
```

7.3.1.1 Addressing the temporal aspect

Temporal information is a key element of coordination, and thus is critical for effective coordination detection. Frequent posts within a short period may represent genuine discussion or deliberate attempts to game trend algorithms (Grimme et al., 2018; Varol et al., 2017b; Assenmacher et al., 2020). We treat the post stream as a series of discrete windows to constrain detection periods. An LCN is constructed from each window (Step 4), and these are aggregated and mined for HCCs (Step 5). We assume posts arrive in order, and assign them to windows by timestamp.

7.3.1.2 A brief example

Figure 7.2 gives an example of searching for co-hashtag and co-mention coordination across Facebook, Twitter, and Tumblr posts. The posts are converted to their interaction primitives in Step 1, shown in Figure 7.2a. The information required from each post is the identity of the post's author,[3] the timestamp of the post for addressing the temporal aspect, and the hashtag or account mentioned (there may be many, resulting in separate records for each). This is done in Figure 7.2b, which shows the filtered mentions (in orange) and hashtag uses (in purple), ordered according to timestamp.

Step 3 in Figure 7.2c involves searching for evidence of coordination through searching for our target coordination strategies through pairwise examination of accounts and

[3] Linking identities across social media platforms is beyond the scope of this work, but the interested reader is referred to Adjali et al. (2020) for a recent contribution to the subject.

their interactions. Here, three accounts co-use a hashtag while only two of them co-mention another account.

By Step 4 in Figure 7.2d, the entire LCN has been constructed, and then Figure 7.2e shows its most highly coordinating communities.

As mentioned above, to account for the temporal aspect, the LCNs produced for each time window in Figure 7.2d can be aggregated and then mined for HCCs, or HCCs could be extracted from each window's LCN and then they can be aggregated, or analysed in near real-time, as dictated by the application domain.

7.3.1.3 Opportunities for fusion

As mentioned above, many of the interaction we consider have analogies on multiple OSNs, so a technique applied to Twitter, for example, may also be effective on Reddit or Tumblr. Misinformation was widely disseminated over Facebook, Tiktok, Twitter, and WhatsApp during the 2021 Israeli/Palestinian conflict as links to misattributed videos, images of blocks of text, and audio files.[4] Our technique could be used to study coordinated link (i.e., URL) sharing across these platforms in an appropriate time period, similar to the work of Giglietto et al. (2020b) and Broniatowski (2021) — all that is required from each platform's posts are the identity of the posting account, the link posted[5] and the post's timestamp. The identities of accounts posting the URLs will differ between platforms, of course, but this technique may also provide a mechanism for cross-platform identity matching, associating accounts that frequently post the same or similar content. Nimmo et al. (2020) essentially performed this task manually by searching for the same article content across different platforms, and then confirming similarity between the account names found. Our technique could be incorporated into the researcher's workflow to make this task easier by searching for duplication of text, and automatically linking instances where it is found, and then highlighting those connections.

7.3.2 Validation methods

As mentioned in Section 7.2.1, the second element of addressing our research challenge is that of validation. Once HCCs have been discovered, it is necessary to confirm that what has been found are examples of genuine coordinating groups. This step is required before addressing the further question of whether the coordination is authentic (e.g., grassroots activism) or inauthentic (e.g., astroturfing).

A number of methods we rely upon for validation have been introduced in Chapter 3, but we discuss some specific approaches in this subsection, including the specifics of our comparison datasets, network visualisations, how we examine the consistency of

[4] https://www.nytimes.com/2021/05/14/technology/israel-palestine-misinformation-lies-social-media.html. Posted 2021-05-14. Accessed 2022-01-19.

[5] More sophisticated content matching can also be used in Step 3, comparing what media the links refer to, rather than just the link itself (*cf.*, Yu, 2021; Pacheco et al., 2021).

HCC coordination, and details of the confirmation provided by ML systems is used to gain confidence that the discovered HCCs are coordinating similar to those in our ground truth.

7.3.2.1 Comparison datasets

In addition to relevant datasets, we make use of a ground truth (GT), in which we expect to find coordination (*cf.*, Keller et al., 2017; Vargas et al., 2020; Fazil and Abulaish, 2020). We introduced this notion in Section 3.3.1. By comparing the evidence of coordination (i.e., HCCs) we find within the ground truth with the coordination we find in the other datasets, we can develop confidence that: a) our method finds coordination where we expect to find it (in the ground truth); and b) our method also finds coordination of a similar type where it was not certain to exist. Furthermore, to represent the broader population (which is not expected to exhibit coordination), similar to Cao et al. (2015), we create a randomised HCC network from the non-HCC accounts in a given dataset, and then compare its HCCs with the HCCs that had been discovered by our method.

7.3.2.2 Network visualisation

We use a variety of parameters to construct visualisations of networks for subjective analysis using the tools introduced in Section 3.2.1.7. In particular, the force-directed layouts aid in clarifying clusters identified with the Louvain method (Blondel et al., 2008). Each connected component is an HCC, node colour is used to represent the number of posts, and edge weight is represented by thickness and, depending on the density of the network, depth of colour. For analyses that involve multiple criteria (e.g., co-conv and co-mention), we use node shape to represent which combination of criteria an HCC is bound by (e.g., just co-mention or a combination of co-mention and co-conv or just co-conv).

By extending the HCC account networks with nodes to represent the ‘reasons’ (described in Section 3.2.2.4, thereby creating a 2-layer *account/reason* network, we investigate how HCCs relate to one another. In this case, the account/reason network has two types of nodes and two types of edges (‘coordinates with’ links between accounts and ‘caused by’ or ‘associated because’ links between ‘reasons’ and accounts). Visualising the 2-layer network by colouring nodes by their HCC and using a force-directed layout highlights how closely HCCs associate with each other, not only revealing what reasons draw HCCs together (i.e., HCCs may be bound by a single reason, or an HCC may be entirely isolated from others in the broader community), but also how many reasons bind them (i.e., many reasons may bind an HCC together or just one). Deeper insights can be revealed from this point using multi-layer network analyses.

7.3.2.3 Variation of content

Part of characterising HCC behaviour is examining the distributions of the content they share, and how the variation observed in detected HCCs differ from that of RANDOM groupings. Highly coordinated behaviour such as co-retweeting involves reusing the same content frequently, resulting in low feature variation (e.g., hashtags, URLs, mentioned accounts), which can be measured via entropy, discussed in Section 3.3.2. A frequency distribution of each HCC's use of each feature type is used to calculate each entropy score. Low levels of feature variation corresponds to low entropy values. As per Cao et al. (2015), we compare the entropy of features used by detected HCCs to RANDOM ones and visualise their cumulative frequency. Entries for HCCs which did not use a particular feature are omitted, as their scores would inflate the number of groups with 0 entropy.

7.3.2.4 Consistency of coordination

The method presented Section 7.3.1 highlights HCCs that coordinate their activity at a high level over an entire collection period. Further steps can be taken to determine which HCCs are coordinating their behaviour repeatedly and consistently across adjacent time windows. In this case, for each time window, we consider not just the nodes and edges from the current LCN, but additionally from previous windows, applying a degradation factor the contribution of their edge weights.

To build an LCN from a sliding frame of T time windows, the new LCN includes the union of the nodes and edges of the individual LCNs from the current and previous windows, but to calculate the edge weights, we apply a decay factor, α, to the weights of edges appearing in windows before the current one. In this way, we apply a multiplier of α^x to the edge weights, where x is the number of windows into the past: the current window is 0 windows into the past, so its edges are multiplied by $\alpha^0 = 1$; the immediate previous window is 1 window back, so its edge multiplier is α^1; the one before that uses α^2, and so on until the farthest window back uses α^{T-1}. Generalising from Step 4, the weight $\omega(e, c, t)$ for an edge $e \in E$ between accounts $v_i, v_j \in V$ for criterion c at window t and a sliding window T windows wide is given by

$$\omega(e, c, t) = \sum_{x=0}^{T-1} \omega(e, c, t - x) \cdot \alpha^x. \tag{7.2}$$

In this way, to create a baseline in which the sliding frame is only one window wide, one only need choose T=1, regardless of the value of α. As $\alpha \to 1$, the contributions of previous windows are given more consideration.

7.3.2.5 Supervised machine learning with one-class classifiers

An approach that aids in the management of data with many features is classification through machine learning (introduced in Section 3.4). This is an approach that has been used extensively in campaign detection, in which tweets are classified, rather than accounts (e.g., Lee et al., 2013; Chu et al., 2012; Wu et al., 2018). Because of its 'black box' nature, its application should be considered carefully, however. Our intent is to use classification to validate that entire HCCs (not just individual tweets or accounts) detected in datasets are similar to those found in ground truth. Such classifiers will not be applicable in the general case, as they rely on ground truth (which is historical by nature) for training data. Tactics and strategies used in information operations will change over time, as shown by Alizadeh et al. (2020); this is not just to avoid detection but also because OSN features change over time. As our focus is only on a positive answer to whether one HCC is similar to others, it is acceptable to rely on one-class classification (i.e., an HCC detected in a dataset is either recognised as COORDINATING/positive or is regarded as NON-COORDINATING/unknown). The more common binary classification approach was used by Vargas et al. (2020), however our approach has two distinguishing features:

1. We rely on one-class classification because we have positive examples of what we regard as COORDINATING from the ground truth, and everything else is regarded as unknown, rather than definitely 'not coordinating'. When our one-class classifier recognises HCC accounts as positive instances, it provides confidence that the HCC members are coordinating their behaviour in the same manner as the accounts in the ground truth. We can therefore prioritise precision over recall (as discussed in Section 3.4.2).

2. We rely on features from both the HCCs and the HCC members and use the HCC members as the instances for classification, given it is unclear how many members an HCC may have, and accounts that are members of HCCs may have traits in common that are distinct from 'normal' accounts. In contrast, Vargas et al. (2020) relied on features of "coordination networks" (i.e., HCCs) alone, as they were their classification instances. For this reason the feature vectors that our classifier is trained and tested on will comprise features drawn from the individual accounts and their behaviour as well as the behaviour of the HCC of which they are a member. Feature vectors for members of the same HCC will naturally share the feature values drawn from their grouping.

Regarding the construction of the feature vector, at a group level, we consider not just features from the HCC itself, which is a weighted undirected network of accounts, but of the activity network built from the interactions of the HCC members within the corpus. The activity network is a multi-network (i.e., supports multiple edges between nodes) with nodes and edges of different types. The three node types are *accounts*, *URLs*, and *hashtags*. Edges represent interactions and the following types are modelled: hashtag uses, URL uses, mentions, repost/retweets, quotes (*cf.*, comments on a Facebook share or Tumblr repost), reply, and 'in conversation' (meaning that one account replied to a post that was transitively connected via replies to an original post by an account in the corpus). This activity network therefore represents not just the members of the HCC but also their degree of activity in context.

Classifier algorithms. We use the GT to train three classifiers. A bagging PU classifier (BPU, Mordelet and Vert, 2014) was used, the implementation[6] for which was based on a Random Forest (RF) classifier configured with 1,000 trees (estimators). We also used a standard 1,000 tree RF, as used by Vargas et al. (2020), to compare directly with BPU. A Support Vector Machine (SVM) classifier was also used, given the technique's known high performance with non-linear recognition problems even with small feature sets due its use of the kernel trick. Furthermore, Mordelet and Vert (2014) employed a variety of SVMs as part of their experimentation, though our choice of implementation differed. Both SVM and RF implementations were drawn from the `scikit-learn` Python library (Pedregosa et al., 2011). Contrasting "unlabelled" training instances were created from the RANDOM dataset. Feature vector values were standardised prior to classification and upsampling was applied to create balanced training sets of approximately 400 positive and random elements each. 10-fold cross validation was used.

The classifiers predict whether instances provided to them are in the positive or unlabelled classes, which, to aid readability, we refer to as 'COORDINATING' and 'NON-COORDINATING', respectively.

The performance metrics relied upon have been introduced in Section 3.4.2, including precision, recall, the corresponding F_1 score, and accuracy.

7.3.3 Complexity analysis

The steps in processing timeline presented in Section 7.3.1 are reliant on two primary factors: the size of the corpus of posts, P, being processed, and the size of the set of accounts, A, that posted them. Therefore $|A| \leq |P|$ and the complexity of Step 1 is linear, $O(|P|)$, because it requires processing each post, one-by-one. The set of interactions, I_{all}, it produces may be larger than $|P|$, because a post may include many hashtags, mentions, or URLs, but given posts are not infinitely long (even long Facebook and Tumblr posts can only include several thousand words), the number of interactions will also be linear, i.e., $|I| = k|P|$, for some constant k. Step 2 filters these interactions down to only those of interest, I_C, based on the type of coordinated activity sought, C, so $|I_{all}| \geq |I_C|$, and again the complexity of this step is also linear, $O(|I_{all}|)$, as it requires each interaction to be considered. Step 3 seeks to find evidence of coordination between the accounts in the dataset, and so requires examining each filtered interaction and building up data structures to associate each account with their interactions ($O(|I_{all}|)$), then emitting pairs of accounts matching the coordination criteria, producing the set M, which requires the pairwise processing of all accounts, and so is $|A|^2$ steps with a subsequent complexity of $O(|A|^2)$. This,

[6] Thanks to Roy Wright for his implementation: https://github.com/roywright/pu_learning/blob/master/baggingPU.py Version as of 2019-08-15. Last accessed 2022-01-11.

however, also depends on the pairwise comparison of each account's interactions, which is likely to be small, practically, but theoretically could be as large as $|I_C|$ if one user is responsible for every single interaction in the corpus (but then $|A|$ would be 1). On balance, as a result, we will regard the processing of each pair of users' interactions as linear with a constant factor k (i.e., $O(k|A|^2) = O(|A|^2)$). In Step 4, producing the LCN, L, from the criteria is a matter of considering each match one-by-one, so is again linear (though potentially large, depending on $|M|$). The final step (5) is to extract the HCCs from the LCN, and its performance and complexity very much depend upon the algorithm employed, but significant research has been applied in this field (as considered in, e.g., Bedru et al., 2020). For FSA_V, which relies on the Louvain algorithm with complexity $O(|A| \log_2 |A|)$ (Blondel et al., 2008), it considers edges within each community to build its HCC candidates, so has a complexity of less than $O(|E|)$, where $|E|$ is the number of edges in the LCN, meaning its complexity is linear. FSA_V's complexity is therefore $O(|A| \log_2 |A| + |E|)$.

We regard the computation complexity of the entire pipeline as the highest complexity of its steps, which are:

1. Extract interactions from posts: $O(|P|)$
2. Filter interactions: $O(|I_{all}|)$
3. Find evidence of coordination: $O(|A|^2)$
4. Build LCN from the evidence: $O(|M|)$
5. Extract HCCs from LCN using, e.g., FSA_V: $O(|A| \log_2 |A| + |E|)$

The maximum of these is Step 3, the search for evidence of coordination, $O(|A|^2)$. Though in theoretical terms the method is potentially very costly, in practical terms we are bound by the number of accounts in the collection (which is determined by the manner in which the data was collected and the nature of the online discussion to which it pertains) and may be managed by constraining the time window, further reducing the number of posts (and therefore accounts) considered, as long as that suits the type of coordination being sought.

7.4 Evaluation

Our approach was evaluated in two phases:

- The first was conducted as an experiment using the validation methods mentioned above and two datasets known to include coordinated behaviour, as well as a ground truth dataset.
- The second phase involved two case studies in which we apply our approach against datasets relating to politically contentious topics expected to include polarised groups.

The first stage of the evaluation involved searching for *Boost* by co-retweet and other strategies while varying window sizes (γ). FSA_V was compared against two other community detection algorithms, when applied to the LCNs built in Step 4 (aggregated). We then validated the resulting HCCs through a variety of network, content, and temporal analyses and machine learning classification, guided by the research questions posed in Section 7.1. Discussion of further applications and performance metrics is also presented.

7.4.1 The experiment datasets

The two real-world datasets selected (shown in Table 7.3) represent two collection techniques: filtering a live stream of posts using keywords direct from the OSN (DS1) and collecting the posts of specific accounts (DS2):

DS1 Tweets relating to a regional Australian election in March 2018, including a ground truth subset (GT); and

DS2 A large subset of the Russian IRA (Chen, 2015; Mueller, 2018) dataset published by Twitter in October, 2018.[7]

TABLE 7.3. Experiment dataset statistics. (Rates are per account per day.)

	Tweets	Retweets (%)		Accounts	Tweet rate	Retweet rate
DS1	115,913	63,164	(54.5%)	20,563	0.31	0.17
(GT)	4,193	2,505	(59.7%)	134	1.74	1.04
DS2	1,571,245	729,937	(56.5%)	1,381	3.12	1.45

DS1 was collected using RAPID (Lim et al., 2019) over an 18 day period (the election was on day 15) in March, 2018. The filter terms included nine hashtags and 134 political handles (candidate and party accounts). The dataset was expanded by retrieving all replied to, quoted and political account tweets posted during the collection period. The political account tweets formed our ground truth. It was our expectation that some of the coordinated political influence techniques observed on the international stage may have been adopted by political parties and issue-motivated groups at the regional level by 2018 (especially given the use of political bots had been reported in the Australian setting five years prior, as reported in Woolley, 2016), and hence would be present in this dataset.

The RU-IRA dataset released by Twitter covers 2009 to 2018, but DS2 is the subset of tweets posted in 2016, the year of the US presidential election. Because DS2 consists entirely of RU-IRA accounts which Twitter believed to be connected with an SIO, it was expected to include evidence of coordinated amplification. It was also much larger than DS1, and our intent was that our findings would complement forensic

[7] https://transparency.twitter.com/en/reports/information-operations.html. Accessed 2022-01-19.

studies of the activity (e.g., Benkler et al., 2018; Jamieson, 2020) and also contrast with techniques from more focused studies (e.g., Dawson and Innes, 2019).

7.4.2 Experimental set up

The size of the window γ was set at $\{15, 60, 360, 1440\}$ (in minutes) and the three community detection methods used on the aggregated LCNs were:

- FSA_V (θ=0.3);
- kNN with k=$ln(|V|)$ (*cf.*, Cao et al., 2015); and
- a simple threshold retaining the edges with a normalised value above 0.1.

7.4.2.1 Parameter selection

Other than a value of k=$ln(|V|)$ for kNN (taken from Cao et al., 2015), the choice of values for parameters γ, θ and the threshold were determined as follows. Our intent was to search for human-driven coordination, i.e., teams of humans manipulating potentially several accounts each, meaning that the timeframes under examination would need to allow for the time required to switch between accounts. As discussed by Dawson and Innes (2019), the motivation for even paid coordinated behaviour may be based on numbers of posts made, rather than how tightly coordinated they are, so by examining a relatively wide 'short' window of 15 minutes allows for such people to react to each others' posts as they see them (rather than the sub-minute coordination sought by others, e.g., Giglietto et al., 2020a; Pacheco et al., 2021; Dawson and Innes, 2019). The 60-minute window allows for people motivated by personal interest as well as paid trolls, who check their social media frequently throughout the day while attending to other duties (e.g., preparing new content, Nimmo et al., 2020). The six hour time frame is of medium length and allows for users who check social media over breakfast, at lunch, and then at dinner who also may be more motivated by personal reasons to coordinate their behaviour. Finally, the long term time frame of a whole day allows for accounts that only check social media in concentrated sessions once a day, but who coordinate their actions with others each day outside of the six-hour window. Furthermore, automated coordinated accounts (i.e., bots) can react to posts very quickly (i.e., within seconds), and simple implementations can be revealed by their consistent short response times rather than relying on the more sophisticated co-activity methods presented here. More complex bot implementations vary their response times to avoid this (Cresci et al., 2017b; Cresci, 2020), however if they wish to game OSN trending algorithms to improve their reach, their posts must occur near to each other in time. Values for γ were also informed by the observation of Zhao et al. (2015) that 75% of retweets occur within six hours of posting. This implies that if attempts were made to boost a tweet, retweeting it in much shorter times would be required for it to stand out from typical traffic. Varol et al. (2017b) checked Twitter's trending hashtags every 10 minutes, which is an indication of how quickly

TABLE 7.4. HCCs by coordination strategy.

			GT			DS1			DS2		
	Strategy	γ	Nodes	Edges	Comp.	Nodes	Edges	Comp.	Nodes	Edges	Comp.
LCN	Boost	15	44	112	5	8,855	80,702	419	855	23,022	14
	Pollute	15	51	154	2	13,831	1,281,134	73	1,203	65,949	5
	Bully	60	70	482	1	16,519	1,925,487	222	1,103	37,368	5
FSA_V	Boost	15	9	6	3	633	753	167	113	758	19
	Pollute	15	9	5	4	135	93	50	24	15	9
	Bully	60	11	7	4	338	280	119	109	1,123	16
kNN	Boost	15	9	21	1	1,041	33,621	1	675	22,494	1
	Pollute	15	11	37	1	724	153,424	1	1,040	65,280	1
	Bully	60	18	135	1	1,713	663,413	1	692	35,136	1
Threshold	Boost	15	11	16	3	85	68	31	8	10	2
	Pollute	15	24	26	3	44	37	10	6	13	1
	Bully	60	15	19	3	25	23	8	10	10	3

a concerted *Boost*ing effort may have an effect. Values chosen for γ therefore ranged from 15 minutes to a day, growing by a factor of approximately four at each increment. Deliberate coordinated retweeting (i.e., covert *Boost*ing masquerading as grassroots activity) was expected to occur in the smaller windows, but then be replaced by coincidental co-retweeting as the window size increases.

Values for θ and the threshold were based on experimenting with values in $[0.1, 0.9]$, maximising the MEW to HCC size ratio, using the DS1 and DS2 aggregated LCNs when γ=$\{15, 1440\}$.

7.4.3 Experimental results

The research questions introduced in Section 7.1 guide our discussion, but we also present follow-up analyses.

7.4.3.1 HCC detection (RQ1)

Detecting different strategies. The three detection methods all found HCCs when searching for *Boost* (via co-retweets), *Pollute* (via co-hashtags), and *Bully* (via co-mentions), details of which are shown in Table 7.4. Notably, *kNN* consistently builds a single large HCC, highlighting the need to filter the network prior to applying it (*cf.*, Cao et al., 2015). The *kNN* HCC is also consistently nearly as large as the original LCN for DS2, perhaps due to the low number of accounts and the fact that *kNN* retains every edge adjacent to the retained vertices, regardless of weight. It is not clear, then, that *kNN* is producing meaningful results used in this way, even if it can extract a community.

Varying window size γ. Different strategies may be executed over different time periods, based on their aims. *Boost*ing a message to game trending algorithms requires the messages to appear close in time, whereas some forms of *Bully*ing exhibit

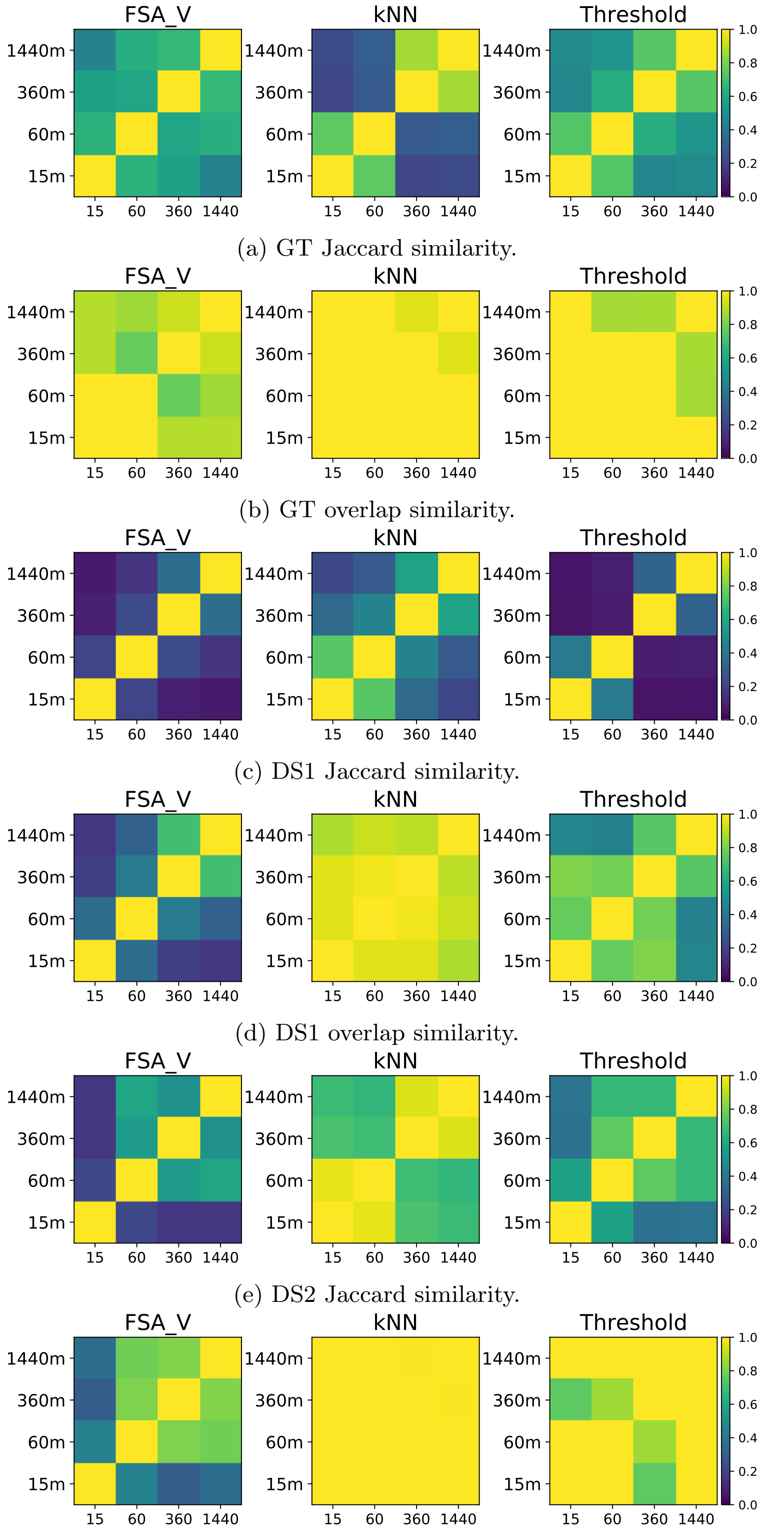
Table 7.5. HCCs by window size γ (Boost, FSA_V).

		Network Attributes			HCC Sizes				Nodes in common			
	γ	Nodes	Edges	HCCs	Min.	Max.	Mean	Std. Dev.	γ=15	γ=60	γ=360	γ=1440
GT	15	9	6	3	3	3	3.00	0.00	9	9	8	8
	60	14	9	5	2	3	2.80	0.40	-	14	10	12
	360	13	9	5	2	3	2.60	0.49	-	-	13	12
	1440	17	12	6	2	3	2.80	0.37	-	-	-	17
DS1	15	633	753	167	2	18	3.79	2.21	633	218	93	100
	60	619	1,293	151	2	13	4.10	2.30	-	619	208	193
	360	503	1,119	127	2	19	3.96	2.58	-	-	503	350
	1440	815	2,019	141	2	110	5.78	12.60	-	-	-	815
DS2	15	113	758	19	2	65	5.95	13.94	113	34	29	25
	60	77	394	18	2	27	4.28	5.64	-	77	62	54
	360	98	775	15	2	32	6.53	9.13	-	-	98	56
	1440	69	380	15	2	27	4.60	6.15	-	-	-	69

only consistency and low variation (mentioning the same account repeatedly). Polluting a user's timeline on Twitter can also be achieved by frequently joining their conversations over a sustained period.

Varying γ searching for *Boost*, we found different accounts were prominent over different time frames (Table 7.5); the overlap in the accounts detected in each time frame differed considerably even though the number of HCCs stayed relatively similar. Figure 7.3 shows the Jaccard and overlap similarity scores (Equations 3.13 and 3.14, respectively) between the sets of accounts appearing in each window size (agnostic of HCC membership). The overlap results for kNN shows very high levels of similarity, but lower levels of Jaccard similarity. For all datasets, as γ grows kNN finds more and more HCC members, including all the ones it found with smaller window sizes (overlap similarity values appear close to 1.0, shown as yellow). The highest Jaccard similarities for kNN seem to group the shorter periods (γ=$\{15, 60\}$) and the medium and long periods (γ=$\{360, 1440\}$). FSA_V finds different sets of members in each time window without significant overlap, though for DS2 it appears that the windows longer than 15 minutes have many members in common, but have very few in common with the γ=15 HCCs. As might be expected, thresholding by LCN edge weight results in the identification of additional accounts as γ increases, and the Jaccard similarity of GT and DS1 (Figure 7.3c) reveals that accounts identified in the shorter time windows (γ=$\{15, 60\}$) are very different to those from the longer time windows, but they still overlap somewhat (Figure 7.3d). This suggests that although there are some accounts that coordinate in short periods, other accounts coordinate *more* over the medium and long time periods. These include media accounts that are consistently highly active over longer periods and differ from the active discussion participants who might log on to Twitter in the evening for a few hours whose behaviour is more bursty in nature.

Other than in GT, which revealed very few HCCs, the sizes of the HCCs found seemed



(a) GT Jaccard similarity.

(b) GT overlap similarity.

(c) DS1 Jaccard similarity.

(d) DS1 overlap similarity.

(e) DS2 Jaccard similarity.

(f) DS2 overlap similarity.

FIGURE 7.3. Similarity matrices of HCC account sets found using different window sizes (FSA_V). The similarity measured here relates to the accounts found not to the similarity in groupings of accounts into HCCs. Yellow implies a high similarity (Jaccard: account sets are identical, Overlap: one set is a subset), while blue implies low similarity (i.e., account sets are disjoint).

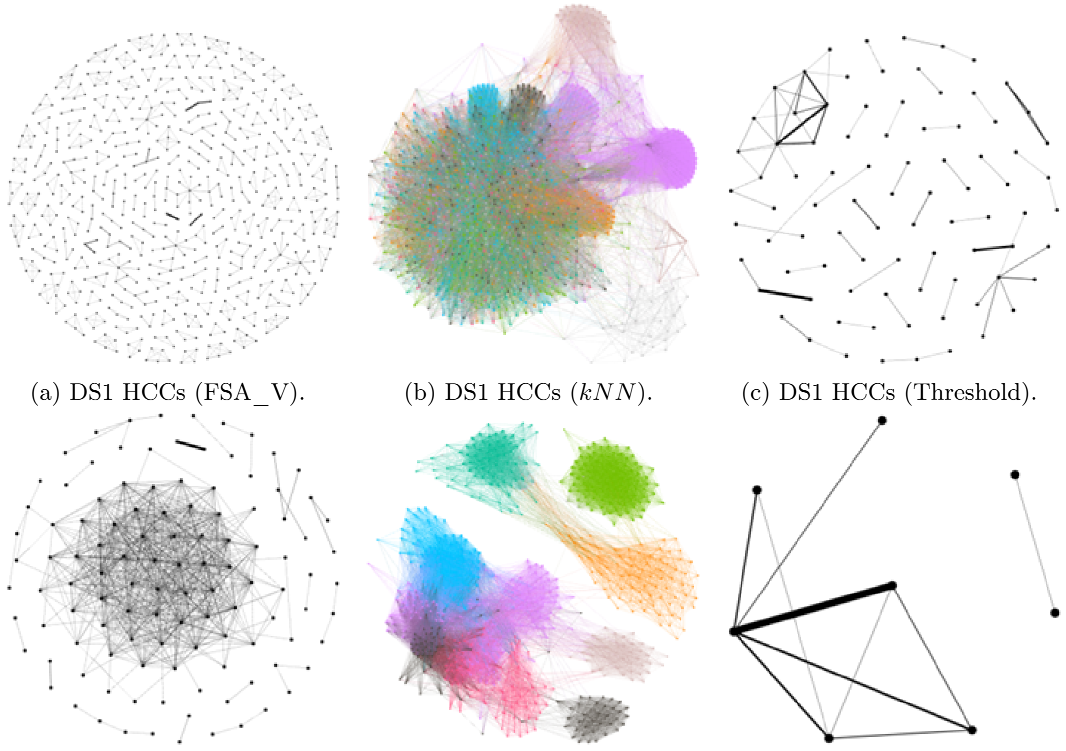
TABLE 7.6. HCCs by detection method (Boost, γ=15).

		Network Attributes			HCC Sizes		Nodes in common		
		Nodes	Edges	HCCs	Min.	Max.	FSA_V	*kNN*	Threshold
DS1	FSA_V	633	753	167	2	18	633	56	36
	kNN	1,041	33,621	1	1,041	1,041	-	1,041	44
	Threshold	85	68	31	2	14	-	-	85
DS2	FSA_V	113	758	19	2	65	113	88	4
	kNN	675	22,494	1	675	675	-	675	8
	Threshold	8	10	2	2	6	-	-	8

to follow a rough power law; most were very small but one or a few were very large (see the HCC Sizes section in Table 7.5). The number of HCCs did not vary significantly nor consistently as γ increased. The number of edges retrieved tells us in DS1, as the window increased, more edges had weights high enough to be retained, whereas DS2 edge counts diminished, implying that the LCNs were progressively dominated by a smaller number of very *heavy* edges, while other remained relatively *light*.

HCC detection methods. Similarly, HCCs discovered by the three community extraction methods (Table 7.6) exhibit large discrepancies, suggesting that whichever method is used, tuning is required to produce interpretable results. This is evident in the literature: Cao et al. conducted significant pre-processing when identifying URL sharing campaigns across two years of Twitter activity (Cao et al., 2015), and Pacheco et al. showed how specific strategies could identify groups in the online narrative surrounding the Syrian White Helmet organisation (Pacheco et al., 2020). Here we present the variation in results while controlling methods and other variables and keeping the coordination strategy constant, as our interest here is to validate the effectiveness of the method.

The networks were visualised using the FR layout in Figure 7.4, revealing further structure within the *kNN* networks, each of which consisted of a single connected component. To examine the structure of the single *kNN* component more closely, we applied Louvain analysis (Blondel et al., 2008) and coloured the largest detected clusters. The clustering reveals distinct communities within both the lone *kNN* HCC found in each of the datasets. It is possible the DS2 ones are more easily discernible either due to the smaller number of accounts (675 compared with 1,041) or because the accounts were, in fact, organised teams of malicious actors acting over a longer time frame. In either case, it makes clear that *kNN*, configured as it was, failed to distinguish communities clearly extractable via other means. This is less an indictment on *kNN* and more an indication that community extraction is likely to be a multi-step process embedded in particular domains and datasets, and in the particular types of networks to which they are applied. The networks in Figures 7.4b and 7.4e bear a passing resemblance to many in, e.g., the deep analysis of the media landscape during the 2016 US presidential election by Benkler et al. (2018) (which relied on simpler methods to build their networks), however these examples are networks of accounts



(a) DS1 HCCs (FSA_V). (b) DS1 HCCs (kNN). (c) DS1 HCCs (Threshold).

(d) DS2 HCCs (FSA_V). (e) DS2 HCCs (kNN). (f) DS2 HCCs (Threshold).

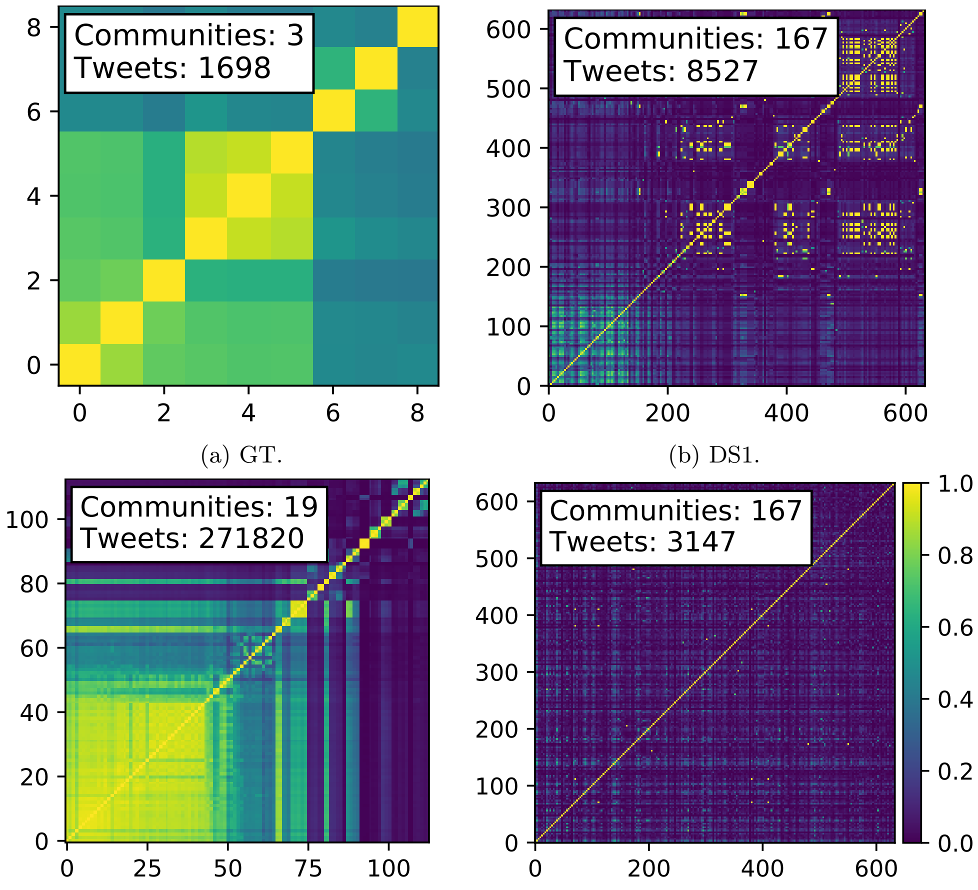
Figure 7.4. HCCs discovered using different methods in DS1 and DS2 (Boost, γ=15). Each kNN network consists of a single connected component, but detected clusters have been coloured to highlight internal structures.

rather than media organisations or sites, and, importantly, are not necessarily directly linked, offering the possibility of uncovering otherwise hidden connections between actors. This could be especially valuable when searching multiple OSNs.

7.4.3.2 HCC differentiation (RQ2)

How similar are the discovered HCCs to each other and to the rest of the corpus? The HCC detection methods used relied on network information; in contrast we examine content, metadata and temporal information to validate the results. We contrast DS1 and DS2 results with GT and a RANDOM dataset, constructed to match the HCC distributions in DS1 (FSA_V, γ=15). As DS2 consisted entirely of bad actors, and GT consisted entirely of political accounts, it was felt non-HCC accounts from DS1 would offer more 'normal' non-coordinating accounts.

Internal consistency. Visualising the similarities between accounts using the method in Section 3.3.5 (Figure 7.5), the HCCs are discernible as being internally similar. The RANDOM groupings demonstrated little to no similarity, internal or external, as expected, while the DS2 HCCs demonstrated high internal similarity, as expected of organised accounts over an extended period. The internal consistency of the DS1 HCCs is not as clear as for DS2, possibly due to the greater number of HCCs. Where HCCs are highly similar to others (indicated by yellow cells off the diagonal),



(a) GT.

(b) DS1.

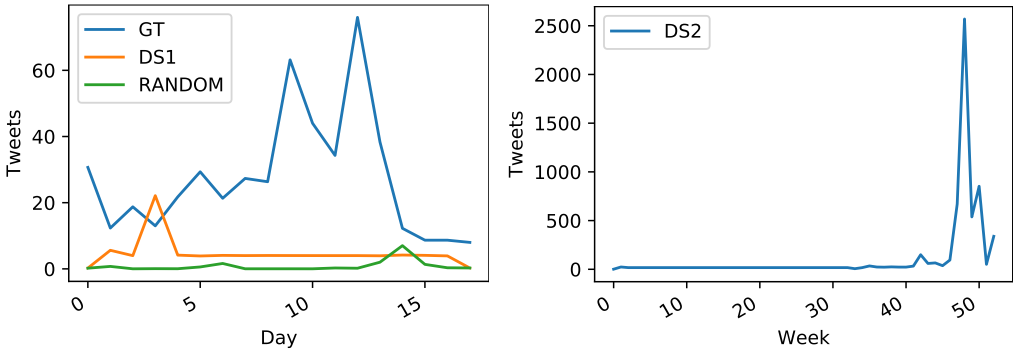
(c) DS2.

(d) RANDOM.

FIGURE 7.5. Similarity matrices of content posted by HCC accounts (FSA_V, γ=15). Each axis has an entry for each account, grouped by HCC. Each cell represents the similarity between the two corresponding accounts' content, calculated using cosine similarity (yellow = high similarity). Each account's content is modelled as a vector of 5-character n-grams of their combined tweets.

it is highly likely these are due to small HCCs (e.g., with two or three members) retweeting the same small set of tweets (fewer than ten). The use of filtering in conjunction with FSA_V may help remove potentially spurious HCCs, as could a final merge phase, joining HCC candidates whose evidence for coordination matches closely (e.g., two small HCCs retweeting 90% of the same tweets, kept separate by FSA_V but clearly similar).

Temporal patterns. We applied the temporal averaging technique described in Section 3.3.7 to compare the daily activities of the HCCs found in GT, DS1 and RANDOM (all of which occur over the same time period) in Figure 7.6a and weekly activities in DS2 in Figure 7.6b. The GT accounts were clearly most active at two points prior to the election (around day 15), during the last leaders' debate and just prior to the mandatory electoral advertising blackout. DS1 and RANDOM HCCs were only consistently active at different times: around the day 3 leaders' debate and on election day, respectively. Inter-HCC variation may have dragged the mean activity value down, as many small HCCs were inactive each day. Reintroducing



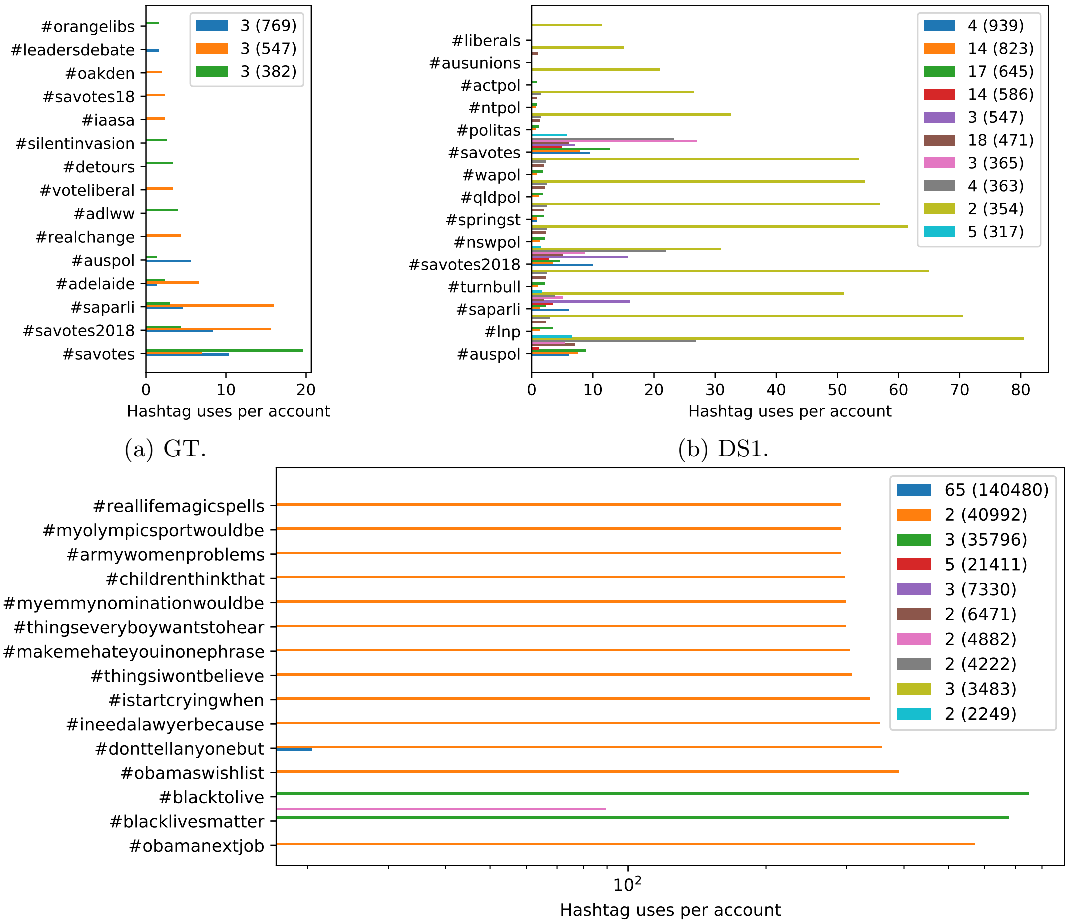
(a) GT, DS1 and RANDOM.

(b) DS2.

FIGURE 7.6. Averaged temporal graphs of HCC activities (FSA_V, γ=15).

FSA's stitching element to FSA_V may avoid this. In DS2, HCC activity increased in the second half of 2016, culminating in a peak around the election, inflated by two very active HCCs, both of which had used many predominantly benign hashtags over the year.

Hashtag use. The most frequent hashtags in the most active HCCs revealed the most in GT (Figure 7.7a). It is possible to assign some HCCs to political parties via the examination of partisan hashtags (e.g., `#VoteLiberals` and `#OrangeLibs`), although the hashtags of contemporaneous cultural events are also prominent; for example, `#SilentInvasion`, `#detours` and `#AdlWW` all relate to a contemporaneous international writers' festival. DS1 hashtags are all politically relevant, but are dominated by a single small HCC (rendered in pale green) which used many hashtags very often (Figure 7.7b). These accounts clearly attempted to widely disseminate their tweets by using 1,621 hashtags in 354 tweets. Furthermore, the hashtags they use relate to political discussions in many regions around the country (all listed hashtags that end in `pol` relate to the political discussion communities for each Australian state or the national community). Their prominence in hashtag use effectively hampers our ability to analyse the hashtag use of other HCCs, however, but seeing the results in context is important, as it helps to confirm that the pale green HCC is probably engaging in inauthentic behaviour. We can still see that a large portion of hashtag use amongst the other listed HCCs relates to `#SaVotes`, `#SaVotes2018`, and `#SaParli`, focusing on the South Australian election. If the hashtags had been irrelevant to the election, that could have provided evidence of accounts attempting to divert the discussion to other topics (because those tweets would still have needed to include the collection filter terms – i.e., ones relating to the election – to have been captured in the first place). Similarly, DS2 hashtags were dominated by a single HCC (using 41,317 relatively general hashtags in 40,992 tweets) and one issue-motivated HCC (Figure 7.7c). Given DS2 covers an entire year, it is unsurprising the largest HCCs use such a variety of hashtags that their hashtags do not appear on the chart (little evidence of most of the HCCs listed in the legend appear visible in the barchart, despite the use of a log scale on the x axis), but it is revealing that at least a few of HCCs devoted much of their content to using hashtags, while the other most active HCCs did not, indicating that different HCCs detected by searching for one coordination strategy (co-retweet) are engaging (perhaps even more strongly) in other strategies. Perhaps these hashtag disseminator HCCs acted as distractors, supporters or even polluters, contributing messages sporadically but not consistently.



(a) GT.

(b) DS1.

(c) DS2.

FIGURE 7.7. Most used hashtags (per account) of the most active HCCs (γ=15, FSA_V). The labels indicate HCC sizes (i.e., in members) and post counts. Many HCCs are too inactive to be visible.

Analysing hashtag co-occurrences can help further explore the HCC discussions to determine if HCCs are truly single groups or merged ones. Applied to GT HCC activities (Figure 7.7a), it was possible to delineate subsets of hashtags in use: e.g., one HCC promoted a political narrative in some tweets with `#OrangeLibs` (a partisan hashtag) and discussed cultural events such as the writers' festival in others with `#AdlWW` (Figure 7.8), but was definitely one group.

Given the great number of hashtags used in even moderate sized datasets such as DS1, using hashtag co-occurrence analysis to examine the broader election discussion in DS1 requires filtering to reveal the core structure underlying the semantic network. We limited the minimum frequency of co-occurrences to 100 and also removed the most frequently occurring hashtags (`#SaVotes`, `#SaVotes2018`, `#SaParli`

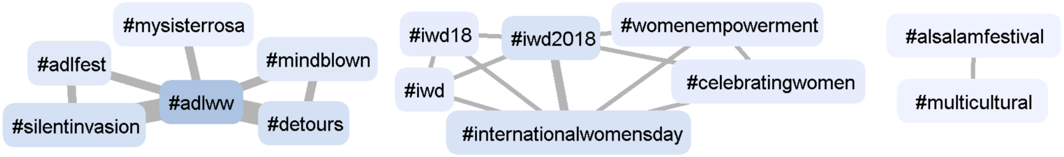


Figure 7.8. Clusters of hashtags relating to non-election events, including a writers festival, International Women's Day, and a multicultural festival, connected only when they appeared in the same tweet (GT). Wider edges represent a higher tweet count. Node colour implies the frequency of hashtag occurrences (darker means more).

and #auspol) to produce Figure 7.9. Application of Louvain cluster detection (Blondel et al., 2008) exposes five clear clusters, though domain knowledge tells us that there is interesting conflation of topics within some of the clusters. The green cluster contains subclusters relating to current affairs television programmes (#PMLive, #abc730, #Insiders, #Outsiders, #qanda and #TheDrum), political parties and advocacy groups (#OneNation, #Labor, #Greens, and #Getup) and relevant issues (#ClimateChange, #ClimateCrisis, and #StopAdani). It also includes political hashtags (e.g., hashtags ending with pol and votes) that might fit better in the yellow cluster, which is dominated by them and forms the core of the semantic network by including the heaviest edges. The purple cluster consists primarily of location names, apart from #RenewableEnergy which hangs off #SouthAustralia (the focus of the election collection).

The other two clusters make apparent the fact that Twitter is an international network and hashtag clashes can draw in content irrelevant to local issues. The hashtag #Liberals in the blue cluster can refer either to the Liberal party in South Australia (the major party that ultimately won the election) but is also used as a focus in American politics, especially rightwing politics, as reflected by the links to #MAGA, #GunControl and #2A (i.e., the 2nd Amendment of the United States' Constitution, which refers to the right to bare arms), as well as #NationalWalkOutDay. During the collection period, high school students in the United States staged a national day of protest against gun violence following a mass school shooting.[8] The red cluster also highlights content from outside the area of interest, with many terms relating to locations in other countries, possibly bound by sports, given the presence of #FullTime, #NrlStormTigers, #AflwDogsDees, and #SydVBri, the last three of which refer to Australian sporting matches between specific teams.

DS2 covers a longer period and seemed to consist of different teams of accounts driving different topics. As a consequence, its semantic network reveals clearly delineated (but often connected) discussion topics, as shown in Figure 7.10. It is immediately notable that although the accounts in the dataset were flagged as trolls implicated in attempting to influence the US election, a lot of content is not in English and, in fact, appears to target other countries. This would be consistent with at least

[8] https://www.nytimes.com/2018/03/14/us/school-walkout.html. Posted 2018-03-14. Accessed 2022-01-19.

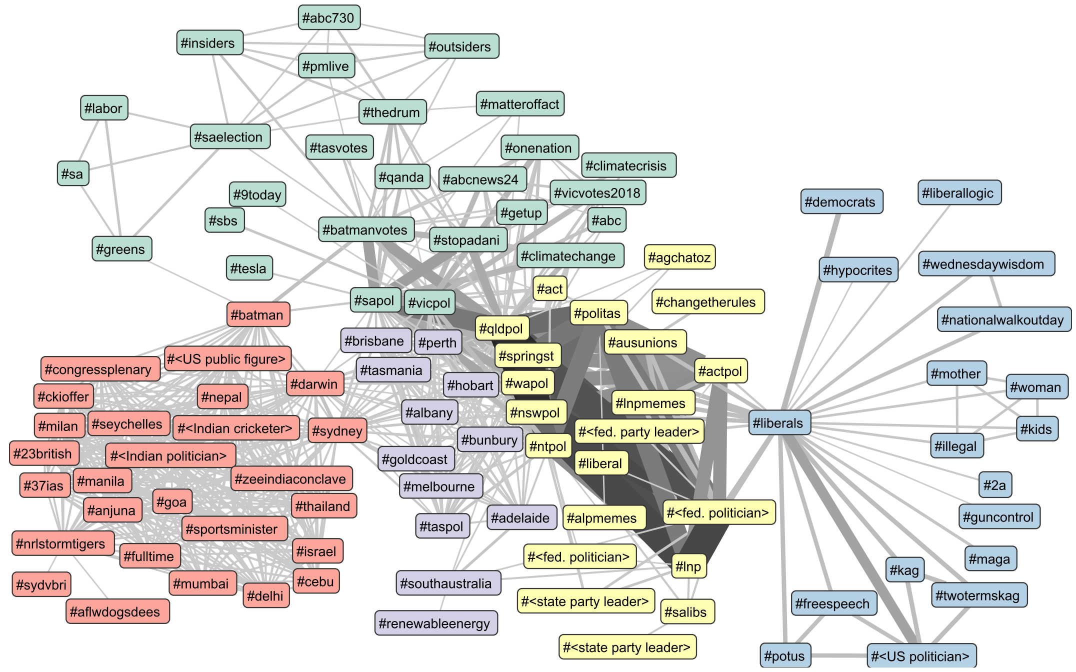


Figure 7.9. Semantic network of hashtags used in DS1, connected only when they appeared in the same tweet. The minimum edge weight is 100 and the most highly co-occurring hashtags (`#SaVotes`, `#SaVotes2018`, `#SaParli` and `#auspol`) have been excluded. Nodes are coloured according to Louvain clustering (Blondel et al., 2008), and some hashtags have been anonymised. Wider and darker edges represent a higher tweet count, and a darker background has been provided to improve contrast.

one other Russian campaign that targeted many Western audiences as well as Russians ("Secondary Infektion", Nimmo et al., 2020). Three non-English examples are apparent:

- The green cluster in the centre consists primarily of Russian news-related hashtags, perhaps aimed at a Russian audience to direct their attention to US election-related content.
- The pale blue central cluster has many hashtags related to the Middle East, including the ISIS terrorist group, but also German politicians and German names for nearby countries, such as Turkey. Germany's response to refugees from Syria escaping ISIS was politically contentious and may have been seen as an opportunity to foster divisions in the European Union and within Germany.
- The green cluster on the lower left is aimed at discussions of the United Kingdom's (UK) exit from the European Union (EU), otherwise referred to as Brexit. The UK held a referendum in 2016 on whether it should leave the EU and the campaigning caused significant division within the UK and Europe.

Other significant communities in the semantic network are the pink Top / Christian Conservatives On Twitter (`#TCOT` and `#CCOT`) cluster, tightly connected to the emerging `#MAGA` cluster supporting Donald Trump, the red cluster focused on American patriotism and the highly active brown cluster including the terms `#news`, `#local`,

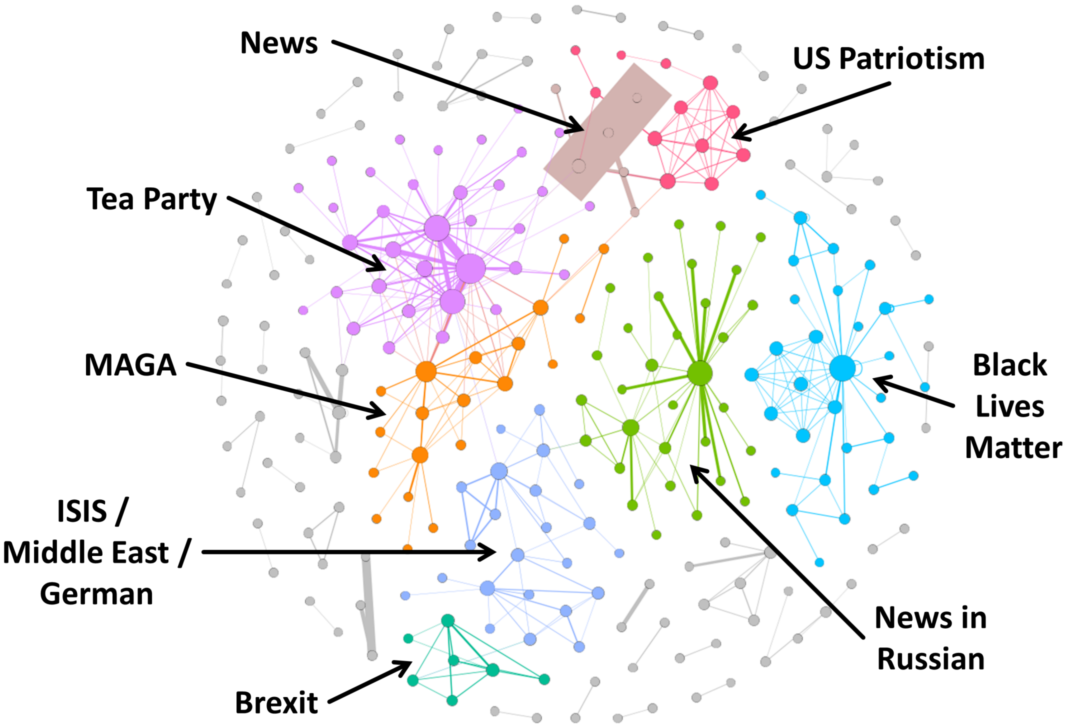
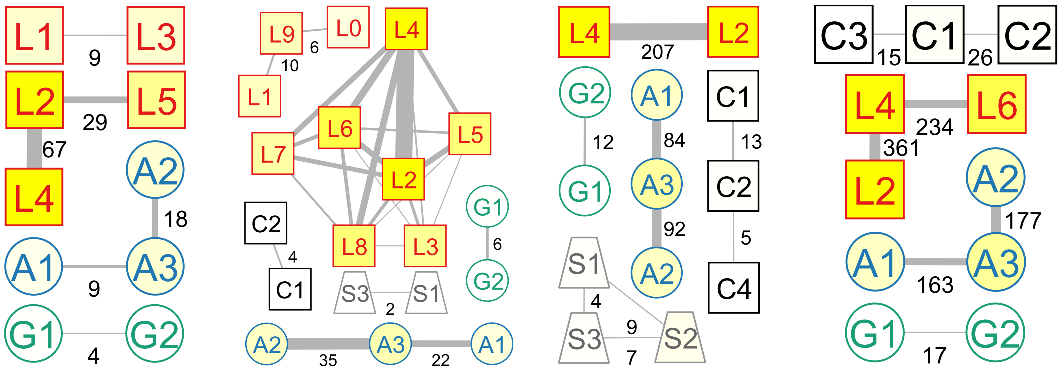


FIGURE 7.10. Clusters of hashtags used in DS2, connected only when they appeared in the same tweet. The minimum edge weight is 100. Nodes are coloured according to Louvain clustering (Blondel et al., 2008), the most prominent of which have been annotated with their topic of discussion. Wider and darker edges represent a higher tweet count.

#business and #world. The activity of HCCs shown previously in Figure 7.7c presents a different and complementary view into hashtag use in the dataset, as very little of it apparent in the semantic network—it is the combination of not only which hashtags are associated together, but also which groups of accounts are using them that provides deeper insights. By finding groups that are using otherwise entirely disjoint sets of hashtags it may be possible to identify changes in narrative, especially if HCCs can be tracked over time to see when they use which sets of hashtags.

Examining the Ground Truth. The importance of having ground truth in context is demonstrated by Keller et al. (2017) and Keller et al. (2019). By analysing the actions of known bad actors in a broad dataset, they could identify not just different subteams within the actors and their strategies, but their effect on the broader discussion. Many datasets comprising only bad actors (e.g., DS2) miss this context.

Considering GT alone, the HCCs identified consist only of members within the same political party, across all values of γ. Some accounts appeared in each window size. HCCs of six major parties were identified. Figure 7.11 shows the HCCs for each γ value. Some accounts and parties appeared at each window size, (e.g., parties L, A, G, and nodes L2, A1, G2), while some only appear in a few (e.g., parties C and S). This shows that different parties exhibited different approaches to retweeting and different



(a) γ=15.

(b) γ=60.

(c) γ=360.

(d) γ=1440.

FIGURE 7.11. Ground truth HCCs identified with FSA_V. Vertex shape = ideology (centre, left, right), colour = activity (brighter = higher), label and border colour = political party (L = red, A = blue, G = green, C = black), label = party and account identifier (e.g., 'G1' is Party G's account #1), link width = co-retweet count (some omitted for clarity).

members were involved over different time frames. Although party S members co-retweeted enough to appear in two time windows, they were not consistently active enough to re-appear in the largest time window, where their activity was overtaken by other accounts. It is particularly noticeable that the L party had two core cooperating accounts, L2 and L4, who were active enough to appear in each time window, and then a large team active in the hour-long window, implying that a deliberate strategy of team-based co-retweeting was employed (rather than a coincidental one). Rather than the posting times being highly coordinated (so that retweets could appear nearly simultaneously), it appears as if the L accounts were simply more attentive to their colleagues' tweets and retweets and retweeted them when they saw them (which often occurred within an hour), as could be expected of any social media-savvy group.

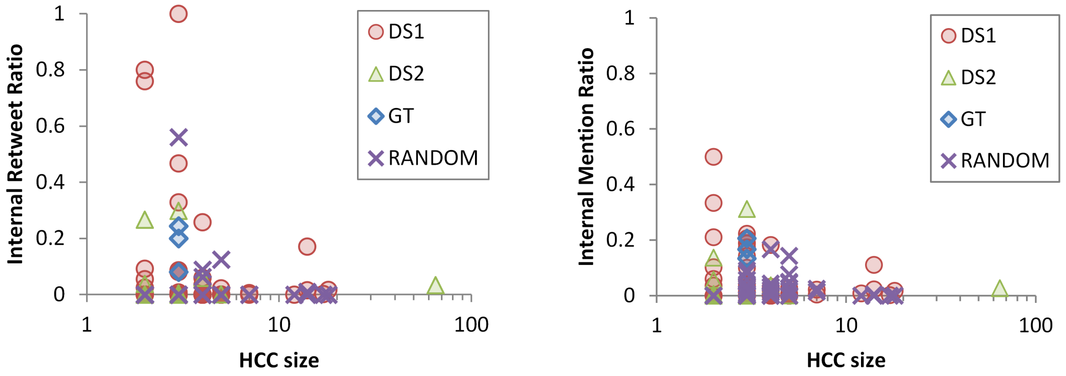
Examining the content of these HCCs confirmed that they were genuine communities engaging in co-retweeting (though not necessarily deliberately). The top retweeted tweets of each HCC (FSA_V, γ=15) are shown in Table 7.7. Using the tweets each HCC posted, it is possible to attribute each to a political affiliation, if not a party, without resorting to inspecting each member's identity.

7.4.3.3 Focus of connectivity (RQ3)

The IRRs and IMRs for the HCCs in the DS1, DS2, GT and RANDOM datasets are shown in Figure 7.12. The larger the HCC size, the greater the likelihood of retweeting or mentioning internally, so it is notable that DS2's largest HCC has IRR and IMR's of around 0, though even the smaller HCCs have low ratios. Ratios for the smallest HCCs seem largest, possibly due to low numbers of posts, many of which may be retweets or include a mention, inflating the ratios. The hypothesis that political accounts would retweet and mention themselves frequently is not confirmed by these

TABLE 7.7. The most retweeted tweet in each GT HCC (FSA_V γ=15). *NB, URLs starting with 'https://t.co/' refer back to the original retweeted tweet's URL, and are obscured here for readability and anonymity.

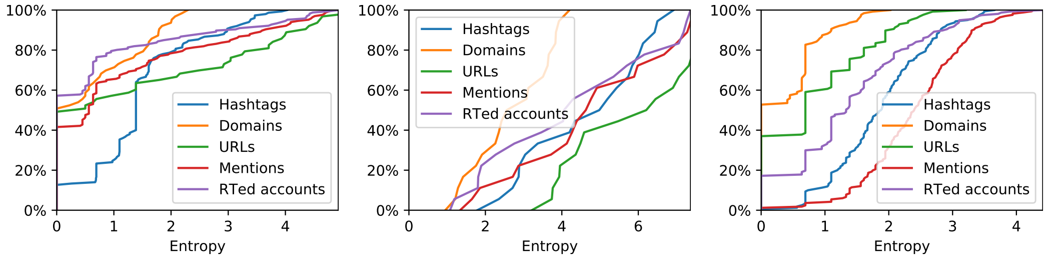
RT `@alpsa`: A message from former `@AustralianLabor` Prime Minister, ⟨REDACTED⟩. https://t.co/⟨URL⟩*
RT `@⟨Reporter⟩`: Liberals promise $40m to tackle elective surgery waiting times in South Australian hospitals. `#SAVotes2018`... https://t.co/⟨URL⟩
RT `@SALibMedia`: Under Labor there aren't enough job opportunities for young South Australians. Here's what they are saying about `@⟨Labor Politician⟩` and `@alpsa` `#saparli` https://t.co/⟨URL⟩
RT `@⟨Greens Politician⟩`: The results of this state election are clear – celebrity candidates and pop up parties come and go, but the Greens... https://t.co/⟨URL⟩



(a) Retweets.

(b) Mentions.

FIGURE 7.12. The proportions of each HCCs retweets and mentions referring to accounts within the HCC (γ=15, FSA_V).



(a) DS1.

(b) DS2.

(c) RANDOM.

FIGURE 7.13. Cumulative frequency of HCCs' entropy scores for five tweet features, comparing DS1 and DS2 with RANDOM (FSA_V, γ=15). Feature variation increases along the x axis.

results, possibly because they are retweeting and mentioning official or party accounts outside the HCCs.

7.4.3.4 Content variation (RQ4)

We compared the entropy of features used by DS1 and DS2 HCCs to RANDOM ones (Figure 7.13). Many of DS1's small HCCs used only one of a particular feature, resulting in an entropy score of 0 (Figure 7.13a). In contrast, DS2's fewer HCCs have

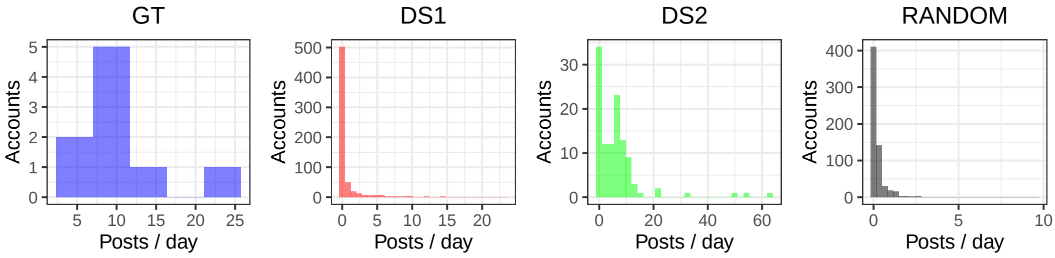
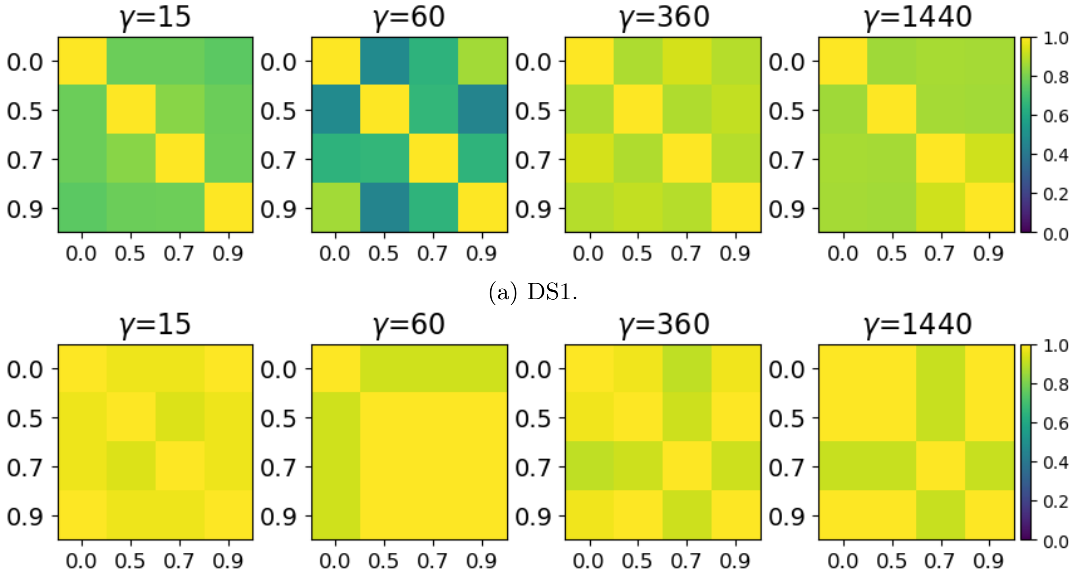


FIGURE 7.14. Histograms of the daily posting rates of accounts in the GT, DS1, DS2, and RANDOM HCCs (FSA_V, γ=15). Because the datasets cover different periods of time, the posting rate enables a fairer comparison. The distributions of DS1 and RANDOM posting rates are similar and notably different to those of DS2, while GT includes a higher proportion of more active accounts than the other datasets.

higher entropy values (Figure 7.13b), likely because more of their activity was collected (365, not 18, days' worth) and they therefore had more opportunity to use more feature values. The majority of HCCs used few hashtags and URL domains, which is to be expected as the dominating domain is *twitter.com*; this domain is embedded in all retweets as part of the link back to the original (retweeted) tweet. Compared to the RANDOM HCCs (Figure 7.13c), DS1 HCCs had lower variation in all features, while the longer activity period of DS2 resulted in distinctly different entropy distributions. Because DS1 HCC activity appears to have been more deliberate, and perhaps coordinated, it may be that the HCCs were more focused on their topic of conversation (especially when contrasted with RANDOM HCCs). Compared with RANDOM HCCs, DS1 HCCs retweeted fewer accounts, used fewer URLs (though they were from a similar distribution of domains), and many fewer mentions and hashtags. Many non-HCC accounts posted only a single retweet as their contribution to the discussion, and so it may be that a relatively high proportion the RANDOM HCC members only posted a single tweet, causing the distributions observed. The RANDOM HCC members posted 3,147 tweets compared with DS1 HCCs' 8,527 tweets, despite having the same number of members, so DS1 HCC members posted more than 2.7 times as often. Although DS1 accounts posted more tweets per individual than the RANDOM accounts, their distribution appears similar, and notably different to those of both DS2 and GT (Figure 7.14).

7.4.3.5 Consistent coordination (RQ5)

The sliding frame technique from Section 7.3.2.4 was applied to DS1 and DS2 to reveal HCCs engaging in coordination consistently in adjacent time windows. The baseline used T=1 (i.e., a sliding frame a single time window wide) and α=0.0. For the three other conditions, T was set to 5 (as γ increases approximately five times each time) and $\alpha=\{0.5, 0.7, 0.9\}$. In this way, the choice of α=0.9 would most strongly consider the contribution of LCNs from preceding time windows. Once applied for each time window, the aggregated LCNs were mined for HCCs and then the membership of these were compared in the same manner as in Section 7.4.3.1 using Jaccard similarity



(a) DS1.

(b) DS2.

Figure 7.15. Jaccard similarity of HCC membership when varying α (0.0 is the Baseline).

(Equation 3.13). As noted earlier, Jaccard similarity is stricter about set matching than the Overlap method (Equation 3.14). Even so, as can be seen in Figure 7.15, changes introduced by using the decaying sliding frame with different α values were insignificant in all cases, except for DS1 and γ=60. The implication, which is borne out when the exact network sizes (in nodes) are compared in Table 7.8, is that the previous windows did not add significant numbers of nodes, but instead increased the weight of existing edges, so the HCCs that were detected consisted of the same members working together over time, rather than splitting into subsets. To hide a team's coordination, one might expect that its members would associate separately in different time windows, but that does not appear to have happened significantly in these datasets, except in the shorter time windows in DS1, the majority of which may very well be coincidental.

Table 7.8. Statistics of discovered HCCs while varying α (FSA_V, Boost). T=5 except in the Baseline condition. N = node count, E = edge count, C = HCC count.

		γ=15			γ=60			γ=360			γ=1440		
		N	E	C	N	E	C	N	E	C	N	E	C
DS1	Baseline	633	753	167	619	1,293	151	503	1,119	127	815	2,019	141
	α=0.5	604	711	168	1,178	2,121	149	519	1,183	129	800	2,037	137
	α=0.7	578	697	160	847	1,569	149	518	1,155	130	792	1,997	136
	α=0.9	596	706	165	585	1,223	145	530	1,188	134	796	1,995	141
DS2	Baseline	113	758	19	77	394	18	98	775	15	69	380	15
	α=0.5	116	760	20	79	396	18	100	776	16	69	380	15
	α=0.7	110	756	18	79	395	18	102	777	17	69	381	15
	α=0.9	113	758	19	79	396	18	100	776	16	69	381	15

The greatest variation in node and edge count occurs in the shorter windows in DS1

Table 7.9. Features selected from both account activity and collective HCC activity based on their activity network (described in Section 7.3.2.5). IRR and IMR are defined in Section 3.3.6.

	Account-level	Group-level
Instances (Uses)	Posts, Reposts, Replies, Mentions, Hashtags, URLs	Posts, Interactions, User nodes, Hashtags, URLs, Reposts, Quotes, Mentions, Replies, In-conversations (see Section 7.3)
Unique	Mentions, Hashtags, URLs	HCC members, Nodes in the network (including URLs and hashtags), hashtags, URLs
Rates	Posts / minute	Reposts of HCC members / all Reposts (*cf.*, IRR, Eq. 3.15), Mentions of HCC members / all Mentions (*cf.*, IMR, Eq. 3.16), Replies to HCC members / all Replies
Profile	Default image (boolean), Characters in description, Characters in URL	–

(γ=$\{15, 60\}$), probably because of the greater number of accounts active in DS1 (compared to DS2): accounts have more alters to form HCCs with in DS1, which has 20.5k accounts, whereas choice in DS2 is limited to 1.3k accounts. The near doubling of accounts in DS1's HCCs when γ=60 implies accounts co-retweeted often just within a single hour, and then not again (at least not for T=5 hours). This effect is swamped by the much more active consistent co-retweeting of a smaller set of users when α is increased to 0.7 and above. Given the membership varies so little in the other conditions, an analysis of how these HCCs form and change over time is required. It is clear, however, that this approach would be best suited to filter-based collections, as they are likely to capture more accounts.

7.4.3.6 Validation via HCC classification (RQ6)

Our final validation method relies on the HCCs in GT as positive examples of coordinating sets of accounts, given it is reasonable to assume that they ought to be coordinating their activities during an election campaign (an intuition shared by Vargas et al., 2020). The purpose of this particular activity is not to build a classifier for coordinated behaviour in general, or coordinated amplification specifically, but to provide a degree of confidence that the HCCs detected in DS1 and DS2 are exhibiting similar behaviour to those in the GT.

Feature selection. As mentioned in Section 7.3.2.5, features are drawn from individual accounts *and* their groupings, specifically based on their individual and collective behaviour and homophily. For this reason, we select account-level features as well as group-level features to make up each account's feature vector, meaning that some of the values for HCC co-members will be identical. The account-level features are all drawn from their activity within the dataset, while the group-level features are drawn from the HCC's activity network (see Section 7.3.2.5) and are included in the feature vector of each member of the HCC. The account- and group-level features used are shown in Table 7.9.

TABLE 7.10. Accuracy (Acc.), positive class (COORDINATING) F$_1$-scores (F_{1P}) and unlabelled class (NON-COORDINATING) F$_1$-scores (F_{1U}) from the HCC classifiers.

		γ=15			γ=60			γ=360			γ=1440		
	Classifier	Acc.	F_{1P}	F_{1U}	Acc.	F_{1P}	F_{1U}	Acc.	F_{1P}	F_{1U}	Acc.	F_{1P}	F_{1U}
DS1	SVM	**0.91**	**0.91**	**0.90**	**0.80**	**0.82**	**0.77**	0.56	0.39	0.65	0.88	0.88	0.88
	RF	0.72	0.76	0.67	0.63	0.63	0.64	0.59	0.40	0.68	**0.90**	**0.89**	**0.91**
	BPU	0.70	0.74	0.64	0.66	0.65	0.67	**0.69**	**0.63**	**0.73**	0.88	0.86	0.89
DS2	SVM	**0.84**	**0.86**	**0.81**	0.73	0.79	0.64	0.81	0.84	0.76	0.81	0.84	0.77
	RF	0.81	0.84	0.77	**0.75**	**0.80**	**0.67**	**0.85**	**0.87**	**0.82**	**0.89**	**0.90**	**0.88**
	BPU	0.81	0.84	0.76	**0.75**	**0.80**	**0.67**	0.84	0.86	0.80	0.88	0.89	0.86

Classification results. After being trained on the GT HCCs, the classifiers were then applied to the HCCs in DS1 and DS2. We use COORDINATING and NON-COORDINATING to represent the positive and unlabelled classes, respectively. A second disjoint subset of RANDOM HCCs were created for this testing by sampling accounts outside the ground truth and training sets. Upsampling was also used to ensure the classes were balanced with at least 400 instances each.

The accuracy of the best classifier for each dataset and time window ranged from 0.69 to 0.91 (shown in Table 7.10), with performance varying between classifiers and window sizes, but mostly recognising HCC members in DS1 slightly better than DS2. This difference may be because the training data was sourced from the same online discussion (though using the behaviour of completely different accounts). F_1 scores (outside γ=360) for the COORDINATING (F_{1P}) and NON-COORDINATING (F_{1U}) instances ranged from 0.80 to 0.91 and 0.67 to 0.91, respectively. Each classifier performed best for DS1 in different time windows, except for γ=360, but all classifiers performed well, with the worst accuracy at 0.69. All classifiers also performed the least well in the six hour time window for DS1, possibly because the GT HCCs' activity coordination was most prominent over the short time frames of an hour or less, and otherwise at the day level. Even so, F_{1U} scores consistently hover around 0.7 when γ=360, which is significantly better than random, though the F_{1P} scores around 0.40 for SVM and RF indicate difficulty identifying all COORDINATING HCC members, a detail which is discussed in more detail below. The accuracy and F_1 results show that the the classifiers could all be successfully trained to recognise GT HCCs in most time windows and that the GT HCCs represented most of the HCCs in DS1 and DS2, despite the different levels of activity (DS2 HCC members interacted more than DS1 or GT HCC members in their corpus, primarily because the collection period was longer).

Table 7.11 shows precision and recall across all classifiers and datasets for the COORDINATING class. (Given our emphasis on recognising COORDINATING instances, we do not present the corresponding results for the NON-COORDINATING class here.) For all time windows, precision is high for the classifiers against DS1 (ranging

Table 7.11. Precision and recall for the positive (COORDINATING) class.

		γ=15		γ=60		γ=360		γ=1440	
	Classifier	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
DS1	SVM	**0.85**	**0.99**	**0.74**	**0.92**	0.62	0.28	0.90	**0.85**
	RF	0.67	0.88	0.64	0.62	0.73	0.28	**1.00**	0.80
	BPU	0.65	0.86	0.67	0.64	**0.78**	**0.54**	**1.00**	0.75
DS2	SVM	**0.76**	**1.00**	0.65	**1.00**	0.72	**1.00**	0.73	**1.00**
	RF	0.73	**1.00**	**0.67**	**1.00**	**0.77**	**1.00**	**0.82**	**1.00**
	BPU	0.72	**1.00**	**0.67**	**1.00**	0.75	**1.00**	0.80	**1.00**

from 0.62 to 1.00) and moderate against DS2 (ranging from 0.67 to 0.82), meaning that the HCCs are clearly discernible from the NON-COORDINATING instances (i.e., if an instance was classified as COORDINATING, then it was almost certainly a member of an HCC). Recall varies significantly for DS1 (between 0.28 and 0.99), but is perfect (i.e., 1.00) for DS2, meaning that some DS1 HCCs were rejected incorrectly, while all DS2 HCC members were identified. The recall scores for γ=360 explain why the F_{1P} scores were so low in Table 7.10, because the corresponding precision scores are still relatively high. As mentioned above, there is something particular about the six hour time window (γ=360), as the GT HCC members (via their account features and group behavioural features) were less easily distinguishable from the randomised NON-COORDINATING accounts, resulting in poorer classifier performance. The reason for this is possibly the choice of window boundaries. The time window boundaries rested at 0000, 0600, 1200, and 1800 hours, while boundaries defined more by work activity (e.g., 0200, 0800, 1400, 2000 hours) may better match human activity patterns. For other, less geographically bound datasets (i.e., ones where the activity comes from around the world, rather than from a single or small group of adjacent timezones), other ground truth may be required.

SVM was the best performing classifier for COORDINATING accounts in DS1 in the shorter time windows (γ=$\{15, 60\}$) and had close to equal top performance in γ=1440, but BPU clearly performed best in the challenging six-hour window, including with moderately better precision and markedly better recall than SVM and RF. For DS2, all classifiers performed well, with RF most often performing best, but only marginally. SVM struggled to compete in the day long period, though still achieved moderate scores for precision and accuracy. For that reason, we can argue that RF performed best overall, but the margin was minimal. Importantly, classifiers found all DS2 HCC members, though they incorrectly included some false positives.

Consequently, by accepting the assumption that the ground truth HCCs exhibited at least one type of coordination, these classifiers provide confidence that the other HCCs appear similar to the GT ones and thus may have behaved in similar ways. The question of intent remains, however. Examining the content subjected to coordination will likely provide clues, but deeper examination of behaviour to identify, e.g., Principal-Agent patterns (Keller et al., 2017; Giglietto et al., 2020b), may also be enlightening. More examples of similar coordination activities as well as other coordination types would bolster the positive training and testing sets, as well as expand knowledge regarding coordination strategies in use online. Furthermore, Vargas et al. (2020) make the point in their work on detecting SIOs that "SIO coordination should be seen as a spectrum and not a binary state... [which could lead to] ... an overestimation of accounts that are part of disinformation campaigns" (p.142, Vargas et al., 2020) potentially silencing those who need their voice to heard the most. For this reason, the application of binary classifiers for SIO detection ought to be part of a larger overall process with strong oversight.

7.4.3.7 Multiple criteria: ***Bully*ing**

Some strategies can involve a combination of actions. Magelinski et al. (2021) explored online campaigning considering only tweets that included both a hashtag and a URL, finding a number of distinct campaigns with different aims. Coordination need not focus on information dissemination, however. Behaviours that contribute to *Bully*ing by dogpiling, for example, include joining conversations started by the target's posts and mentioning the target repeatedly, within a confined time frame. As DS1 included all replied to tweets, we investigated it inferring links via co-mentions and co-convs (FSA_V, θ=0.001, γ=10 minutes), having maximised the ratio of MEW to HCC size. Of 142 HCCs discovered, the largest had five accounts and most only had two. Only 32 had more than ten inferred connections, but five had more than 1,000. These heavily connected accounts, after deep analysis, were simply very active Twitter users who engaged others in conversation via mentions, which outweighed the more strict co-conv criterion of participants *reply*ing into the same conversation reply tree.

A larger window size was considered (γ=360) in case co-conv interactions were more prevalent. FSA_V (θ=0.01) exposed little further evidence of co-conv (Figure 7.16), finding 98 small HCCs again dominated by co-mentions, not many of which had more than one inferred connection, implying most links were incidental; FSA_V did not filter these out.

This provides an argument for a more sophisticated approach to combining LCN edge weights for analysis than Equation (7.1), and that FSA_V could be modified to better balance HCC size and edge weight. Furthermore, it is likely that bullying accounts will not just co-mention accounts frequently, but have low diversity in the accounts they co-mention, i.e., they will repeatedly co-mention a small set of accounts, and spend a disproportional number of their tweets doing so. A further consideration is that participants in long discussions (reply trees) often include the author of the original tweet that sparked the discussion, and it would be misleading to include their account in results, implying that they *bull*ied themselves. Finally, patterns of behaviour that would clearly qualify as conversations were observed in the datasets that did not fit the strict 'conversation tree' model: accounts would mention several collocutors at the start of every tweet, but only reply to a tweet of one of them

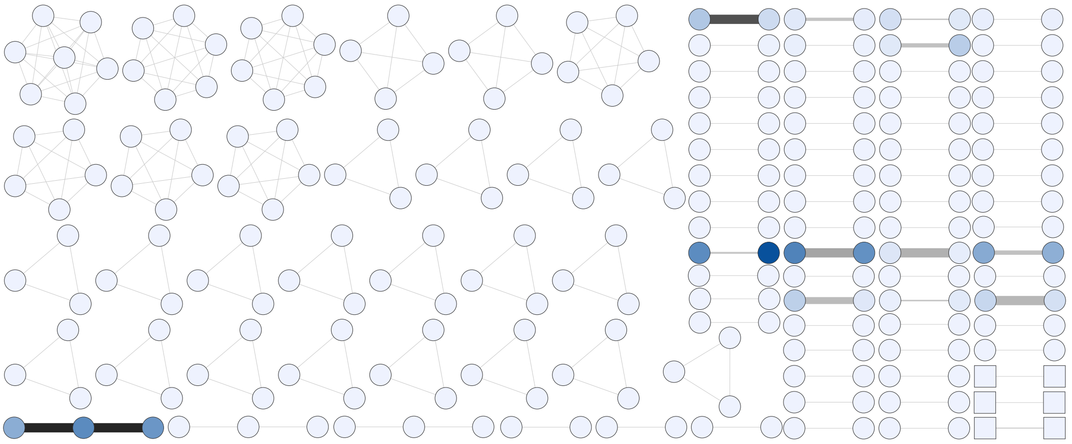


FIGURE 7.16. While searching for *Bully*ing behaviour in DS1, these are HCCs of accounts found engaging in co-mentions (circles) and co-mentions plus co-convs, i.e., engaged in both (square vertices in bottom right) (γ=360, FSA_V, θ=0.01). Edge thickness and darkness = inferred connections (darker = more). Vertex colour = tweets posted by that account (darker = more).

while continuing the conversation. Importantly, sometimes the mentioned accounts included in tweets were prominent individuals whose names were included not because they were active participants in the conversation, but because the tweeter wanted to draw their attention to the conversation (regardless of the likelihood that the attempt would succeed; e.g., some tweets included references to prominent and busy politicians who would be unlikely to wade into arbitrary online discussions).

7.4.3.8 HCC inter-relationships

To study the relationships between HCCs, we create 2-layer networks starting with the HCC network and then adding nodes representing the elements of evidence linking them, known as *reason* nodes (e.g., the tweets they co-retweet or the hashtags they use in common). Figure 7.17 shows the largest component after such expansion was conducted on the HCCs in Figure 7.16. HCC accounts (circles) share colours and the distribution of the reasons for their connection (diamonds) show which other accounts are uniquely mentioned by an HCC and which are mentioned by more than one HCC. Heavy links between HCC accounts with few adjacent reason vertices imply these accounts are mentioning a small set of other accounts on many occasions.

7.4.3.9 *Boost*ing accounts, not just posts

It is possible to *Boost* an account rather than just a post. Returning to DS2, we sought HCCs from accounts retweeting the same account (FSA_V, γ=15), and found that the hashtag use revealed further insights (Figure 7.18). No longer does one HCC dominate the hashtags. Instead clear themes are exhibited by different HCCs, but again, they are not the largest HCCs. The red HCC uses `#BlackLivesMatter` and other Black rights-related hashtags (including `#BLM`, `#BlackToLive`, `#BlackSkinIsNotaCrime`,

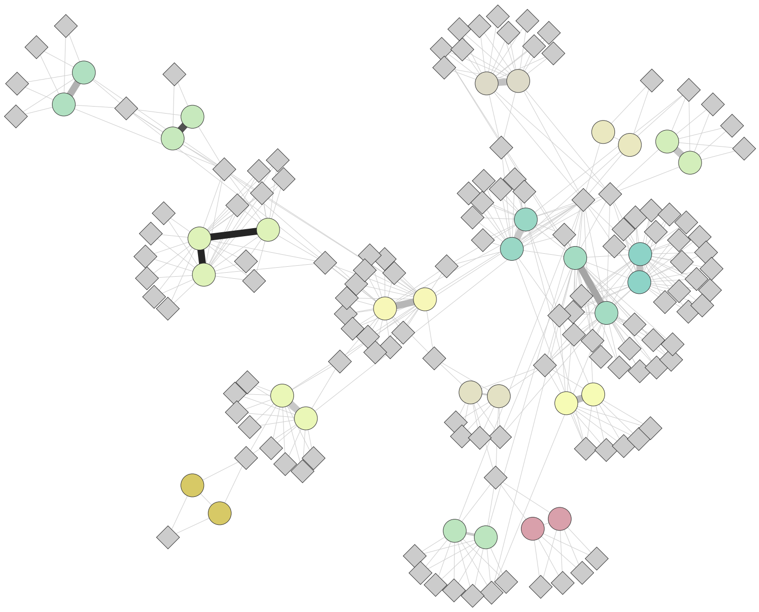


Figure 7.17. A network of DS1 HCC accounts (circle vertices) connected to the accounts they mention or conversations they join (diamonds). Accounts in the same HCC share a colour. Clear communities surrounding HCCs indicate who they converse with, and which conversants are co-mentioned by multiple HCC accounts. The width and darkness of the edges between HCC accounts indicates the weight of evidence linking them (darker implies more).

#PoliceBrutality and #BTP[9]), while the purple HCC uses pro-Republican ones (#MAGA and #TCOT), and the green HCC is more general. Given the number of tweets these HCCs posted over 2016 (at least 16,849), it is clear they concentrated their messaging on particular topics, some politically charged. It is arguable that their contributions helped inflame tensions and stoke divisions in socially sensitive topics, not just in the United States, but in the UK as well, and at the very least sought to draw the attention of others.

The green HCC may be acting in distractor or polluter roles, as previously suggested, given their contribution of 72,428 tweets over the year (an average of nearly 40 tweets per account every day).

7.4.3.10 Validation of inauthentic behaviour detection

The approach presented can be used to perform analytics similar to the Rapid Retweet Network used by Pacheco et al. (2020), who used it to expose tight clusters of bot-like accounts, which retweeted the same tweet within 10 seconds of it appearing. We varied this for the DS1 dataset (due to its small nature) and searched for accounts which retweeted the same tweet within 10 seconds, regardless of the age of the original tweet. We discovered a tight cluster of accounts, most with relatively high Botometer CAP

[9] BTP refers to the British Transport Police, the conduct of which was discussed in accounts of the arrest of a Black man at a London train station in mid-2016, e.g., https://www.theguardian.com/uk-news/2016/jul/28/man-complains-after-police-place-spit-hood-over-head-during-arrest-london-bridge. Posted 2016-07-28. Accessed 2022-01-11.

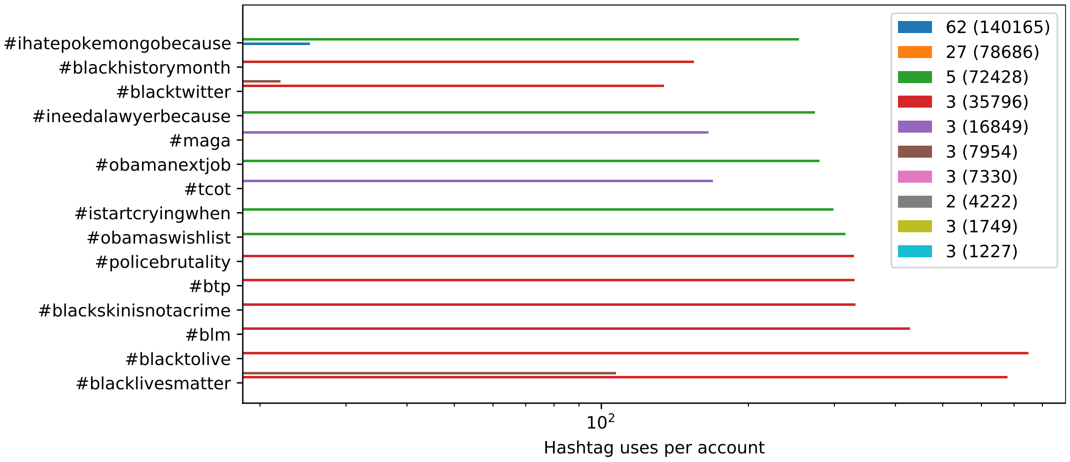


FIGURE 7.18. Hashtag uses of the most active HCCs boosting accounts (FSA_V, γ=15). The labels indicate HCC member and tweet counts. Many HCCs are too inactive to be visible.

scores[10] (Davis et al., 2016), shown in Figure 7.19. The scores were as follows: node 26: 0.787; node 22: 0.381; node 2: 0.949; and node 17: 0.464. All were high relative to the other accounts in the corpus, most of which had scores well below 0.2; all four were had scores well above 0.2, but the scores of two were also well above the 'bot' threshold of 0.6. On further inspection, they appeared to support vocational training and left-wing issues and posted retweets almost exclusively, but the content all related to the election. This finding enhances the bot ratings by making it clear which bots (or bot-like accounts) appear to work together. It also raises further questions regarding bot detection systems, however, as some of the accounts appeared to be genuinely human, though unusually active. These accounts appeared to work together to actively disseminate messages aligned with their preferred narrative, though with a very low IRR (just shy of 10%) despite most of their activity being retweets (97.7%), so to a certain degree it matters not whether they are automated or genuinely human-driven, but whether they are engaging in astroturfing or other inauthentic behaviour. In this circumstance, they may be genuine agenda-driven users, but they were definitely all highly attentive to the same sources. Alternatively, when we consider their bot ratings more closely, it is possible that there is a mixture of account types, with node 26, in particular, acting as an automated 'cheerleader' for nodes 22 and 17. Examining relative timings of their posts (to answer whether node 26 consistently was the second co-retweeter when paired with nodes 22 and 17) could reveal support for this hypothesis.

7.4.3.11 Performance

In Table 7.12 we present the timings observed for the stages of processing for DS1 and DS2 conducted on a Dell Precision 5520 laptop equipped with an Intel Core i7-7820HQ CPU (2.9GHz), 32Gb RAM, and an NVMe PC300 480Gb SSD, running Windows 10.

[10] The English score variant was used as both the datasets were either primarily in English or aimed at English speaking audiences.

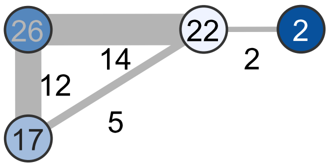


FIGURE 7.19. The most active DS1 co-retweet HCC (γ=10 seconds). Node label = post count, node colour = Botometer scores (higher = darker), link thickness and label = co-retweet occurrences.

TABLE 7.12. Execution times (in seconds).

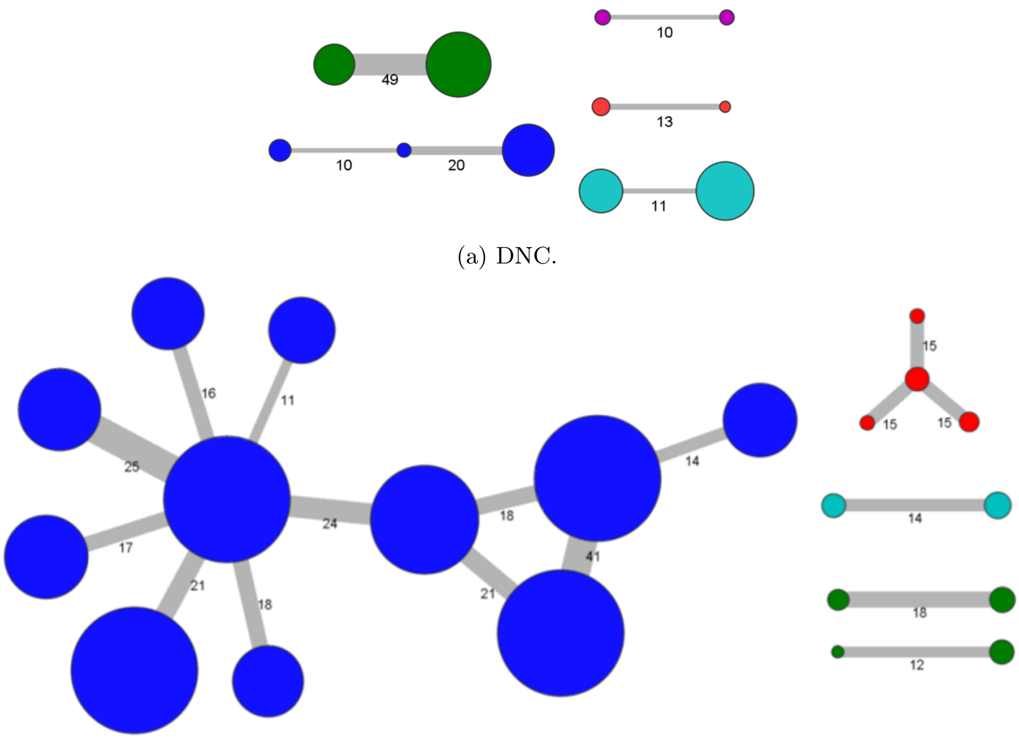
	DS1				DS2			
Tweets	115,913				1,571,245			
Parse raw (Step 1)	19.0 (from JSON)				74.0 (from CSV)			
Window size γ (minutes)	15	60	360	1440	15	60	360	1440
Find evidence and build LCNs	15.0	28.0	123.0	427.0	121.0	106.0	246.0	567.0
Aggregate LCNs	27.0	65.6	168.5	170.7	70.4	55.2	35.6	22.7
HCCs: FSA_V	28.3	58.2	126.1	209.3	6.3	4.2	5.8	5.0
HCCs: kNN	9.0	22.7	97.5	206.4	4.3	4.3	4.7	4.6
HCCs: Threshold	5.2	11.9	34.6	64.0	2.2	2.3	2.7	2.7

Parsing raw data is relatively cheap, with DS2's 1.5m tweets processed in just over a minute, and LCN construction is dependent on the degree of activity and the number of accounts. DS1's larger account pool increased the size of the networks generated, and all associated post-processing. The size of DS1 LCNs were an order of magnitude greater than DS2's (in nodes and edges), resulting in increasing execution times for aggregation and HCC extraction.

7.4.4 Case study: The 2020 US political conventions

Complementing the detailed validation presented above, we offer a case study to demonstrate the practical utility of our method. The technique was also used to examine coordinated amplification in the `#ArsonEmergency` case study presented in Chapter 5. For that analysis, see Section 5.3.4.

This case study relates to the search for social bots attempting to influence US political discussions in the lead up to the 2020 US presidential election. We specifically examined the online discussion surrounding the Democratic and Republican National Conventions in August 2020, at which the parties formally nominated their candidates for president. For a 96 hour period over each 4-day convention, tweets were filtered using RAPID (Lim et al., 2019), starting with `#DemConvention` and `#Rnc2020` as seed hashtags for the Democratic National Convention (DNC) and the Republican National Convention (RNC), respectively. For the three hours prior to the formal collection period, RAPID's topic tracking feature was enabled, adding hashtags that appeared frequently in the tweets observed, bolstering the filter terms for each convention:



(a) DNC.

(b) RNC.

FIGURE 7.20. Co-retweeting HCCs detected during the August 2020 DNC and RNC (Threshold, t=0.1, γ=10 seconds). Nodes are HCC member accounts, sized by the number of tweets they contributed to the discussion, and joined by edges sized and labelled according to the number of times they retweeted the same tweet. The nodes are coloured by Louvain cluster for convenience, but any matching colours between the DNC and RNC subfigures has no meaning.

- DNC: `#DemConvention`, `#BidenHarris`, `#BidenHarris2020`, `#KHive`, `#SignsAcrossAmerica`, `#UnitedForBiden`, and `#WeWantJoe`;
- RNC: `#Rnc2020`, `#RncConvention`, and `#NeverTrump`.

Despite the disparity in hashtags, each dataset ultimately comprised approximately 1.5 million tweets by over 400 thousand unique users at each convention. Bots are often used to boost tweets, reaching other accounts that follow them, or by flooding hashtag communities or gaming trending algorithms (Woolley, 2016; Hegelich and Janetzko, 2016; Keller et al., 2019; Graham and Keller, 2020; Graham et al., 2020c). Social bots are specifically designed to mimic genuine human users, hiding the fact they are automated (Ferrara et al., 2016; Grimme et al., 2018; Cresci, 2020). They do this to avoid detection, and in doing so can contribute to astroturfing campaigns, artificially boosting narratives while making them appear as simple popular grass roots movements.

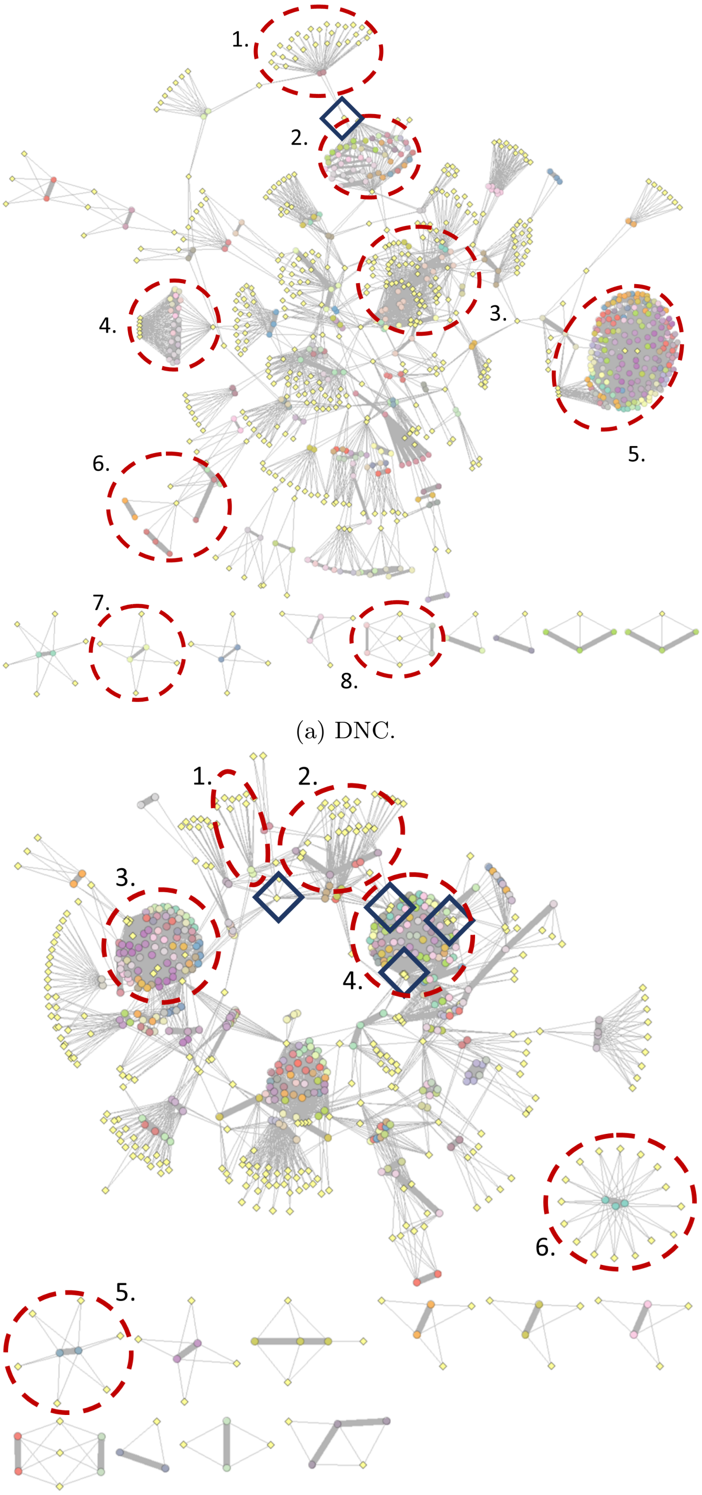
By searching for *Boost*ing via co-retweet (Threshold, t=0.1, γ=10 seconds), several HCCs were identified in each convention (see Figure 7.20). Analysis of the HCC members using Botometer (Davis et al., 2016) found the majority had CAP scores above 0.6, indicating a high probability that they made use of automation. Further analysis of the HCCs' content provided some indication of their agendas, and examination of their account age and posting rates enabled categorisation into official accounts (verified by Twitter), unofficial reposters (topic-focused aggregators), and accounts that gave the appearance of typical human users. These 'normal people', however, posted at very high average daily rates for years, often at far greater rates than previous automation detection methods have used (e.g., 50 tweets a day, Neudert, 2018).

The largest HCC (the large blue HCC in Figure 7.20b) consisted of a cluster of potential social bot accounts supporting an official political campaign account, `@TrumpWarRoom`, responsible for 2,085 tweets during the Republican Convention. For each pair of members in each HCC, we considered the proportion of time that one account retweeted a tweet before the other, to determine if both accounts were potentially working together (in which case, they would be equally likely to retweet a tweet first), or if one was a 'cheerleader' for the other (in which case the cheered account would always retweet first, quickly followed by the other account). We found strong evidence that at least three of the accounts were cheerleaders for `@TrumpWarRoom`, retweeting the same tweet within ten seconds on 214, 229, and 89 occasions over the four day collection period. These particular accounts had daily tweeting rates of 78.7, 209.4 and 147.4 tweets per day for 0.9, 8.5 and 3.6 years, respectively. Given the age of these accounts, it is clear that they have successfully avoided Twitter's bot scanning processes for some considerable time.

We also applied co-hashtag analysis (FSA_V, θ=0.3, γ=10 seconds) to the two datasets and plotted 2-level networks of the resulting HCCs with the hashtags they used (Figure 7.21). Regardless of the content, a number of structures are immediately apparent. These include:

- clusters that are bound by a few yellow diamond hashtag nodes (e.g., DNC clusters 5, 6 and 8) or lie between hashtags (e.g., DNC clusters 2 and 4);
- fan shapes that consist of a small number of accounts using a wide variety of hashtags (e.g., DNC clusters 1 and 7);
- island clusters that are bound by the hashtags they use but are isolated from the broader community which has ignored the hashtags they are using (e.g., DNC clusters 7 and 8).

The fact that the clusters are coloured according to their HCC in Figure 7.21 highlights what FSA_V regards as distinct clusters are, in fact, bound together by the topics they are discussing (by the hashtags they are co-using). This indicates that there may be benefit in re-introducing the re-stitching step in FSA (Şen et al., 2016) that FSA_V avoids, or also experimenting further with FSA itself. Using conductance cutting (Brandes et al., 2008) for cluster detection aligned better with the visible clusters, but these clusterings may be somewhat misleading, as it may combine polarised HCCs, as can be seen on closer inspection below.



(a) DNC.

(b) RNC.

Figure 7.21. Account/hashtag 2-layer networks of co-hashtag HCCs and the hashtags they used during the August 2020 DNC and RNC (FSA_V, θ=0.3, γ=10 seconds). Circular nodes are HCC member accounts, coloured by HCC, and hashtags are yellow diamond nodes. The links between accounts are sized by their co-hashtag frequency (i.e., how often they used the same hashtag in the same time window). *visone*'s *stress minimisation* layout was used for both networks. Notable clusters have been highlighted with red dashed ovals and numbered, while particular hashtag clusters have been highlighted with blue diamonds.

Several co-hashtag clusters in Figure 7.21a provide insight into the nature of parts of the online discussion.

- Cluster 5 is closely centred on two hashtags (`#Goodyear` and `#Ohio`) that relate to then US President Donald Trump's call for a boycott of Goodyear tires,[11] though it is unclear whether the surrounding accounts are for or against the boycott. Several hashtags linked on the left edge of the cluster indicate that some are against, as they refer to support for the then Democratic candidate Vice President Joe Biden.
- The fan-shaped cluster 1 at the top consists of two accounts that are attempting to disseminate their message across America, as each hashtag is a US state code (e.g., `#GA` for Georgia) or a minority group (e.g., `#Latinos`). These hashtags are all apparently unique, apart from the one highlighted just below cluster 1 surrounded by a blue diamond (`#BLM`) linking cluster 1 to cluster 2, and the one to the left (another state code, `#NC`, for North Carolina).
- Cluster 2 binds a number of HCCs spanning two relatively disjoint hashtags, one being `#vote` (below the cluster) and the other being the name of a musician who had recently encouraged his fans to vote.
- Cluster 3 is more diffuse than the others and appears to relate to a discussion of data science and big data in the context of the election campaign.
- Cluster 4 appears to join a number of potentially opposed HCCs, as they refer to `#Trump2020Landslide` and `#snowflakes` as well as `#Epstein`[12] and `#TrumpVirus` (a condemnation of the Trump administration's handling of the response to the COVID-19 pandemic), the final hashtag which links the cluster into the broader community.
- The island clusters 7 and 8 are focused on groups of particular politicians, which were not picked up by the broader community: Republicans who had pledged to vote for the Democratic presidential candidate and US Congress members known to campaign for social equality, respectively.

The links between the clusters are sometimes deceptive. Already, we observed that some single clusters include polarised HCCs, however it is also possible to see internally (politically) consistent clusters that are linked but also contrary in their views. DNC cluster 1 (in Figure 7.21a) is linked to the left by `#NC` to another left-leaning cluster (calling for gun control), which itself is linked to the left by `#America` to another small

[11] The Goodyear factory in Ohio banned clothing with political messaging, including the Trump campaign's MAGA caps, during the election campaign: https://www.abc.net.au/news/2020-08-20/donald-trump-calls-for-goodyear-boycott-over-alleged-maga-ban/12577372. Posted 2020-08-20. Accessed 2022-01-11.

[12] Jeffrey Epstein was a billionaire arrested for sex crimes before dying in custody, however he was known to Donald Trump, and therefore this hashtag's use can be seen as an attack on his political campaign: https://www.forbes.com/sites/lisettevoytko/2020/10/18/spider-book-excerpt-how-trumps-presidency-helped-expose-jeffrey-epstein/. Posted 2020-10-18. Accessed 2022-01-11.

cluster, which is clearly right-leaning (one of its hashtags is #VoteRedToSaveAmerica). These visualisations may highlight how HCCs can be merged, but care must be taken when interpreting them.

Analysis of the RNC co-hashtag HCCs and their hashtags in Figure 7.21b offers further examples of these observations and offers new insights. Clusters 1 and 2 are joined by the blue diamond-highlighted hashtag, #BlackLivesMatter, but cluster 1 is a detractor group (using #AllLivesMatter) while cluster 2 is a supporter group using several Black rights-related hashtags. Cluster 4 discusses riots following Black Lives Matter protests in Kenosha, Wisconsin, however, while the two sets of hashtags highlighted at the top of the cluster relate mostly to current events (e.g., #Kenosha and #COVID19 on the left, and #KenoshaRiots and #ThursdayThoughts, plus #WalkAway, which links to a small fan, as it is a pro-Republican statement to avoid conflict), the hashtags at the bottom of the cluster are more clearly right wing or conservative in nature, referring to a relevant media organisation, #Kag2020 (Keep America Great, a pro-Trump slogan) and #CCOT. Whereas cluster 4 in Figure 7.21a includes polarised HCCs, the placement of the hashtag nodes they are linked to offers no guidance on how they might be separated. Cluster 4 in Figure 7.21b indicates that an alternative layout algorithm may aid analysis. Cluster 6 represents a concerted anti-Trump effort with many attacking hashtags, but the isolation of the HCC at the cluster's centre makes it clear that not many of the others tweeting during the RNC took its lead. Cluster 5 is an effort to draw attention to an instance of police brutality, which also did not gain traction with the broader co-hashtag community.

7.5 Conceptual Comparison and Critique

Methods to discover coordinated behaviour by inferring links between accounts based on related interactions is not unique. Cao et al. (2015) and Giglietto et al. (2019) identified groups of accounts based on the URLs they shared in common, while Lee et al. (2013), Keller et al. (2019), Dawson and Innes (2019) and Graham et al. (2020c) relied on the similarity of the content posted by accounts to do the same. Giglietto et al. explicitly added a temporal element by considering potential links only between accounts that share a URL within a constrained time frame. Their "rationale is that, while it may be common that several entities share the same URLs, it is unlikely, unless a consistent coordination exists, that this occurs within the time threshold and repeatedly."[13] To the knowledge of the authors, only three other proposed approaches appear to generalise the idea to allow links between accounts to be inferred based on a variety of behaviours common to the major OSNs: Pacheco et al. (2021), Graham et al. (2020c), and Nizzoli et al. (2021).

[13] Quoted from the README of Giglietto et al. (2019)'s open source code (version at 2021-01-19): https://github.com/fabiogiglietto/CooRnet. Accessed 2022-01-11.

Pacheco et al. (2021)'s method creates strong ties between accounts that share similar behavioural traits. Behavioural traits are extracted from social media data (e.g., hashtags or URLs) and, together with the accounts using them, a bipartite graph is created, similar to our account/reason networks. A weighted account network is projected from this bipartite network, linking accounts that have edges to the same trait node. The more shared traits, the heavier the edge between accounts. Finally, the account network undergoes cluster analysis specific to the nature of coordination sought. In their examples, Twitter accounts linked by using the same account handle are divided into clusters by virtue of the connected component in which they appear. A second example examining share market "pump and dump" scams links accounts based on the similarity of the text they post, using *text frequency–inverse document frequency* (*tf-idf*), and then clusters are discovered by simply filtering out edges with a final weight less than 0.9. A third example connects accounts that use multiple hashtags in the same order in their tweets. The approach was employed searching for co-retweeting communities spreading propaganda attacking the Syrian White Helmet movement by linking accounts that retweeted tweets within 10 seconds Pacheco et al., 2020.

In contrast, Graham et al. (2020c)'s "coordination network toolkit"[14] (CNT) is written in Python (as is ours), and relies on a database populated with information extracted from tweets to carry out searches for: coordinated retweeting (retweeting the same tweet); co-tweeting (tweeting identical text); co-*similarity* (tweeting similar text); co-linking (sharing the same URL); and co-replying (replying to the same tweet). The database implementation uses an inner join to improve the performance of searching for evidence of coordination between pairs of accounts (which, similar to our approach, requires pairwise comparison of all accounts in the dataset). This implementation would need to be modified to suit a streaming data source, but could theoretically be applied to data from a variety of OSNs as it employs a technique similar to our Steps 1 and 2.

The approach of Nizzoli et al. (2021) is very similar to ours, however it explicitly begins by selecting a set of users of interest, whereas we begin with a corpus of posts and our set of users is defined by those present in it. Nizzoli et al. make clear that the users may be defined by using the corpus in the same way at the outset, or may be otherwise nominated by virtue of being superproducers or superspreaders or followers of a prominent account. They also introduce a filter step before the extraction of HCCs. Pacheco et al. (2021) filter their user similarity network with an arbitrary filter, which, as pointed out by Nizzoli et al. (2021), results in a binary classification of coordinating and non-coordinating users, but importantly disregards the effect of the network structure. Instead, Nizzoli et al. (2021) rely on multiscale filtering approaches for complex networks, which retain network structures (not just individual edges) based on statistical significance. Furthermore, these can be scaled to retain

[14] https://pypi.org/project/coordination-network-toolkit/. Accessed 2022-01-11.

more or less of the network, permitting examination of the 'degree' of coordination, not just a binary answer to whether or not it is present. They propose an iterative algorithm at this point for detecting clusters of coordinating users, which makes use of an increasingly strict definition of user similarity (i.e., coordination) and each time relies on the communities found in the previous step as the starting point, guaranteeing they are kept in some form. This makes it possible to track communities at different levels of coordination, similar to how k-core decomposition provides insight into how deeply particular nodes and structures are embedded within a network. Finally, they apply a validation step, studying the resulting networks with network measures, and text analysis of the posts of the HCCs, but all as a function of the resolution at which the HCCs were detected. The FSA_V algorithm is our alternative to their filtering and cluster detection steps. The ability for Nizzoli et al. (2021) to examine different degrees of coordination is a distinguishing factor, however they also (just like Pacheco et al., 2021) must decide beforehand what similarity measure to connect users with – this is equivalent to the behaviours that underpin the coordination strategies we discussed in Section 7.2, however they make the point that the similarity measure may involve any relevant information about the user profiles, not just their behaviour within the corpus. The temporal aspect of the coordination is not discussed, presumably as it is assumed to be a component of the user similarity measure.

Giglietto et al. (2019)'s CoorNet R package does not allow specification of a time window directly, but instead uses a proportion threshold to determine what to regard as an anomalously small but active time window, and thus requires access to an entire dataset. It is designed to study Coordinated Link Sharing Behaviour (Giglietto et al., 2020a) and thus only considers URLs in posts, however, it accepts URLs from a variety of sources, including via CrowdTangle[15] and MediaCloud.[16]

Our method is similar to all of these but is described in greater detail, relies upon a discrete window-based approach to apply temporal constraints, and we provide and evaluate a novel cluster extraction algorithm, and an open source implementation is available. By applying time constraints in discrete windows, connections may be missed across windows, but this makes it easier to apply in near real-time streaming settings. If one were to infer connections between accounts as each new tweet is posted, it could create a potentially significant, ongoing processing cost depending on the number of unique accounts observed in the current time window. As new posts arrive, new nodes may need to be added to the account network, while others may need to be removed, along with their adjacent edges (which, it is important to recall, represent indirect evidence of coordination, not the individual timestamped interactions as one might find in a social network based on direct retweets, mentions or replies). Furthermore, this constantly updated account network must be complete, i.e.,

[15] https://www.crowdtangle.com/. Accessed 2022-01-11.

[16] https://www.media.mit.edu/projects/media-cloud/overview/. Accessed 2022-01-11.

edges should always be added in case the evidence they represent may be consolidated by future posts.

If the choice of time window is very short (e.g., 10 seconds, as per Pacheco et al., 2020), and LCNs from adjacent windows are aggregated (as per our method), the absence of a truly sliding window like Graham et al. (2020c)'s may not significantly affect results, as ongoing high levels of coordination will appear over multiple windows. In contrast, if the time window is longer (e.g., five or more minutes), then the hard boundary between windows may cause coordinated activities to be missed. The question is, then, what kind of coordination is being sought. Teams of bots tweeting or retweeting the same tweet within small time frames will be vulnerable to detection, however a deliberate covert human team with sockpuppet accounts may escape detection (at least initially) by varying the time frame over which retweets are posted (e.g., spread them unevenly over an hour or more), but if the same accounts cooperate for extended periods, our method will find them once their activities are aggregated. One type of coordination that is very difficult to detect is single event boosts of a post: e.g., when, say, 1,000 paid accounts retweet or reply to a single tweet or comment on an online review. In a large discussion, 1,000 tweets will not stand out, but, depending on how connected the paid accounts are to the broader discussion, they may spread the content a considerable distance through the network. Furthermore, gaming OSN trending algorithms may not be difficult,[17] and even a thousand retweets may result in a valuable degree of influence in comparatively smaller communities (e.g., Australia).

As a final comment, all methods discussed in this section are suited to post-collection analysis. Graham et al. (2020c)'s relies on the power of database systems to build the LCN but avoids exploring clustering analysis for HCCs. Giglietto et al. (2019)'s relies on R's expressivity and filtering based on anomaly detection, while our implementation uses Python and batch mode processing to enable flexibility in the choice of cluster analysis technique. Pacheco et al. (2021)'s implementation is in Python,[18] but has been applied to very large datasets, managing them with *Big Data* file formats. Nizzoli et al. (2021) do not mention the availability of their implementation, only that their test dataset will be forthcoming.

Our paper is the only one of these to address the concept of searching for multiple coordination criteria, and how to treat the combination of their evidence, and the attendant complications explored in Step 4 of Section 7.3.1. Magelinski et al. (2021) have proposed second-order interactions to address this, but only in combination (e.g., connect accounts using the same hashtag+URL in tweets). In fact, the other papers primarily treat the coordination criteria (i.e., user similarity measure) as

[17] OSN gaming efforts of the form "Let's get X trending" are quite common in Australia, e.g., https://twitter.com/Timothyjgraham/status/1351742513044807680. Posted 2021-01-20. Accessed 2022-01-11.

[18] https://github.com/IUNetSci/coordination-detection/. Last updated 2020-12-08. Accessed 2022-01-29.

entirely dependent on the current investigation and no generalisation of the concepts is discussed.

7.6 Conclusion

As coordinated online influence activities grow in sophistication, so must our automation and campaign detection methods also improve in order to expose the accounts covertly engaging in “orchestrated activities” (Grimme et al., 2018). We have described several strategies for coordinated amplification, their purpose and execution methods, and demonstrated a novel pipeline-based approach to finding sets of accounts engaging in such behaviours in two politically relevant Twitter datasets. We have also explained and provided examples of how our method is conceptually applicable to a range of OSNs based on common features and functionality. Using discrete time windows, we temporally constrain potentially coordinated activities, successfully identifying groups operating over various time frames. Guided by research questions posed in Section 7.1, our results were validated by using a variety of techniques, including developing three one-class classifiers to compare the HCCs found in two relevant datasets, plus a randomised one, with HCCs from a ground truth subset. Two case studies of contentious online discussion were also presented, in which our technique was applied to reveal insights into the activity of polarised groups in one and the activity of social bots and bot-like accounts in the other. The algorithmic complexity of our approach was discussed, as well as comparison with several similar contemporary approaches.

This technique provides a valuable addition to the suite of analytical tools used in deep forensic investigations of SIOs, such as Benkler et al. (2018), Jamieson (2020) and Nimmo et al. (2020), as well as law enforcement and open source investigation groups – in particular, this technique can help reveal entities that deliberately avoid direct connections to hide their cooperation.

The temporal analysis of HCC evolution and their impact on the broader discussion, theoretical questions of the semantics of edges in LCNs, the ability to distinguish between authentic and inauthentic coordinated behaviour, improvement of HCC extraction and validation techniques and application in near real-time processing environments all provide opportunities for future research in this increasingly important field.

Coordinated amplification remains a simple but key strategy in the toolbox of those running disinformation campaigns (Paul and Matthews, 2016).

7.7 Part Summary

In this Part, we have begun to address **TRQ4** by developing and demonstrating a method to find groups coordinating their behaviour to amplify content, which can be used to amplify and normalise fringe voices to permit them into the mainstream discourse (Woolley and Guilbeault, 2018). With a case study of US political discussion, we identified groups promoting propaganda, but in doing so highlighted the difficulty in distinguishing genuine enthusiasm and support from malicious inauthentic coordination. Nevertheless, the method presented forms a solid foundation on which to build and conduct further research into CIB.

Chapter 8

Conclusion

In this thesis, we explored the extent to which groups engaging in coordinated social media behaviour can be identified and studied with a computational social science approach. We also examined the information environment and polarised discussions vulnerable to misinformation and disinformation, in which such groups operate. In particular, groups active in coordinated *inauthentic* behaviour (CIB), whether it be for ideological, political or other influence campaigns, present a clear and present danger to the stability of modern society and national security. Techniques to identify these coordinated groups improve our capabilities to counter such modern information disorders.

8.1 Findings and Contributions

In addressing our first thesis research question **TRQ1** posed in Section 1.1, we evaluated the information environment of social media as a challenging one for research, due to commercial encumbrances placed upon the data by the social media platforms (Part I). The subsequent lack of transparency has implications for trust in the results of social media research. This is because it affects the exchange of datasets for benchmarking, and because the completeness of datasets is never known nor fixed. There are variations in the social media data that researchers can obtain due to at least two factors: the hidden sampling biases of the platforms that provide the data and filtering features in collection tools intended to add value for the user. We systematically demonstrated the extent to which these variations in data cause flow-on variations in social network analyses, potentially impacting interpretations of results and decisions based upon them.

To better understand the context in which CIB is a threat, we explored contentious online discussions (Part II). While directly considering **TRQ2**, this exploration revealed two polarised communities, whose behaviour we characterised through their interactions and their effect on the broader discussion participants. We revealed that the communities' different communication strategies produced different effects, both in the content dissemination patterns and the social networks that evolved. Our longitudinal study addressing **TRQ3** found that the polarised participants re-appeared in other contentious discussions and largely remained polarised in terms of their interactions with other accounts. Their discussion of partisan topics and the fact that apparent political stances did not always align with the polarised groupings, however, suggested there may be opportunities to bridge the gap between communities.

Finally, in addressing **TRQ4** we proposed and systematically demonstrated a novel network-based approach to identifying coordinated amplification of content (Part III), which is generally applicable to many types of common interactions and is designed to facilitate multi-platform investigations. The findings were validated with a variety of computational analysis techniques, augmenting the more commonly used method of manual inspection, which relies on the expertise of domain experts. Additionally, the approach is amenable to near real-time processing, such as those required by the military and national security agencies.

8.2 Future Work

Only Twitter data was considered in our study, but the techniques derived are easily transitioned to data from other platforms, with Gab and Parler being obvious candidates due to both their relevance and similar data models. A more comprehensive relevant dataset including knowledge of offline events that cause corresponding online activity would provide a strong basis for future investigations. Such a dataset could comprise a multi-month collection of relevant Twitter discussions during an Australian federal election, along with details of party policy announcements and debates. Regarding the study of homophily, adding content analyses to cross-community communications will help reveal whether they are friendly or antagonistic. This nuance of sentiment is lost in the homophily measures used in Part II but clearly visible in the aggressive behaviour noted in Chapter 5.

The study of coordination, in particular, provides a wealth of avenues for further exploration and research. From a theoretical point of view, social theory applying to networks derived from online interactions remains understudied compared with real-world studies and requires attention. Importantly, this theory will need to emerge in parallel with findings in real-world data, as researchers continue to focus the immediate need to address information disorders on social media and how they interact with real-world events, including domestic and international disputes. The platforms are also ever-changing, and some have shown a willingness to support researchers to investigate information disorders.[1] As a result, there will always be new information from which to build social networks. Advances in theory will also affect analyses of polarisation: the binary concept of polarisation (i.e., polarised / non-polarised, member / non-member) requires more nuance to better model real communities, yet researchers will remain constrained by the data offered by each of the platforms.

[1] E.g., Twitter recently announced its APIs now provide lists of *all* accounts who like or retweet a tweet in response to researcher requests. Source: https://twittercommunity.com/t/updates-to-retweets-lookup-and-likes-lookup-endpoints/165327. Posted 2022-01-21. Accessed 2022-01-24.

From a practical point of view, the kinds of coordination employed by modern information operations are much broader than simple amplification, and very different approaches will be required to support analysts uncovering such long-term multi-platform activities. This broader concept of coordination requires a willingness to observe and report on information cycles involving not just social media, but also the news media and political realms, to measure the effects of and distinguish between foreign interference, domestic disinformation campaigns and grassroots activism. Regarding amplification, however, the focus of research should now shift to incorporating coordination detection methods into near real-time social media analysis systems, such as RAPID (Lim et al., 2019).

Appendix A

Extra #ArsonEmergency Analysis

A.1 Location Analysis

In exploring the discussion of any contentious regional topic on social media, it is sensible to consider from where contributors come. People from different countries may bring different opinions to the table, and when such discussions may help shape public policy, there is the potential for malign foreign interference. The simplest approach is to consider the 'lang' field in the tweet metadata,[1] which is assigned by Twitter. Across every group and phase, roughly 99% of the tweets had a language code of 'en' (English) or 'und' (undefined). Manual inspection of the largest 'und' proportion (1,007 tweets by Supporters in Phase 3, 19.1% of those tweets) revealed the tweets' content comprised almost entirely of @mentions and hashtags.

Table A.1. The self-reported locations of accounts, categorised by country by hand. Only non-empty locations were used, and only those used multiple times by Unaffiliated accounts were considered (i.e., unique Unaffiliated locations were ignored).

	Opposer		**Supporter**		**Unaffilated**	
Country	Counts	Proportion	Counts	Proportion	Counts	Proportion
Australia	393	88.7%	273	76.9%	3,642	72.0%
USA	4	0.9%	19	5.4%	586	11.6%
UK	4	0.9%	5	1.4%	287	5.7%
Canada	2	0.5%	7	2.0%	146	2.9%
NZ	2	0.5%	5	1.4%	51	1.0%
Miscellaneous	35	7.9%	41	11.5%	143	2.8%
Other	3	0.7%	5	1.4%	204	4.0%
Total	443	100.0%	355	100.0%	5,059	100.0%

To learn more, we examined the 'location' field in the 'user' objects in the tweets. This is a free text field users can populate as they wish and contains a great variety of information, not all of which is accurate, but the majority of populated fields are at least meaningful locations (88%). We manually coded the 'location' for each Supporter and Opposer account and then the 'location' values that appeared more

[1] The 'language', 'utc_offset' and 'timezone' fields within the 'user' field of tweets have been deprecated: https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/user. Accessed 2022-01-29.

than once for the Unaffiliated accounts (Table A.1). The majority of contributors in each group is from Australia, but the Supporters and Unaffiliated accounts included more non-Australian but English-speaking contributions than Opposers. The larger proportion of American and UK contributions in the Unaffiliated accounts may be due to an influx of highly-motivated users who joined the discussion after Graham's analysis (Stilgherrian, 2020) reached the MSM. It is thought that climate change is less settled in those countries.[2] This is borne out by the increased number of unique Unaffiliated accounts in Phase 3.

A.2 Inauthentic Behaviour Patterns

Aggressive and profane language was observed in content posted by both Supporters and Opposers, but our observations includes behaviour that could be regarded as inauthentic (Gleicher, 2018), including trolling. We examined the frequency of hashtags and mentions appearing in tweets by Supporters, Opposers and the remainder of accounts, as well as identifying inflammatory behaviour through manual inspection.

The 288 Supporters and 149 Opposers in the mention network connected to Opposers and Supporters, respectively, slightly more than they mentioned themselves, with 710 edges (E-I index of -0.14). When Unaffiliated accounts are considered (resulting in a mention network of 3,206 nodes and 5,825 edges, a subset of the one shown in Figure 5.8b which omits Unaffiliated—Unaffiliated edges), the combined E-I index for Supporters and Opposers rises to 0.7, suggesting a clear preference to mention Unaffiliated accounts.

An analysis of contemporaneous co-mentions also reveals that Supporter accounts mentioned the same accounts in quick succession much more frequently than Opposers, but that one prominent Opposer account was mentioned by many other accounts (Figure A.1). It is clear the highly mentioned Opposer is a target for accounts, with many pairs of co-mentioners mentioning only the Opposer. A second (Unaffiliated) account is also highly mentioned, lying just below the Opposer account, though it appears mentioned more often by Supporter accounts, while the Opposer is more often mentioned by Unaffiliated accounts. The Opposer account is a prominent left-wing online personality mentioned more than 2400 times in the dataset, while the Unaffiliated account had been suspended by the end of January 2020, just after the collection period, and was mentioned over 350 times in the dataset. The largest Unaffiliated mentioning account (circular green node, on the right of the large connected component) appears to support the arson narrative and also promotes a number of QAnon-related hashtags (The Soufan Center, 2021b).

Tweets that include many hashtags or mentions can stand out in a timeline, because the vast majority of tweets include very few, if any. By including many hashtags, a

[2] https://www.theguardian.com/environment/2019/may/07/us-hotbed-climate-change-denial-international-poll. Posted 2019-05-07. Accessed 2022-01-29.

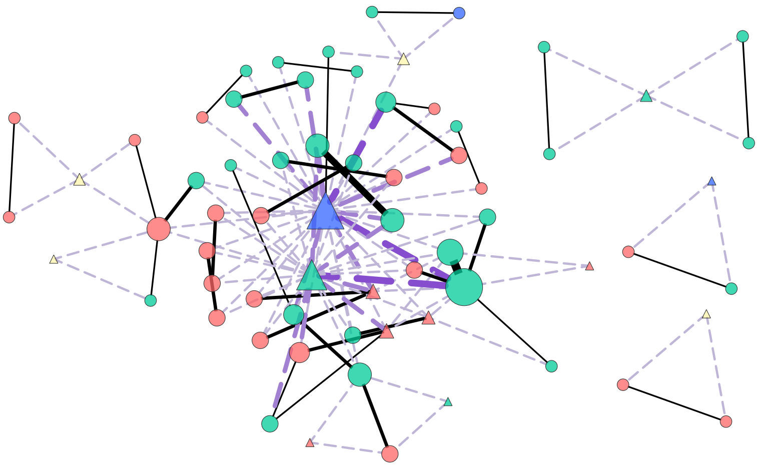


FIGURE A.1. The account/mention 2-layer network resulting from co-mention analysis, connecting accounts with black edges when they mentioned the same account within 60 seconds. Purple edges connect accounts with the accounts they mention, which are shown as triangles. Node colour indicates affiliation: red nodes are Supporters; blue nodes are Opposers; green nodes are Unaffiliated accounts; and yellow nodes are accounts that were mentioned but did not post a tweet in the dataset. Node size indicates the number of tweets they contributed to the corpus or, for mentioned accounts, their degree (reflecting the number of times they were mentioned).

tweet may be seen by anyone searching by those hashtags, thereby increasing its potential audience. Including many mentions may be a way to draw other participants into an ongoing conversation or at least inform them of an opinion or other information. Figure A.2 shows that all groups trended similarly, and that Supporters posted more tweets with many hashtags than Opposers did (although they tweeted nearly twice as often). Unaffiliated accounts used the most hashtags in tweets, with more than 100 Unaffiliated tweets including 19 or more hashtags. Given the great numbers of Unaffiliated accounts and tweets, these can be regarded as outliers (making up less than 1% of their contribution).

Supporters used many more mentions than Opposers more often (Figure A.3). Opposers only used a maximum of 5 mentions on fewer than 10 occasions, while Supporters did the same more than 50 times. In fact, Supporters used more than 5 mentions in 369 tweets. In a few tweets, 45 or more mentions appear, however analysis of this phenomenon has revealed that Twitter accumulates mentions from tweets that have been replied to. One reply tweet including 50 mentions was a simple reply into a reply chain that stretched back to 2018. Many replies in the chain had mentioned one or two other accounts, and they were then incorporated as implicit mentions in any replies to them. Unfortunately, from the point of view of the data provided by the Twitter API, it is unclear whether mentions in a reply are manually added by the respondent or included implicitly, as they simply appear at the start of the tweet text.

Although using many hashtags and mentions may expose inauthentic behaviour, trolling involves broad or direct attacks or simple provocation, and is exposed through

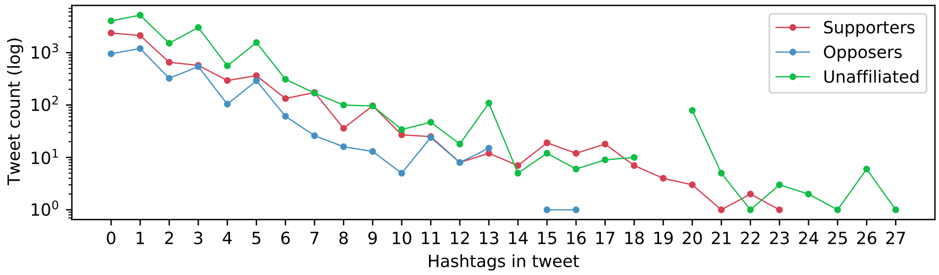


FIGURE A.2. The distribution of hashtag uses amongst all ArsonEmergency tweets.

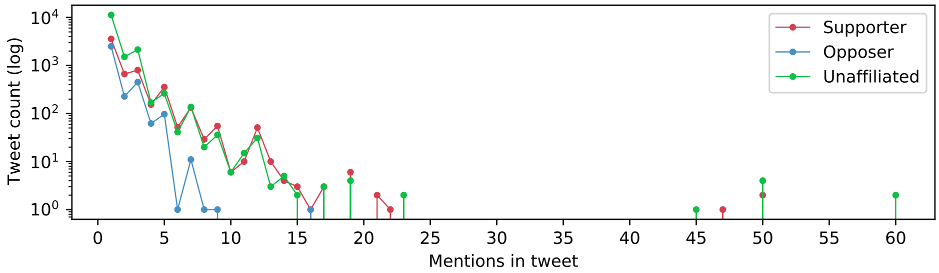


FIGURE A.3. The distribution of mention uses amongst all ArsonEmergency tweets.

use of platform features as well as the content of posts. Patterns of activity that appeared provocative included repetitions of tweets consisting of only:

- one or more hashtags;
- one or more hashtags and a trailing URL;
- one or more mentions with one or more hashtags; and
- one or more mentions with one or more hashtags and a trailing URL.

The frequencies of the occurrence of these text patterns in tweets by each group, in each phase and overall, is shown above in Table 5.10. The majority of these behaviours were present in Phase 3. Although Unaffiliated accounts certainly used some of these patterns, Supporters made much more use of them, particularly more than Opposers (Figure A.4). Many of the instances of hashtags followed by a URL are instances of quote tweets, where the URL is the link to the quoted tweet. These are attempts to disseminate the quoted tweet to a broader audience (engaged through the hashtags).

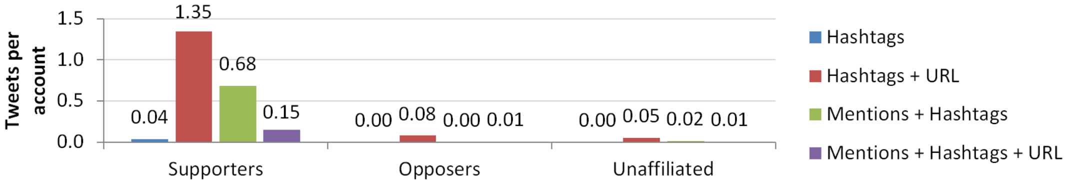


FIGURE A.4. Rates of use of inauthentic tweet text patterns per account for the 497 Supporters, 593 Opposers and 11,782 Unaffiliated accounts over the entire ArsonEmergency dataset.

Finally, inspection of the ten most retweeted tweet contributors revealed that three were Supporters, one was Unaffiliated, and the remainder were Opposers (including five of the top six).

A.2.1 Switching Names

Name switching had been observed in other discussions (Mariconti et al., 2017; Ferrara, 2017), so we examined the accounts for such behaviour. We found only 13 examples, including one Opposer and five Supporters (see Table A.2). Manual inspection of the Unaffiliated, four clearly aligned with the Supporter discussion opinions and themes, based on their content, one was clearly an Opposer, and, of the remaining two, one was raising money for koalas and used hashtags to increase their reach and the other was reporting their research into the number of arson reports (referring to facts more than opinions). The behaviour of the Supporter-aligned accounts used a high proportion of retweets (12 of 18 tweets) though one of them aggressively engaged with other accounts with their six tweets. Some of the changes in screen name appeared to reflect a new 'personality' (*cf.*, Dawson and Innes, 2019), but not in a particularly deceptive way – instead, the changes of name seemed whimsical.

Table A.2. Behaviour of Unaffiliated accounts that changed screen names.

Account	Inclination	Original	Reply	Retweets	Total
u_1	Supporter	2	4	0	6
u_2	Supporter	0	0	4	4
u_3	Supporter	0	0	4	4
u_4	Supporter	0	0	4	4
u_5	Opposer	1	1	0	2
u_6	Unaffiliated	4	0	0	4
u_7	Unaffiliated	2	7	2	11

A.3 Hashtag Use

As expected, the most prominently used hashtag for all communities was `#ArsonEmergency`, however it is clear that there are other commonly occurring hashtags. Table A.3 shows the top ten hashtags used by the Supporters, Opposers and Unaffiliated in each phase, as well as the number of tweets in which they appeared.

In Phase 1, it is clear that the Supporters are trying to engage with existing climate change emergency discussion communities, as well as the media (`#7News`) and broader political discussion (`#auspol`). The few Opposer tweets seem to be poking fun at the discussion (e.g., `#RelevanceDepravationEmergency`, `#PoliticalBSEmergency`), while the Unaffiliated tweets are very broadly about the bushfires, but `#ClimateChangeHoax` is the third most used hashtag.

In the brief Phase 2, Supporters appear to be more concentrated in their promotion of the arson narrative (using `#ClimateCriminals` and `#ecoterrorism`) into the

Table A.3. Top ten hashtags used by the Supporters, Opposers, and Unaffiliated communities in each phase. Hashtags have been compared without considering case in the same way Twitter does. The tag *anon*$_1$ in Phase 3 refers to the same redacted identity in Figure 5.11b.

Phase 1	**Supporters**		**Opposers**		**Unaffiliated**	
	1,573 Tweets		*33 Tweets*		*1,961 Tweets*	
	Hashtag	Count	Hashtag	Count	Hashtag	Count
	arsonemergency	2,086	arsonemergency	43	arsonemergency	2,534
	auspol	574	auspol	9	auspol	1,012
	climatechangehoax	232	bushfires	7	climatechangehoax	682
	climateemergency	230	tresspassemergency	6	climatechange	611
	climatechange	191	lootingemergency	6	australiaburns	307
	7news	126	bandeemergency	6	australiaburning	227
	vicfires	111	theftemergency	5	climateemergency	186
	victoria	107	relevancedepravationemergency	4	australiabushfires	142
	nswfires	90	politicalbsemergency	4	bushfireemergency	133
	globalwarming	84	denialmachine	4	australianfires	78
Phase 2	*121 Tweets*		*327 Tweets*		*759 Tweets*	
	Hashtag	Count	Hashtag	Count	Hashtag	Count
	arsonemergency	142	arsonemergency	487	arsonemergency	1,135
	auspol	79	auspol	36	auspol	194
	bushfiresaustralia	51	climateemergency	11	bushfiresaustralia	110
	climateemergency	26	scottyfrommarketing	9	climateemergency	53
	climatecriminals	23	australianbushfires	9	climatecriminals	34
	climatechange	8	australiaisburning	9	climatechange	23
	victoria	7	dontgetderailed	7	climatechangehoax	18
	ecoterrorism	6	arsonmyarse	7	scottyfrommarketing	16
	australiaisburning	6	stupidemergency	6	australianbushfires	15
	australiaburning	6	australiabushfire	6	astroturfing	15
Phase 3	*5,278 Tweets*		*3,227 Tweets*		*14,267 Tweets*	
	Hashtag	Count	Hashtag	Count	Hashtag	Count
	arsonemergency	7,731	arsonemergency	5,070	arsonemergency	21,194
	auspol	534	australiafires	649	australiafires	2,747
	climateemergency	477	climateemergency	601	climateemergency	2,566
	itsthegreensfault	270	*anon*$_1$	427	*anon*$_1$	1,778
	climatechangehoax	270	bushfires	251	australianbushfiredisaster	1,101
	climatechange	226	auspol	210	auspol	1,011
	climatehoax	220	australianbushfiredisaster	152	climatechangehoax	758
	climatecriminals	177	climatechange	140	australianbushfires	739
	bushfires	176	fakenews	137	climatechange	721
	arsondeniers	169	australianbushfires	101	bushfires	664

`#auspol` political discussion. Opposers seem to focus almost exclusively on using `#ArsonEmergency` rather than any other hashtags, while the Unaffiliated still follow, to some extent, the Supporters' lead with hashtags related to the arson narrative.

Finally, in Phase 3, Supporters focus mostly on just `#ArsonEmergency`, briefly linking to blaming an environmental political party and references to hoaxes, and even reversing the attack and accusing others of being `#ArsonDeniers`. Opposers are firmly focused on `#ArsonEmergency` but start referring to an individual prominent in the media industry commonly seen as advocating against dealing with climate change. By this stage, the Unaffiliated accounts are starting to follow the Opposers' lead discussing emergency- and fire-related hashtags.

Bibliography

Ackland, Robert (Nov. 2020). *Using Semantic Network Analysis to Identify Meaning Structures on Twitter*. Talk presented at the Australian Social Network Analysis Conference, ASNAC '20.

Adjali, Omar, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, and Brigitte Grau (2020). "Multimodal Entity Linking for Tweets". In: *Lecture Notes in Computer Science*. Springer International Publishing, pp. 463–478. DOI: 10.1007/978-3-030-45439-5_31.

Aggarwal, Anupama, Saravana Kumar, Kushagra Bhargava, and Ponnurangam Kumaraguru (Apr. 2018). "The follower count fallacy: Detecting Twitter users with manipulated follower count". In: *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*. SAC '18. ACM, pp. 1748–1755. DOI: 10.1145/3167132.3167318.

Aggarwal, Anupama and Ponnurangam Kumaraguru (July 2015). "What they do in shadows: Twitter underground follower market". In: *13th Annual Conference on Privacy, Security and Trust*. PST '15. IEEE, pp. 93–100. DOI: 10.1109/pst.2015.7232959.

Aiello, Luca Maria, Martina Deplano, Rossano Schifanella, and Giancarlo Ruffo (June 2012). "People are Strange when you're a Stranger: Impact and Influence of Bots on Social Networks". In: *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*. ICWSM '12, pp. 10–17. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4523.

Albada, Katja, Nina Hansen, and Sabine Otten (Apr. 2021). "Polarization in attitudes towards refugees and migrants in the Netherlands". In: *European Journal of Social Psychology* 51.3, pp. 627–643. DOI: 10.1002/ejsp.2766.

Ali, S. Harris and Fuyuki Kurasawa (Mar. 2020). "#COVID19: Social media both a blessing and a curse during coronavirus pandemic". In: *The Conversation*. Accessed 2022-01-31. URL: https://theconversation.com/covid19-social-media-both-a-blessing-and-a-curse-during-coronavirus-pandemic-133596.

Aliapoulios, Max, Emmi Bevensee, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, and Savvas Zannettou (June 2021). "A Large Open Dataset from the Parler Social Network". In: *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media*. ICWSM '21. AAAI Press, pp. 943–951. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/18117.

Alizadeh, Meysam, Jacob N. Shapiro, Cody Buntain, and Joshua A. Tucker (July 2020). "Content-based features predict social media influence operations". In: *Science Advances* 6.30, eabb5824. DOI: 10.1126/sciadv.abb5824.

Amrhein, Valentin, Fränzi Korner-Nievergelt, and Tobias Roth (July 2017). "The earth is flat ($p > 0.05$): significance thresholds and the crisis of unreplicable research". In: *PeerJ* 5, e3544. DOI: 10.7717/peerj.3544.

Angwin, Julia, Madeleine Varner, and Ariana Tobin (Sept. 2017). "Facebook Enabled Advertisers to Reach 'Jew Haters'". In: *ProPublica*. Accessed 2018-05-29. URL: https://www.propublica.org/article/facebook-enabled-advertisers-to-reach-jew-haters.

ASIO (Sept. 2021). *ASIO Annual Report 2020-21*. Annual Report. Australian Security Intelligence Organisation. URL: https://www.asio.gov.au/sites/default/files/AnnualReport2020-21WEB.pdf.

Assenmacher, Dennis, Lena Adam, Heike Trautmann, and Christian Grimme (May 2020). "Towards Real-Time and Unsupervised Campaign Detection in Social Media". In: *Proceedings of the Thirty-Third International Florida Artificial Intelligence Research Society Conference*. FLAIRS '20. AAAI Press, pp. 303–307. URL: https://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS20/paper/view/18452.

Assenmacher, Dennis, Derek Weber, Mike Preuss, André Calero Valdez, Alison Bradshaw, Björn Ross, Stefano Cresci, Heike Trautmann, Frank Neumann, and Christian Grimme (May 2021). "Benchmarking Crisis in Social Media Analytics: A Solution for the Data-Sharing Problem". In: *Social Science Computer Review*, pp. 1–27. DOI: 10.1177/08944393211012268.

Bacco, Caterina De, Eleanor A. Power, Daniel B. Larremore, and Cristopher Moore (Apr. 2017). "Community detection, link prediction, and layer interdependence in multilayer networks". In: *Physical Review E* 95.4, p. 042317. DOI: 10.1103/physreve.95.042317.

Badawy, Adam and Emilio Ferrara (Apr. 2018). "The rise of Jihadist propaganda on social networks". In: *Journal of Computational Social Science* 1.2, pp. 453–470. DOI: 10.1007/s42001-018-0015-z.

Badham, Van (Oct. 2021). "No, Australia is not actually an evil dictatorship". In: *The New York Times*. Accessed 2022-01-30. URL: https://www.nytimes.com/2021/10/14/opinion/australia-far-right-america.html.

Bagavathi, Arunkumar, Pedram Bashiri, Shannon Reid, Matthew Phillips, and Siddharth Krishnan (Aug. 2019). "Examining Untempered Social Media: Analyzing Cascades of Polarized Conversations". In: *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ASONAM'19. ACM, pp. 625–632. DOI: 10.1145/3341161.3343695.

Bagrow, James P., Xipei Liu, and Lewis Mitchell (Jan. 2019). "Information flow reveals prediction limits in online social activity". In: *Nature Human Behaviour* 3.2, pp. 122–128. DOI: 10.1038/s41562-018-0510-5.

Bail, Christopher A., Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and

Alexander Volfovsky (Aug. 2018). “Exposure to opposing views on social media can increase political polarization”. In: *Proceedings of the National Academy of Sciences* 115.37, pp. 9216–9221. DOI: 10.1073/pnas.1804840115.

Baker, Monya (May 2016). “1,500 scientists lift the lid on reproducibility”. In: *Nature* 533.7604, pp. 452–454. DOI: 10.1038/533452a.

Ball, Philip and Amy Maxmen (May 2020). “The epic battle against coronavirus misinformation and conspiracy theories”. In: *Nature* 581.7809, pp. 371–374. DOI: 10.1038/d41586-020-01452-z.

Barberá, Pablo, John T. Jost, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau (Aug. 2015). “Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber?” In: *Psychological Science* 26.10, pp. 1531–1542. DOI: 10.1177/0956797615594620.

Baronchelli, Andrea (Feb. 2018). “The emergence of consensus: a primer”. In: *Royal Society Open Science* 5.2. DOI: 10.1098/rsos.172189.

Barry, Paul (Feb. 2020). *Media Watch: News Corps Fire Fight.* Australian Broadcasting Corporation. Broadcast 2020-02-03. URL: https://iview.abc.net.au/show/media-watch/series/0/video/FA1935H001S00.

Baumann, Fabian, Philipp Lorenz-Spreen, Igor M. Sokolov, and Michele Starnini (Jan. 2021). “Emergence of Polarized Ideological Opinions in Multidimensional Topic Spaces”. In: *Physical Review X* 11.1. DOI: 10.1103/physrevx.11.011012.

Bedru, Hayat Dino, Shuo Yu, Xinru Xiao, Da Zhang, Liangtian Wan, He Guo, and Feng Xia (Aug. 2020). “Big networks: A survey”. In: *Computer Science Review* 37, p. 100247. DOI: 10.1016/j.cosrev.2020.100247.

Bellutta, Daniele, Catherine King, and Kathleen M. Carley (Mar. 2021). “Deceptive accusations and concealed identities as misinformation campaign strategies”. In: *Computational and Mathematical Organization Theory* 27.3, pp. 302–323. DOI: 10.1007/s10588-021-09328-x.

Benkler, Yochai, Robert Farris, and Hal Roberts (Oct. 2018). *Network Propaganda.* Oxford University Press. DOI: 10.1093/oso/9780190923624.001.0001.

Berger, Jonathan M. (June 2014). “How ISIS Games Twitter”. In: *The Atlantic*. Accessed 2021-12-03. URL: https://www.theatlantic.com/international/archive/2014/06/isis-iraq-twitter-social-media-strategy/372856/.

Berger, Jonathon M. and Jonathon Morgan (Mar. 2015). *The ISIS Twitter Census: Defining and describing the population of ISIS supporters on Twitter*. Analysis Paper 20. The Brookings Institution. URL: https://www.brookings.edu/research/the-isis-twitter-census-defining-and-describing-the-population-of-isis-supporters-on-twitter/.

Bergmann, Eiríkur (July 2020). “Populism and the politics of misinformation”. In: *Safundi: The Journal of South African and American Studies* 21.3, pp. 251–265. DOI: 10.1080/17533171.2020.1783086.

Beskow, David M. and Kathleen M. Carley (Aug. 2018a). “Bot Conversations are Different: Leveraging Network Metrics for Bot Detection in Twitter”. In: *Proceedings* *of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ASONAM '18. IEEE. DOI: 10.1109/asonam.2018.8508322.

Beskow, David M. and Kathleen M. Carley (July 2018b). "Bot-hunter: a tiered approach to detecting & characterizing automated activity on Twitter". In: *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Vol. 3. SBP-BRiMS '18, p. 8. URL: http://www.casos.cs.cmu.edu/publications/papers/LB_5.pdf.

Bessi, Alessandro and Emilio Ferrara (2016). "Social bots distort the 2016 U.S. Presidential election online discussion". In: *First Monday* 21.11. DOI: 10.5210/fm.v21i11.7090.

Bittman, Ladislav (1985). *The KGB and Soviet disinformation: an insider's view*. Washington: Pergamon-Brassey's. ISBN: 9780080315720.

Blondel, Vincent D, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre (Oct. 2008). "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10, P10008. DOI: 10.1088/1742-5468/2008/10/p10008.

Borgatti, Stephen P., Martin G. Everett, and Jeffrey C. Johnson (May 2013). *Analyzing Social Networks*. 1st Edition. SAGE PUBN. 296 pp. ISBN: 1446247414.

Borgatti, Stephen P, Ajay Mehra, Daniel J Brass, and Giuseppe Labianca (2009). "Network analysis in the social sciences". In: *Science* 323.5916, pp. 892–895. DOI: 10.1126/science.1165821.

Bot Sentinel (Oct. 2021). *Twitter hate accounts targeting Meghan and Harry, Duke and Duchess of Sussex*. Report. Bot Sentinel Inc. URL: https://botsentinel.com/reports/documents/duke-and-duchess-of-sussex/report-10-26-2021.pdf.

Boutyline, Andrei and Robb Willer (May 2016). "The Social Structure of Political Echo Chambers: Variation in Ideological Homophily in Online Networks". In: *Political Psychology* 38.3, pp. 551–569. DOI: 10.1111/pops.12337.

boyd, danah (Jan. 2017). "Hacking the Attention Economy". In: *Data & Society: Points*. Accessed 2022-01-30. URL: https://points.datasociety.net/hacking-the-attention-economy-9fa1daca7a37.

Bradshaw, Samantha, Hannah Bailey, and Philip N. Howard (Jan. 2021). *Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation*. Working Paper 2021.1. Programme on Democracy & Technology, Oxford, UK. URL: https://demtech.oii.ox.ac.uk/research/posts/industrialized-disinformation/.

Brandes, Ulrik and Thomas Erlebach, eds. (Feb. 2005). *Network Analysis: Methodological Foundations*. Springer-Verlag GmbH. ISBN: 978-3-540-31955-9.

Brandes, Ulrik, Marco Gaertler, and Dorothea Wagner (June 2008). "Engineering graph clustering: Models and experimental evaluation". In: *ACM Journal of Experimental Algorithmics* 12.1.1, pp. 1–26. DOI: 10.1145/1227161.1227162.

Brazil, Rachel (July 2020). "Fighting flat-Earth theory". In: *Physics World*. Accessed 2022-01-30. URL: https://physicsworld.com/a/fighting-flat-earth-theory/.

Breck, Eric, Neoklis Polyzotis, Sudip Roy, Steven Euijong Whang, and Martin Zinkevich (Mar. 2019). "Data Validation for Machine Learning". In: *Proceedings of Machine Learning and Systems*. Vol. 1. MLSys '19, pp. 334–347. URL: https://proceedings.mlsys.org/paper/2019/file/5878a7ab84fb43402106c575658472fa-Paper.pdf.

Brin, Sergey and Lawrence Page (Apr. 1998). "The Anatomy of a Large-Scale Hypertextual Web Search Engine". In: *Computer Networks and ISDN Systems* 30.1-7, pp. 107–117. DOI: 10.1016/s0169-7552(98)00110-x.

Broniatowski, David A. (May 2021). *Towards Statistical Foundations For Detecting Coordinated Inauthentic Behavior On Facebook*. IDDP Report pre-print. Institute for Data, Democracy and Politics, The George Washington University. URL: https://iddp.gwu.edu/towards-statistical-foundations-detecting-coordinated-inauthentic-behavior-facebook.

Broniatowski, David A., Amelia M. Jamison, SiHua Qi, Lulwah AlKulaib, Tao Chen, Adrian Benton, Sandra C. Quinn, and Mark Dredze (Oct. 2018). "Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate". In: *American Journal of Public Health* 108.10, pp. 1378–1384. DOI: 10.2105/ajph.2018.304567.

Brooking, Emerson T. and P. W. Singer (Nov. 2016). "War Goes Viral: How social media is being weaponized across the world". In: *The Atlantic*. Accessed 2022-01-30. URL: https://www.theatlantic.com/magazine/archive/2016/11/war-goes-viral/501125/.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (May 2020). "Language Models are Few-Shot Learners". In: *arXiv preprint*. arXiv: 2005.14165 [cs.CL].

Bruns, Axel (July 2019a). "After the 'APIcalypse': social media platforms and their fight against critical scholarly research". In: *Information, Communication & Society* 22.11, pp. 1544–1566. DOI: 10.1080/1369118x.2019.1637447.

Bruns, Axel (Sept. 2019b). *Are Filter Bubbles Real?* Polity Press. ISBN: 1509536442.

Bruns, Axel (Nov. 2019c). "Filter bubble". In: *Internet Policy Review* 8.4. DOI: 10.14763/2019.4.1426.

Bruns, Axel and Jean Burgess (Jan. 2012). *#qldfloods and @QPSMedia: Crisis Communication on Twitter in the 2011 South East Queensland Floods*. Research Report 48241. ARC Centre of Excellence for Creative Industries and Innovation. URL: https://eprints.qut.edu.au/48241/.

Bruns, Axel, Stephen Harrington, and Edward Hurcombe (Aug. 2020). "'Corona? 5G? or both?': the dynamics of COVID-19/5G conspiracy theories on Facebook". In: *Media International Australia* 177.1, pp. 12–29. DOI: 10.1177/1329878x20946113.

Bruns, Axel, Stephen Harrington, and Edward Hurcombe (2021). "Coronavirus Conspiracy Theories: Tracing Misinformation Trajectories from the Fringes to the Mainstream". In: *Communicating COVID-19*. Springer International Publishing, pp. 229–249. DOI: 10.1007/978-3-030-79735-5_12.

Bruns, Axel and Yuxian Eugene Liang (Apr. 2012). "Tools and methods for capturing Twitter data during natural disasters". In: *First Monday* 17.4. DOI: 10.5210/fm.v17i4.3937.

Burgess, Jean and Ariadna Matamoros-Fernández (Apr. 2016). "Mapping sociocultural controversies across digital media platforms: one week of #gamergate on Twitter, YouTube, and Tumblr". In: *Communication Research and Practice* 2.1, pp. 79–96. DOI: 10.1080/22041451.2016.1155338.

Cao, Cheng, James Caverlee, Kyumin Lee, Hancheng Ge, and Jinwook Chung (Oct. 2015). "Organic or Organized?: Exploring URL Sharing Behavior". In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. CIKM '15. ACM, pp. 513–522. DOI: 10.1145/2806416.2806572.

Cao, Qiang, Xiaowei Yang, Jieqi Yu, and Christopher Palow (Nov. 2014). "Uncovering Large Groups of Active Malicious Accounts in Online Social Networks". In: *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. CCS '14. ACM. DOI: 10.1145/2660267.2660269.

Carley, Kathleen M. (Nov. 2020). "Social cybersecurity: an emerging science". In: *Computational and Mathematical Organization Theory* 26.4, pp. 365–381. DOI: 10.1007/s10588-020-09322-9.

Carnein, Matthias, Dennis Assenmacher, and Heike Trautmann (Nov. 2017). "Stream Clustering of Chat Messages with Applications to Twitch Streams". In: *Lecture Notes in Computer Science*. Vol. 10651. Springer International Publishing, pp. 79–88. DOI: 10.1007/978-3-319-70625-2_8.

Carvin, Andy (2012). *Distant Witness: Social media, the Arab Spring and a journalism revolution*. CUNY Journalism Press. ISBN: 9781939293022.

Chavoshi, Nikan, Hossein Hamooni, and Abdullah Mueen (2017). "Temporal Patterns in Bot Activities". In: *Proceedings of the 26th International Conference on World Wide Web Companion*. WWW '17. ACM Press, pp. 1601–1606. DOI: 10.1145/3041021.3051114.

Chen, Adrian (June 2015). "The Agency". In: *The New York Times Magazine*. Accessed 2022-01-30. URL: https://www.nytimes.com/2015/06/07/magazine/the-agency.html.

Cheng, Justin, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec (Feb. 2017). "Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions". In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. CSCW '17. ACM. DOI: 10.1145/2998181.2998213.

Chessen, Matt (Sept. 2017). *The MADCOM Future*. Research Report. The Atlantic Council. URL: http://www.jstor.org/stable/resrep03728.

Chirwa, Candice and Zimkhitha Manyana (Sept. 2021). “The rise of fake news: Surveying the effects of social media on informed democracy”. In: *The Thinker* 88.3. Accessed 2021-11-23. URL: https://journals.uj.ac.za/index.php/The_Thinker/article/view/604.

Cho, Charles H., Martin L. Martens, Hakkyun Kim, and Michelle Rodrigue (July 2011). “Astroturfing Global Warming: It Isn’t Always Greener on the Other Side of the Fence”. In: *Journal of Business Ethics* 104.4, pp. 571–587. DOI: 10.1007/s10551-011-0950-6.

Chu, Zi, Indra Widjaja, and Haining Wang (2012). “Detecting Social Spam Campaigns on Twitter”. In: *Applied Cryptography and Network Security*. Springer Berlin Heidelberg, pp. 455–472. DOI: 10.1007/978-3-642-31284-7_27.

Cialdini, Robert (2007). *Influence: the Psychology of Persuasion*. Collins Business. ISBN: 9780061241895.

Ciampaglia, Giovanni Luca, Azadeh Nematzadeh, Filippo Menczer, and Alessandro Flammini (Oct. 2018). “How algorithmic popularity bias hinders or promotes quality”. In: *Scientific Reports* 8.1. DOI: 10.1038/s41598-018-34203-2.

Cockburn, Andy, Pierre Dragicevic, Lonni Besançon, and Carl Gutwin (July 2020). “Threats of a Replication Crisis in Empirical Computer Science”. In: *Communications of the ACM* 63.8, pp. 70–79. DOI: 10.1145/3360311.

Confessore, Nicholas, Gabriel J. X. Dance, Richard Harris, and Mark Hansen (Jan. 2018). “The Follower Factory”. In: *The New York Times*. Accessed 2022-01-30. URL: https://www.nytimes.com/interactive/2018/01/27/technology/social-media-bots.html.

Conover, Michael, Jacob Ratkiewicz, Matthew R Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini (July 2011). “Political polarization on Twitter”. In: *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. ICWSM ’11. The AAAI Press, pp. 89–96. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2847.

Cresci, Stefano (Sept. 2020). “A decade of social bot detection”. In: *Communications of the ACM* 63.10, pp. 72–83. DOI: 10.1145/3409116.

Cresci, Stefano, Fabrizio Lillo, Daniele Regoli, Serena Tardelli, and Maurizio Tesconi (Apr. 2019). “Cashtag Piggybacking: Uncovering Spam and Bot Activity in Stock Microblogs on Twitter”. In: *ACM Transactions on the Web* 13.2, pp. 1–27. DOI: 10.1145/3313184.

Cresci, Stefano, Marinella Petrocchi, Angelo Spognardi, and Stefano Tognazzi (Nov. 2021). “Adversarial machine learning for protecting against online manipulation”. In: *arXiv preprint*. arXiv: 2111.12034 [cs.LG].

Cresci, Stefano, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi (Sept. 2016). “DNA-Inspired Online Behavioral Modeling and Its Application to Spambot Detection”. In: *IEEE Intelligent Systems* 31.5, pp. 58–64. DOI: 10.1109/mis.2016.29.

Cresci, Stefano, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi (July 2017a). "Social fingerprinting: detection of spambot groups through DNA-inspired behavioral modeling". In: *IEEE Transactions on Dependable and Secure Computing* 15.4, pp. 561–576. DOI: 10.1109/tdsc.2017.2681672.

Cresci, Stefano, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi (2017b). "The Paradigm-Shift of Social Spambots: Evidence, Theories, and Tools for the Arms Race". In: *Proceedings of the 26th International Conference on World Wide Web Companion*. WWW '17. ACM Press, pp. 963–972. DOI: 10.1145/3041021.3055135.

CREST (Dec. 2017). *Russian interference and influence measures following the 2017 UK terrorist attacks*. Policy Brief 17-081-02. Accessed 2022-01-30. Centre for Research, Evidence on Security Threats, Cardiff University Crime, and Security Research Institute. URL: https://crestresearch.ac.uk/resources/russian-influence-uk-terrorist-attacks/.

Damashek, M. (1995). "Gauging Similarity with n-Grams: Language-Independent Categorization of Text". In: *Science* 267.5199, pp. 843–848. DOI: 10.1126/science.267.5199.843.

Dancey, Christine P and John Reidy (2011). *Statistics without maths for psychology*. 5th. Prentice Hall/Pearson. ISBN: 9780273726029.

Datta, Srayan and Eytan Adar (June 2019). "Extracting Inter-Community Conflicts in Reddit". In: *Proceedings of the Thirteenth International AAAI Conference on Web and Social Media*. ICWSM '19. AAAI Press, pp. 146–157. URL: https://aaai.org/ojs/index.php/ICWSM/article/view/3217.

Davis, Clayton Allen, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer (2016). "BotOrNot: A System to Evaluate Social Bots". In: *Proceedings of the 25th International Conference on World Wide Web Companion*. WWW '16. ACM Press, pp. 273–274. DOI: 10.1145/2872518.2889302.

Dawkins, Richard (1989). *The Selfish Gene*. Oxford University Press. ISBN: 0192860925.

Dawson, Andrew and Martin Innes (May 2019). "How Russia's Internet Research Agency Built its Disinformation Campaign". In: *The Political Quarterly* 90.2, pp. 245–256. DOI: 10.1111/1467-923x.12690.

DeGroot, Morris H. (Mar. 1974). "Reaching a Consensus". In: *Journal of the American Statistical Association* 69.345, pp. 118–121. DOI: 10.1080/01621459.1974.10480137.

Demszky, Dorottya, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Dan Jurafsky (June 2019). "Analyzing Polarization in Social Media: Method and Application to Tweets on 21 Mass Shootings". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, pp. 2970–3005. DOI: 10.18653/v1/n19-1304.

Deusser, Clemens, Nora Jansen, Jan Reubold, Benjamin Schiller, Oliver Hinz, and Thorsten Strufe (Apr. 2018). “Buzz in Social Media”. In: *Companion Proceedings of the The Web Conference 2018*. WWW ’18. ACM Press. DOI: 10.1145/3184558.3191591.

DFAT (Nov. 2017). *2017 Foreign Policy White Paper*. Department of Foreign Affairs and Trade. ISBN: 9781743224113. URL: https://www.dfat.gov.au/publications/minisite/2017-foreign-policy-white-paper/fpwhitepaper/foreign-policy-white-paper.html.

DiResta, Renée (Oct. 2021). “It’s not misinformation. It’s amplified propaganda.” In: *The Atlantic*. Accessed 2022-01-30. URL: https://www.theatlantic.com/ideas/archive/2021/10/disinformation-propaganda-amplification-ampliganda/620334/.

DNI (Apr. 2021). *Annual Threat Assessment of the US Intelligence Community*. Annual Threat Assessment. Office of the Director of National Intelligence. URL: https://www.dni.gov/files/ODNI/documents/assessments/ATA-2021-Unclassified-Report.pdf.

Dodds, Leigh (May 2017). “Can you publish tweets as open data?” In: *Lost Boy: The blog of @ldodds*. Accessed 2021-12-08. URL: https://blog.ldodds.com/2017/05/19/can-you-publish-tweets-as-open-data/.

Domingos, Pedro (Oct. 2012). “A few useful things to know about machine learning”. In: *Communications of the ACM* 55.10, pp. 78–87. DOI: 10.1145/2347736.2347755.

Douek, Evelyn (July 2020). “What Does “Coordinated Inauthentic Behaviour” Actually Mean?” In: *Slate*. Accessed 2021-11-24. URL: https://slate.com/technology/2020/07/coordinated-inauthentic-behavior-facebook-twitter.html.

Drenten, Jenna and Lauren Gurrieri (Sept. 2018). “Crossing the #BikiniBridge: Exploring the Role of Social Media in Propagating Body Image Trends”. In: Routledge. Chap. 4, pp. 49–70. ISBN: 9781138052567.

Dunn, Adam G, Julie Leask, Xujuan Zhou, Kenneth D Mandl, and Enrico Coiera (June 2015). “Associations Between Exposure to and Expression of Negative Opinions About Human Papillomavirus Vaccines on Social Media: An Observational Study”. In: *Journal of medical Internet research* 17.6, e144. DOI: 10.2196/jmir.4343.

Echeverria, Juan and Shi Zhou (July 2017). “Discovery, Retrieval, and Analysis of ’Star Wars’ botnet in Twitter”. In: *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ASONAM ’17. ACM, pp. 1–8. DOI: 10.1145/3110025.3110074.

Edwards, Chad, Autumn Edwards, Patric R Spence, and Ashleigh K Shelton (Apr. 2014). “Is that a bot running the social media feed? Testing the differences in perceptions of communication quality for a human agent and a bot agent on Twitter”. In: *Computers in Human Behavior* 33, pp. 372–376. DOI: 10.1016/j.chb.2013.08.013.

Edwards, Michelle, Jonathan Tuke, Matthew Roughan, and Lewis Mitchell (Dec. 2020). “The one comparing narrative social network extraction techniques”. In: *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and*

Mining. ASONAM '20. IEEE, pp. 905–913. DOI: 10.1109/asonam49781.2020.9381346.

Emani, Cheikh Kacfah, Nadine Cullot, and Christophe Nicolle (Aug. 2015). "Understandable Big Data: A survey". In: *Computer Science Review* 17, pp. 70–81. DOI: 10.1016/j.cosrev.2015.05.002.

Eswaran, Dhivya, Christos Faloutsos, Sudipto Guha, and Nina Mishra (July 2018). "SpotLight: Detecting Anomalies in Streaming Graphs". In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '18. ACM, pp. 1378–1386. DOI: 10.1145/3219819.3220040.

Fair, Gabriel and Ryan Wesslen (June 2019). "Shouting into the Void: A Database of the Alternative Social Media Platform Gab". In: *Proceedings of the Thirteenth International AAAI Conference on Web and Social Media*. ICWSM '19. AAAI Press, pp. 608–610. URL: https://aaai.org/ojs/index.php/ICWSM/article/view/3258.

Falzon, Lucia, Caitlin McCurrie, and John Dunn (July 2017). "Representation and Analysis of Twitter Activity: A Dynamic Network Perspective". In: *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ASONAM '17. ACM, pp. 1183–1190. DOI: 10.1145/3110025.3122118.

Falzon, Lucia, Eric Quintane, John Dunn, and Garry Robins (Mar. 2018). "Embedding time in positions: Temporal measures of centrality for social network analysis". In: *Social Networks* 54, pp. 168–178. DOI: 10.1016/j.socnet.2018.02.002.

Fang, Yixiang, Xin Huang, Lu Qin, Ying Zhang, Wenjie Zhang, Reynold Cheng, and Xuemin Lin (July 2019). "A survey of community search over big graphs". In: *The VLDB Journal* 29.1, pp. 353–392. DOI: 10.1007/s00778-019-00556-x.

Fazil, Mohd and Muhammad Abulaish (Mar. 2020). "A socialbots analysis-driven graph-based approach for identifying coordinated campaigns in Twitter". In: *Journal of Intelligent & Fuzzy Systems* 38.3, pp. 2961–2977. DOI: 10.3233/JIFS-182895.

Feld, Scott L. (Mar. 1981). "The Focused Organization of Social Ties". In: *American Journal of Sociology* 86.5, pp. 1015–1035. DOI: 10.1086/227352.

Feng, Shangbin, Herun Wan, Ningnan Wang, Jundong Li, and Minnan Luo (Oct. 2021). "TwiBot-20: A Comprehensive Twitter Bot Detection Benchmark". In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. CIKM '21. ACM, pp. 4485–4494. DOI: 10.1145/3459637.3482019.

Ferrara, E. (Aug. 2017). "Disinformation and social bot operations in the run up to the 2017 French presidential election". In: *First Monday* 22.8. DOI: 10.5210/fm.v22i8.8005.

Ferrara, Emilio, Stefano Cresci, and Luca Luceri (Nov. 2020). "Misinformation, manipulation, and abuse on social media in the era of COVID-19". In: *Journal of Computational Social Science* 3.2, pp. 271–277. DOI: 10.1007/s42001-020-00094-5.

Ferrara, Emilio, Onur Varol, Clayton A. Davis, Filippo Menczer, and Alessandro Flammini (2016). "The rise of social bots". In: *Communications of the ACM* 59.7, pp. 96–104. DOI: 10.1145/2818717.

Fisher, Ali (Oct. 2018). *Netwar in Cyberia: Decoding the Media Mujahadin.* CPD Perspectives Paper 5. USC Center on Public Diplomacy. URL: https://uscpublicdiplomacy.org/sites/uscpublicdiplomacy.org/files/NetwarinCyberiaWebReady_withdisclosurepage11.08.18.pdf.

Flew, Terry, Axel Bruns, Jean Burgess, Kate Crawford, and Frances Shaw (Nov. 2014). "Social media and its impact on crisis communication: Case studies of Twitter use in emergency management in Australia and New Zealand". In: *Communication and Social Transformation.* International Communication Association. URL: https://eprints.qut.edu.au/63707/.

Foidl, Harald and Michael Felderer (Aug. 2019). "Risk-based data validation in machine learning-based software systems". In: *Proceedings of the 3rd ACM SIGSOFT International Workshop on Machine Learning Techniques for Software Quality Evaluation.* MaLTeSQuE '19. ACM Press, pp. 13–18. DOI: 10.1145/3340482.3342743.

Freelon, Deen (June 2019). *Post-API Social Media Research: A Trip to the Twilight Zone.* Keynote Address at the 13th International AAAI Conference on Web and Social Media, ICWSM '19.

Freeman, David Mandell (Apr. 2017). "Can You Spot the Fakes?: On the Limitations of User Feedback in Online Social Networks". In: *Proceedings of the 26th International Conference on World Wide Web.* WWW '17. ACM, pp. 1093–1102. DOI: 10.1145/3038912.3052706.

Freitas, Carlos, Fabricio Benevenuto, Saptarshi Ghosh, and Adriano Veloso (Aug. 2015). "Reverse Engineering Socialbot Infiltration Strategies in Twitter". In: *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.* ASONAM '15. ACM. DOI: 10.1145/2808797.2809292.

Garfinkel, Harold (Nov. 2005). *Seeing Sociologically: The Routine Grounds of Social Action.* 1st Edition. Paradigm Publishers. ISBN: 9781594510939.

Garimella, Kiran, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis (June 2017). "The Effect of Collective Attention on Controversial Debates on Social Media". In: *Proceedings of the 2017 ACM Web Science Conference.* WebSci '17. ACM, pp. 43–52. DOI: 10.1145/3091478.3091486.

Garimella, Kiran, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis (Apr. 2018a). "Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship". In: *Proceedings of the 2018 World Wide Web Conference.* WWW '18. ACM Press, pp. 913–922. DOI: 10.1145/3178876.3186139.

Garimella, Kiran and Ingmar Weber (May 2017). "A Long-Term Analysis of Polarization on Twitter". In: *Proceedings of the Eleventh International Conference on Web and Social Media.* ICWSM '17. AAAI Press, pp. 528–531. URL: https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15592.

Garimella, Venkata Rama Kiran, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis (Apr. 2018b). "Polarization on Social Media". In: *Proceedings of the 2018 World Wide Web Conference Tutorials Track*. WWW '18. ACM. URL: https://www2018.thewebconf.org/program/tutorials-track/tutorial-202/.

Giglietto, Fabio, Nicola Righetti, and Giada Marino (Sept. 2019). "Understanding Coordinated and Inauthentic Link Sharing Behavior on Facebook in the Run-up to 2018 General Election and 2019 European Election in Italy". In: *SocArXiv*. DOI: 10.31235/osf.io/3jteh.

Giglietto, Fabio, Nicola Righetti, Luca Rossi, and Giada Marino (June 2020a). "Coordinated Link Sharing Behavior as a Signal to Surface Sources of Problematic Information on Facebook". In: *International Conference on Social Media and Society*. ACM, pp. 85–91. DOI: 10.1145/3400806.3400817.

Giglietto, Fabio, Nicola Righetti, Luca Rossi, and Giada Marino (Mar. 2020b). "It takes a village to manipulate the media: Coordinated link sharing behavior during 2018 and 2019 Italian elections". In: *Information, Communication & Society*, pp. 1–25. DOI: 10.1080/1369118x.2020.1739732.

Giles, Jim (Dec. 2005). "Internet encyclopaedias go head to head". In: *Nature* 438.7070, pp. 900–901. DOI: 10.1038/438900a.

Gillmor, Dan (Feb. 2006). *We the Media: Grassroots Journalism by the People, for the People*. O'Reilly Media. ISBN: 0596102275.

Gleicher, Nathaniel (Dec. 2018). "Coordinated Inauthentic Behaviour Explained". In: *Meta*. Retrieved 2022-01-21. URL: https://about.fb.com/news/2018/12/inside-feed-coordinated-inauthentic-behavior/.

Gonzalez-Bailon, Sandra, Andreas Kaltenbrunner, and Rafael E Banchs (June 2010). "The Structure of Political Discussion Networks: A Model for the Analysis of Online Deliberation". In: *Journal of Information Technology* 25.2, pp. 230–243. DOI: 10.1057/jit.2010.2.

González-Bailón, Sandra, Ning Wang, Alejandro Rivero, Javier Borge-Holthoefer, and Yamir Moreno (2014). "Assessing the bias in samples of large online networks". In: *Social Networks* 38, pp. 16–27. DOI: 10.1016/j.socnet.2014.01.004.

González-Cabañas, José, Ángel Cuevas, Rubén Cuevas, Juan López-Fernández, and David García (Nov. 2021). "Unique on Facebook: Formulation and Evidence of (Nano)targeting Individual Users with non-PII Data". In: *ACM Internet Measurement Conference*. IMC '21. ACM, pp. 464–479. DOI: 10.1145/3487552.3487861.

Goodman, Leo A. (Mar. 1961). "Snowball Sampling". In: *The Annals of Mathematical Statistics* 32.1, pp. 148–170. DOI: 10.1214/aoms/1177705148.

Gorwa, Robert and Douglas Guilbeault (July 2017). "Tinder nightmares: the promise and peril of political bots". In: *Wired*. Accessed 2022-01-31. URL: https://www.wired.co.uk/article/tinder-political-bots-jeremy-corbyn-labour.

Graham, Jesse, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto (2013). "Moral foundations theory: The pragmatic validity of moral pluralism". In: *Advances in experimental social psychology.* Vol. 47. Elsevier, pp. 55–130. DOI: 10.1016/B978-0-12-407236-7.00002-4.

Graham, Tim (June 2020). *Detecting and Analysing Coordinated Inauthentic Behaviour on Social Media.* Online. QUT: Digital Media Research Centre presentation. URL: https://research.qut.edu.au/qutcds/wp-content/uploads/sites/257/2020/06/Coordinated_behaviour_seminar.pdf.

Graham, Tim, Robert Ackland, and Lewis Mitchell (Nov. 2020a). *A novel network-based approach to detecting and analysing coordinated inauthentic behaviour on Twitter.* Talk presented at the Australian Social Network Analysis Conference, ASNAC '20. URL: https://pypi.org/project/coordination-network-toolkit/.

Graham, Tim, Marian-Andrei Rizoiu, Axel Bruns, and Dan Angus (Sept. 2021). *Discovering the Strategies and Promotion Schedules of Coordinated Disinformation via Hawkes Intensity Processes.* Talk presented at the European Communication Conference. URL: https://www.behavioral-ds.science/theme2_content/coordinated_disinfo_hawkes/.

Graham, Timothy and Robert Ackland (2017). "Do Socialbots Dream of Popping the Filter Bubble? The role of socialbots in promoting participatory democracy in social media". In: *Socialbots and Their Friends: Digital Media and the Automation of Sociality.* Routledge. Chap. 10, pp. 187–206. ISBN: 9781138639409.

Graham, Timothy, Axel Bruns, Daniel Angus, Edward Hurcombe, and Sam Hames (Dec. 2020b). "#IStandWithDan versus #DictatorDan: the polarised dynamics of Twitter discussions about Victoria's COVID-19 restrictions". In: *Media International Australia* 179.1, pp. 127–148. DOI: 10.1177/1329878x20981780.

Graham, Timothy, Axel Bruns, Guangnan Zhu, and Rod Campbell (June 2020c). *Like a virus: the coordinated spread of coronavirus misinformation.* Report. Centre for Responsible Technology, The Australia Institute. URL: https://apo.org.au/node/305864.

Graham, Timothy and Tobias R. Keller (Jan. 2020). "Bushfires, bots and arson claims: Australia flung in the global disinformation spotlight". In: *The Conversation.* Accessed 2020-02-07. URL: https://theconversation.com/bushfires-bots-and-arson-claims-australia-flung-in-the-global-disinformation-spotlight-129556.

Granovetter, Mark (Nov. 1985). "Economic Action and Social Structure: The Problem of Embeddedness". In: *American Journal of Sociology* 91.3, pp. 481–510. DOI: 10.1086/228311.

Grassegger, Hannes and Mikael Krogerus (Jan. 2017). "The Data That Turned the World Upside Down". In: *Vice – Motherboard.* Accessed 2022-01-31. URL: https://motherboard.vice.com/en_us/article/mg9vvn/how-our-likes-helped-trump-win.

Gray, Caitlin, Lewis Mitchell, and Matthew Roughan (2020). "Bayesian inference of network structure from information cascades". In: *IEEE Transactions on Signal and Information Processing over Networks* 6, pp. 371–381. DOI: 10.1109/TSIPN.2020.2990276.

Grimme, Christian, Dennis Assenmacher, and Lena Adam (July 2018). "Changing Perspectives: Is It Sufficient to Detect Social Bots?" In: *Lecture Notes in Computer Science*. Vol. 10913. Springer International Publishing, pp. 445–461. DOI: 10.1007/978-3-319-91521-0_32.

Grimme, Christian, Mike Preuss, Lena Adam, and Heike Trautmann (Dec. 2017). "Social Bots: Human-Like by Means of Human Control?" In: *Big Data* 5.4, pp. 279–293. DOI: 10.1089/big.2017.0044.

Gruzd, Anatoliy (2011). "Imagining Twitter as an Imagined Community". In: *American Behavioral Scientist* 55.10, pp. 1294–1318. DOI: 10.1177/0002764211409378.

Guilbeault, Douglas (2016). "Automation, Algorithms, and Politics | Growing Bot Security: An Ecological View of Bot Agency". In: *International Journal of Communication* 10.0, pp. 5003–5021. URL: https://ijoc.org/index.php/ijoc/article/view/6135.

Gupta, Sonu, Ponnurangam Kumaraguru, and Tanmoy Chakraborty (Jan. 2019). "MalReG: Detecting and Analyzing Malicious Retweeter Groups". In: *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*. CoDS-COMAD '19. ACM, pp. 61–69. DOI: 10.1145/3297001.3297009.

Géron, Aurélien (Oct. 2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly UK Ltd. ISBN: 1492032646.

Habermas, Jürgen (1996). *Between facts and norms: Contributions to a discourse theory of law and democracy*. MIT Press. ISBN: 9780262581622.

Hagberg, Aric A., Daniel A. Schult, and Pieter J. Swart (2008). "Exploring Network Structure, Dynamics, and Function using NetworkX". In: *Proceedings of the 7th Python in Science Conference*, pp. 11–15. URL: https://www.osti.gov/biblio/960616.

Häussler, Thomas (July 2018). "Heating up the debate? Measuring fragmentation and polarisation in a German climate change hyperlink network". In: *Social Networks* 54, pp. 303–313. DOI: 10.1016/j.socnet.2017.10.002.

Hawkes, Alan G. (1971). "Spectra of some self-exciting and mutually exciting point processes". In: *Biometrika* 58.1, pp. 83–90. DOI: 10.1093/biomet/58.1.83.

Heaven, Will Douglas (Oct. 2020). "A GPT-3 bot posted comments on Reddit for a week and no one noticed". In: *MIT Technology Review*. Accessed on 2021-05-24. URL: https://www.technologyreview.com/2020/10/08/1009845/a-gpt-3-bot-posted-comments-on-reddit-for-a-week-and-no-one-noticed/.

Hegelich, Simon (Mar. 2020). "Facebook needs to share more with researchers". In: *Nature* 579.7800, pp. 473–473. DOI: 10.1038/d41586-020-00828-5.

Hegelich, Simon and Dietmar Janetzko (May 2016). "Are Social Bots on Twitter Political Actors? Empirical Evidence from a Ukrainian Social Botnet". In: *Proceedings of the Tenth International Conference on Web and Social Media*. ICWSM '16. AAAI Press. URL: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13015.

Hine, Gabriel Emile, Jeremiah Onaolapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn (2017). "Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan's Politically Incorrect Forum and Its Effects on the Web". In: *Proceedings of the Eleventh International Conference on Web and Social Media*. ICWSM '17. AAAI Press, pp. 92–101. URL: https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15670.

Holme, Petter and Jari Saramäki (2012). "Temporal networks". In: *Physics Reports* 519.3, pp. 97–125. DOI: 10.1016/j.physrep.2012.03.001.

Holzmann, Helge, Avishek Anand, and Megha Khosla (Dec. 2018). "Delusive PageRank in Incomplete Graphs". In: *Complex Networks and Their Applications VII - Volume 1 Proceedings The 7th International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2018*. Vol. 812. Studies in Computational Intelligence. Springer, pp. 104–117. DOI: 10.1007/978-3-030-05411-3_9.

Howard, Philip N. (Nov. 2018). "The rise of computational propaganda". In: *IEEE Spectrum* 55.11, pp. 28–33. DOI: 10.1109/MSPEC.2018.8513781.

Howard, Philip N. and Bence Kollanyi (June 2016). *Bots, #StrongerIn, and #Brexit: Computational propaganda during the UK-EU referendum*. Research Note 2016.1. The Computational Propaganda Research Project, Oxford, UK. URL: http://comprop.oii.ox.ac.uk/wp-content/uploads/sites/89/2016/06/COMPROP-2016-1.pdf.

Hristakieva, Kristina, Stefano Cresci, Giovanni Da San Martino, Mauro Conti, and Preslav Nakov (Sept. 2021). "The Spread of Propaganda by Coordinated Communities on Social Media". In: *arXiv preprint*. arXiv: 2109.13046 `[cs.SI]`.

Hubert, Lawrence and Phipps Arabie (Dec. 1985). "Comparing partitions". In: *Journal of Classification* 2.1, pp. 193–218. DOI: 10.1007/bf01908075.

Hui, Pik-Mai, Kai-Cheng Yang, Christopher Torres-Lugo, Zachary Monroe, Marc McCarty, Benjamin Serrette, Valentin Pentchev, and Filippo Menczer (Oct. 2019). "BotSlayer: real-time detection of bot amplification on Twitter". In: *Journal of Open Source Software* 4.42, p. 1706. DOI: 10.21105/joss.01706.

Huszár, Ferenc, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt (Oct. 2021). *Algorithmic Amplification of Politics on Twitter*. Report. Twitter. URL: https://cdn.cms-twdigitalassets.com/content/dam/blog-twitter/official/en_us/company/2021/rml/Algorithmic-Amplification-of-Politics-on-Twitter.pdf.

Hwang, Tim (July 2020). *Deepfakes: A Grounded Threat Assessment*. Analysis Report. Center for Security and Emerging Technology. DOI: 10.51593/20190030.

Hwang, Tim, Ian Pearce, and Max Nanis (Mar. 2012). "Socialbots: Voices from the Fronts". In: *Interactions* 19.2, pp. 38–45. DOI: 10.1145/2090150.2090161.

IPCC (Aug. 2021). *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.

Isenberg, Daniel J. (June 1986). "Group polarization: A critical review and meta-analysis." In: *Journal of Personality and Social Psychology* 50.6, pp. 1141–1151. DOI: 10.1037/0022-3514.50.6.1141.

Jackson, Sarah J., Moya Bailey, and Brooke Foucault Welles (2020). *#HashtagActivism*. The MIT Press. DOI: 10.7551/mitpress/10858.001.0001.

Jamieson, Kathleen Hall (Aug. 2020). *Cyberwar*. Oxford University Press. DOI: 10.1093/oso/9780190058838.001.0001.

Janetzko, Dietmar (Dec. 2017). "Nonreactive Data Collection Online". In: *The SAGE Handbook of Online Research Methods*. Ed. by Nigel G. Fielding, Raymond M. Lee, and Grant Blank. 2nd Edition. London: SAGE Publications Ltd. Chap. 5, pp. 76–91. ISBN: 9781473918788. DOI: 10.4135/9781473957992.

Jansen, Nora (June 2019). "The Fiery, the Lovely, and the Hot - Analysis of Online Viral Phenomena in Social Media". In: *27th European Conference on Information Systems – Information Systems for a Sharing Society*. ECIS '19. URL: https://aisel.aisnet.org/ecis2019_rp/43.

Jiang, Julie, Xiang Ren, and Emilio Ferrara (Aug. 2021). "Social Media Polarization and Echo Chambers in the Context of COVID-19: Case Study". In: *JMIRx Med* 2.3, e29570. DOI: 10.2196/29570.

Jones, Matthew W., Adam J. P. Smith, Richard Betts, Josep G. Canadell, I. Colin Prentice, and Corinne Le Quéré (Jan. 2020). "Climate Change Increases the Risk of Wildfires: January 2020". In: *ScienceBrief*. URL: https://sciencebrief.org/briefs/wildfires.

Joseph, Kenneth, Peter M. Landwehr, and Kathleen M. Carley (Apr. 2014). "Two 1%s Don't Make a Whole: Comparing Simultaneous Samples from Twitter's Streaming API". In: *Social Computing, Behavioral-Cultural Modeling and Prediction – 7th International Conference*. SBP '14. Springer International Publishing, pp. 75–83. DOI: 10.1007/978-3-319-05579-4_10.

Jost, John T. (Mar. 2017). "Ideological Asymmetries and the Essence of Political Psychology". In: *Political Psychology* 38.2, pp. 167–208. DOI: 10.1111/pops.12407.

Jost, John T., Jack Glaser, Arie W. Kruglanski, and Frank J. Sulloway (2003). "Political conservatism as motivated social cognition." In: *Psychological Bulletin* 129.3, pp. 339–375. DOI: 10.1037/0033-2909.129.3.339.

Karell, Daniel, Andrew M. Linke, and Edward C. Holland (May 2021). "Right-Wing Social Media and Unrest Correspond Across the United States". In: *SocArXiv*. DOI: 10.31235/osf.io/pna5u.

Kavanagh, Jennifer and Michael D. Rich (2018). *Truth Decay: An Initial Exploration of the Diminishing Role of Facts and Analysis in American Public Life*. RAND Corporation. DOI: 10.7249/rr2314.

Keller, Franziska B., David Schoch, Sebastian Stier, and JungHwan Yang (May 2017). "How to Manipulate Social Media: Analyzing Political Astroturfing Using Ground

Truth Data from South Korea". In: *Proceedings of the Eleventh International Conference on Web and Social Media*. ICWSM '17. AAAI Press, pp. 564–567. URL: https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15638.

Keller, Franziska B., David Schoch, Sebastian Stier, and JungHwan Yang (Oct. 2019). "Political Astroturfing on Twitter: How to Coordinate a Disinformation Campaign". In: *Political Communication* 37.2, pp. 256–280. DOI: 10.1080/10584609.2019.1661888.

Keller, Tobias, Tim Graham, Dan Angus, Axel Bruns, Nahema Marchal, Lisa-Maria Neudert, Rolf Nijmeijer, Kristoffer Laigaard Nielbo, Marie Damsgaard Mortensen, Anja Bechmann, Patricia Rossini, Erica Anita Baptista, and Vanessa Veiga de Oliveira (Oct. 2020). *'Coordinated Inauthentic Behaviour' and Other Online Influence Operations in Social Media Spaces*. Panel presented at the Annual Conference of the Association of Internet Researchers, AoIR 2020. URL: https://spir.aoir.org/ojs/index.php/spir/article/view/11132/9763.

Kemp, Simon (Jan. 2021). *Digital 2021: Global Overview Report — DataReportal – Global Digital Insights*. Accessed 2021-05-30. URL: https://datareportal.com/reports/digital-2021-global-overview-report.

Kent, Thomas (Sept. 2020). *Striking Back: Overt and Covert Options to Combat Russian Disinformation*. The Jamestown Foundation. ISBN: 0998666092.

King, Gary, Jennifer Pan, and Margaret E. Roberts (2017). "How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument". In: *American Political Science Review* 111.3, 484—501. DOI: 10.1017/S0003055417000144.

Kivela, M., A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter (July 2014). "Multilayer networks". In: *Journal of Complex Networks* 2.3, pp. 203–271. DOI: 10.1093/comnet/cnu016.

Kligler-Vilenchik, Neta, Christian Baden, and Moran Yarchi (2020). "Interpretative polarization across platforms: How political disagreement develops over time on Facebook, Twitter, and WhatsApp". In: *Social Media + Society* 6.3. DOI: 10.1177/2056305120944393.

Kosinski, Michal, David Stillwell, and Thore Graepel (Mar. 2013). "Private traits and attributes are predictable from digital records of human behavior". In: *Proceedings of the National Academy of Sciences* 110.15, pp. 5802–5805. DOI: 10.1073/pnas.1218772110.

Krackhardt, David and Robert N. Stern (June 1988). "Informal Networks and Organizational Crises: An Experimental Simulation". In: *Social Psychology Quarterly* 51.2, pp. 123–140. DOI: 10.2307/2786835.

Kumar, Srijan, Justin Cheng, and Jure Leskovec (Apr. 2017a). "Antisocial Behavior on the Web: Characterization and Detection". In: *Proceedings of the 26th International Conference on World Wide Web Companion*. WWW '17. ACM Press, pp. 947–950. DOI: 10.1145/3041021.3051106.

Kumar, Srijan, Justin Cheng, Jure Leskovec, and V.S. Subrahmanian (Apr. 2017b). "An Army of Me: Sockpuppets in Online Discussion Communities". In: *Proceedings of the 26th International Conference on World Wide Web*. WWW '17. ACM Press, pp. 857–866. DOI: 10.1145/3038912.3052677.

Kumar, Srijan, William L. Hamilton, Jure Leskovec, and Dan Jurafsky (Apr. 2018). "Community Interaction and Conflict on the Web". In: *Proceedings of the 2018 World Wide Web Conference*. WWW '18. ACM Press, pp. 933–943. DOI: 10.1145/3178876.3186141.

Kumar, Srijan and Neil Shah (Apr. 2018). "False Information on Web and Social Media: A Survey". In: *arXiv preprint*. arXiv: 1804.08559 [cs.SI].

Kuran, Timur and Cass R. Sunstein (Apr. 1999). "Availability Cascades and Risk Regulation". In: *Stanford Law Review* 51.4. DOI: 10.2307/1229439.

Kušen, Ema and Mark Strembeck (2020). "You talkin' to me? Exploring Human/Bot Communication Patterns during Riot Events". In: *Information Processing & Management* 57.1, p. 102126. DOI: 10.1016/j.ipm.2019.102126.

Lapowsky, Issie (Jan. 2021). "Doxxing insurrectionists: Capitol riot divides online extremism researchers". In: *Protocol*. Accessed 2022-01-31. URL: https://www.protocol.com/doxxing-capitol-rioters.

Latah, Majd (Aug. 2020). "Detection of malicious social bots: A survey and a refined taxonomy". In: *Expert Systems with Applications* 151, p. 113383. DOI: 10.1016/j.eswa.2020.113383.

Laub, Patrick J., Thomas Taimre, and Philip K. Pollett (July 2015). "Hawkes Processes". In: *arXiv preprint*. arXiv: 1507.02822 [math.PR].

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani (Mar. 2014). "The Parable of Google Flu: Traps in Big Data Analysis". In: *Science* 343.6176, pp. 1203–1205. DOI: 10.1126/science.1248506.

Lee, Kyumin, James Caverlee, Zhiyuan Cheng, and Daniel Z. Sui (Dec. 2013). "Campaign extraction from social media". In: *ACM Transactions on Intelligent Systems and Technology* 5.1, 9:1–9:28. DOI: 10.1145/2542182.2542191.

Leetaru, Kalev (Mar. 2019). "Twitter users mostly retweet politicians and celebrities. That's a big change." In: *The Washington Post*. Accessed 2022-01-31. URL: https://www.washingtonpost.com/politics/2019/03/08/twitter-users-mostly-retweet-politicians-celebrities-thats-big-change/.

Lelkes, Yphtach (Mar. 2016). "Mass Polarization: Manifestations and Measurements". In: *Public Opinion Quarterly* 80.S1, pp. 392–410. DOI: 10.1093/poq/nfw005.

Li, Huayi, Geli Fei, Shuai Wang, Bing Liu, Weixiang Shao, Arjun Mukherjee, and Jidong Shao (Apr. 2017a). "Bimodal Distribution and Co-Bursting in Review Spam Detection". In: *Proceedings of the 26th International Conference on World Wide Web*. WWW '17. ACM, pp. 1063–1072. DOI: 10.1145/3038912.3052582.

Li, Ping, Benjamin Schloss, and D. Jake Follmer (July 2017b). "Speaking two "Languages" in America: A semantic space analysis of how presidential candidates and their supporters represent abstract political concepts differently". In: *Behavior Research Methods* 49.5, pp. 1668–1685. DOI: 10.3758/s13428-017-0931-5.

Lim, Kwan Hui, Sachini Jayasekara, Shanika Karunasekera, Aaron Harwood, Lucia Falzon, John Dunn, and Glenn Burgess (2019). "RAPID: Real-time Analytics Platform for Interactive Data Mining". In: *Machine Learning and Knowledge Discovery in Databases*. Springer International Publishing, pp. 649–653. DOI: 10.1007/978-3-030-10997-4_44.

Loomba, Sahil, Alexandre de Figueiredo, Simon J. Piatek, Kristen de Graaf, and Heidi J. Larson (Feb. 2021). "Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA". In: *Nature Human Behaviour* 5.3, pp. 337–348. DOI: 10.1038/s41562-021-01056-1.

Lou, Xiaodan, Alessandro Flammini, and Filippo Menczer (July 2019). "Manipulating the Online Marketplace of Ideas". In: *arXiv preprint*. arXiv: 1907.06130v2 [cs.CY].

Loucaides, Darren, Alessio Perrone, and Josef Holnburger (June 2021). "How Germany became ground zero for the COVID infodemic". In: *openDemocracy*. Accessed 2022-01-31. URL: https://www.opendemocracy.net/en/germany-ground-zero-covid-infodemic-russia-far-right/.

Mackintosh, Eliza (Oct. 2021). "Facebook knew it was being used to incite violence in Ethiopia. It did little to stop the spread, documents show". In: *CNN* The Facebook papers. URL: https://edition.cnn.com/2021/10/25/business/ethiopia-violence-facebook-papers-cmd-intl/index.html.

Magelinski, Thomas and Kathleen M. Carley (Nov. 2020). *Detecting Coordinated Behavior in the Twitter Campaign to Reopen America*. Talk presented at the Center for Informed Democracy & Social-cybersecurity (IDeaS) annual conference. URL: https://www.cmu.edu/ideas-social-cybersecurity/events/conference-archive/2020papers/magelinski_ideas_abstract_reopen.pdf.

Magelinski, Thomas, Lynnette Hui Xian Ng, and Kathleen M. Carley (May 2021). "A Synchronized Action Framework for Responsible Detection of Coordination on Social Media". In: *arXiv preprint*. arXiv: 2105.07454 [cs.SI].

Malone, Thomas W. and Kevin Crowston (Mar. 1994). "The Interdisciplinary Study of Coordination". In: *ACM Computing Surveys* 26.1, pp. 87–119. DOI: 10.1145/174666.174668.

Mariconti, Enrico, Jeremiah Onaolapo, Syed Sharique Ahmad, Nicolas Nikiforou, Manuel Egele, Nick Nikiforakis, and Gianluca Stringhini (Apr. 2017). "What's in a Name?: Understanding Profile Name Reuse on Twitter". In: *Proceedings of the 26th International Conference on World Wide Web*. WWW '17. ACM, pp. 1161–1170. DOI: 10.1145/3038912.3052589.

Mariconti, Enrico, Guillermo Suarez-Tangil, Jeremy Blackburn, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Jordi Luque Serrano, and Gianluca Stringhini (Nov. 2019). ""You Know What to Do": Proactive Detection of YouTube Videos Targeted by Coordinated Hate Attacks". In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW, pp. 1–21. DOI: 10.1145/3359309.

Marozzo, Fabrizio and Alessandro Bessi (Nov. 2017). “Analyzing polarization of social media users and news sites during political campaigns”. In: *Social Network Analysis and Mining* 8.1. DOI: 10.1007/s13278-017-0479-5.

Massanari, Adrienne (July 2016). “#Gamergate and The Fappening: How Reddit’s algorithm, governance, and culture support toxic technocultures”. In: *New Media & Society* 19.3, pp. 329–346. DOI: 10.1177/1461444815608807.

Mazza, Michele, Stefano Cresci, Marco Avvenuti, Walter Quattrociocchi, and Maurizio Tesconi (June 2019). “RTbust: Exploiting Temporal Patterns for Botnet Detection on Twitter”. In: *Proceedings of the 10th ACM Conference on Web Science*. WebSci’19. ACM, pp. 183–192. DOI: 10.1145/3292522.3326015.

McGregor, Andrew (May 2014). “Graph stream algorithms: a survey”. In: *ACM SIGMOD Record* 43.1, pp. 9–20. DOI: 10.1145/2627692.2627694.

McKew, Molly K. (Feb. 2018). “How Twitter bots and Trump fans made #ReleaseTheMemo go viral”. In: *POLITICO*. Accessed 2022-01-31. URL: https://www.politico.eu/article/how-twitter-bots-and-trump-fans-made-releasethememo-go-viral/amp/.

McPherson, Miller, Lynn Smith-Lovin, and James M. Cook (Aug. 2001). “Birds of a Feather: Homophily in Social Networks”. In: *Annual Review of Sociology* 27.1, pp. 415–444. DOI: 10.1146/annurev.soc.27.1.415.

Merrill, Jeremy B. and Will Oremus (Oct. 2021). “Five points for anger, one for a ‘like’: How Facebook’s formula fostered rage and misinformation”. In: *The Washington Post*. Accessed 2022-01-31. URL: https://www.washingtonpost.com/technology/2021/10/26/facebook-angry-emoji-algorithm/.

Metaxas, Panagiotis T. and Eni Mustafaraj (Oct. 2012). “Social Media and the Elections”. In: *Science* 338.6106, pp. 472–473. DOI: 10.1126/science.1230456.

Metaxas, Panagiotis Takis, Eni Mustafaraj, Kily Wong, Laura Zeng, Megan O’Keefe, and Samantha Finn (May 2015). “What Do Retweets Indicate? Results from User Survey and Meta-Review of Research”. In: *Proceedings of the Ninth International Conference on Web and Social Media*. ICWSM ’15. AAAI Press, pp. 658–661. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14661.

Miller, Greg (2018). *The Apprentice: Trump, Russia, and the Subversion of American Democracy*. London: William Collins. ISBN: 9780008325756.

Mitra, Tanushree, Graham P. Wright, and Eric Gilbert (Feb. 2017). “A Parsimonious Language Model of Social Media Credibility Across Disparate Events”. In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. CSCW ’17. ACM, pp. 126–145. DOI: 10.1145/2998181.2998351.

Mooney, Carl H. and John F. Roddick (Feb. 2013). “Sequential pattern mining – approaches and algorithms”. In: *ACM Computing Surveys* 45.2, pp. 1–39. DOI: 10.1145/2431211.2431218.

Mordelet, Fantine and Jean-Philippe Vert (Feb. 2014). “A bagging SVM to learn from positive and unlabeled examples”. In: *Pattern Recognition Letters* 37, pp. 201–209. DOI: 10.1016/j.patrec.2013.06.010.

Morstatter, Fred, Jürgen Pfeffer, Huan Liu, and Kathleen M. Carley (July 2013). "Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose". In: *Proceedings of the Seventh International Conference on Weblogs and Social Media*. ICWSM '13. AAAI Press, pp. 400–408. URL: https://ojs.aaai.org/index.php/ICWSM/article/download/14401/14250/17919.

Morstatter, Fred, Yunqiu Shao, Aram Galstyan, and Shanika Karunasekera (Apr. 2018). "From *Alt-Right* to *Alt-Rechts*: Twitter Analysis of the 2017 German Federal Election". In: *The 2018 Web Conference Companion*. WWW '18. ACM Press, pp. 621–628. DOI: 10.1145/3184558.3188733.

Mueller, Robert (Feb. 2018). "Indictment, United States v. Internet Research Agency LLC et al." In: *U.S. District Court for the District of Columbia*. Docket entry 1, Feb. 16, 2018, Case no. 18-cr-00032-DLF.

Nasim, Mehwish (2016). "Inferring Social Relations from Online and Communication Networks". PhD thesis. Konstanz, Germany: Computer and Information Science, University of Konstanz.

Nasim, Mehwish (Nov. 2019). *Polarisation on social media: modelling and evaluation*. Talk presented at the Australian Social Network Analysis Conference, ASNAC '19.

Nasim, Mehwish, Raphaël Charbey, Christophe Prieur, and Ulrik Brandes (2016). "Investigating Link Inference in Partially Observable Networks: Friendship Ties and Interaction". In: *IEEE Transactions on Computational Social Systems* 3.3, pp. 113–119. DOI: 10.1109/TCSS.2016.2618998.

Nasim, Mehwish, Muhammad Usman Ilyas, Aimal Rextin, and Nazish Nasim (May 2013). "On commenting behavior of Facebook users". In: *Proceedings of the 24th ACM Conference on Hypertext and Social Media*. HT '13. ACM Press, pp. 179–183. DOI: 10.1145/2481492.2481513.

Nasim, Mehwish, Andrew Nguyen, Nick Lothian, Robert Cope, and Lewis Mitchell (2018). "Real-time Detection of Content Polluters in Partially Observable Twitter Networks". In: *Companion Proceedings of the The Web Conference 2018*. WWW '18. ACM Press, pp. 1331–1339. DOI: 10.1145/3184558.3191574.

Nasim, Mehwish, Jonathan Tuke, Nigel Bean, and Lewis Mitchell (May 2019). *Emergence of echo chambers in social relations and sentiments on Twitter: An observational study*. Talk presented at NetSci 2019 conference.

Nasim, Mehwish, Derek Weber, Tobin South, Jonathan Tuke, Nigel Bean, Lucia Falzon, and Lewis Mitchell (Jan. 2022). "Are we always in strife? A longitudinal study of the echo chamber effect in the Australian Twittersphere". In: *arXiv preprint*. arXiv: 2201.09161 `[cs.SI]`.

Nastos, James and Yong Gao (July 2013). "Familial groups in social networks". In: *Social Networks* 35.3, pp. 439–450. DOI: 10.1016/j.socnet.2013.05.001.

Neudert, Lisa-Maria N. (2018). "Germany: A Cautionary Tale". In: *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media*. Oxford University Press. Chap. 7, pp. 153–184. DOI: 10.1093/oso/9780190931407.003.0008.

Newitz, Annalee (Aug. 2015). “Ashley Madison code shows more women, and more bots”. In: *Gizmodo*. Accessed 2021-11-24. URL: https://gizmodo.com/ashley-madison-code-shows-more-women-and-more-bots-1727613924.

Newman, M. E. J. (Feb. 2003). “Mixing patterns in networks”. In: *Physical Review E* 67.2, p. 026126. DOI: 10.1103/physreve.67.026126.

Newman, M. E. J. (2010). *Networks: An Introduction*. Oxford University Press. ISBN: 9780199206650.

Ng, Lynnette Hui Xian, Iain Cruickshank, and Kathleen M. Carley (Sept. 2021). “Coordinating Narratives and the Capitol Riots on Parler”. In: *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation Disinformation Challenge 2021*. arXiv: 2109.00945 `[cs.SI]`.

Nguyen, C. Thi (Sept. 2018). “ECHO CHAMBERS AND EPISTEMIC BUBBLES”. In: *Episteme* 17.2, pp. 141–161. DOI: 10.1017/epi.2018.32.

Nick, Bobo, Conrad Lee, Pádraig Cunningham, and Ulrik Brandes (Aug. 2013). “Simmelian backbones: Amplifying hidden homophily in Facebook networks”. In: *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ASONAM ’13. IEEE. ACM, pp. 525–532. DOI: 10.1145/2492517.2492569.

Nikolov, Dimitar, Alessandro Flammini, and Filippo Menczer (Feb. 2021). “Right and left, partisanship predicts (asymmetric) vulnerability to misinformation”. In: *Harvard Kennedy School Misinformation Review*. DOI: 10.37016/mr-2020-55.

Nimmo, Ben (Nov. 2017). “How A Russian Troll Fooled America”. In: *DFRLab*. Accessed 2022-01-31. URL: https://medium.com/dfrlab/how-a-russian-troll-fooled-america-80452a4806d1.

Nimmo, Ben (Sept. 2020). *The Breakout Scale: Measuring the Impact of Influence Operations*. Report. The Brookings Institution. URL: https://www.brookings.edu/research/the-breakout-scale-measuring-the-impact-of-influence-operations/.

Nimmo, Ben, Camille François, C. Shawn Eib, Lea Ronzaud, Rodrigo Ferreira, Chris Hernon, and Tim Kostelancik (June 2020). *Exposing Secondary Infektion*. Report. Graphika. URL: https://secondaryinfektion.org/.

Nizzoli, Leonardo, Serena Tardelli, Marco Avvenuti, Stefano Cresci, and Maurizio Tesconi (June 2021). “Coordinated Behavior on Social Media in 2019 UK General Election”. In: *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media*. ICWSM ’21. AAAI Press, pp. 443–454. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/18074.

Nocaj, Arlind, Mark Ortmann, and Ulrik Brandes (2014). “Untangling Hairballs”. In: *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Springer International Publishing, pp. 101–112. DOI: 10.1007/978-3-662-45803-7_9.

Notley, Tanya, Simon Chambers, Sora Park, and Michael Dezuanni (Apr. 2021). *Adult Media Literacy: Attitudes, Experiences and Needs*. Report. Western Sydney University, Queensland University of Technology and University of Canberra. URL: https://westernsydney.edu.au/ics/news/report_adult_media_literacy_in_australia.

NSW Bushfire Inquiry (July 2020). *Final Report of the NSW Bushfire Inquiry*. State Inquiry Report. NSW State Government. URL: https://www.dpc.nsw.gov.au/assets/dpc-nsw-gov-au/publications/NSW-Bushfire-Inquiry-1630/Final-Report-of-the-NSW-Bushfire-Inquiry.pdf.

OECD (July 2021). *Government at a Glance 2021*. OECD. DOI: 10.1787/1c258f55-en.

Oentaryo, Richard J., Arinto Murdopo, Philips K. Prasetyo, and Ee-Peng Lim (Oct. 2016). "On Profiling Bots in Social Media". In: *International Conference on Social Informatics*. SocInfo '16. Springer, pp. 92–109. DOI: 10.1007/978-3-319-47880-7_6.

Olteanu, Alexandra, Carlos Castillo, Fernando Diaz, and Emre Kıcıman (July 2019). "Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries". In: *Frontiers in Big Data* 2. DOI: 10.3389/fdata.2019.00013.

Pacheco, Diogo, Alessandro Flammini, and Filippo Menczer (Apr. 2020). "Unveiling Coordinated Groups Behind White Helmets Disinformation". In: *Companion Proceedings of the Web Conference 2020*. WWW '20. ACM, pp. 611–616. DOI: 10.1145/3366424.3385775.

Pacheco, Diogo, Pik-Mai Hui, Christopher Torres-Lugo, Bao Tran Truong, Alessandro Flammini, and Filippo Menczer (June 2021). "Uncovering Coordinated Networks on Social Media: Methods and Case Studies". In: *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media*. ICWSM '21. AAAI Press, pp. 455–466. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/18075.

Paik, Jiaul H and Jimmy Lin (Aug. 2015). "Do multiple listeners to the public Twitter sample stream receive the same tweets?" In: *SIGIR 2015 Workshop on Temporal, Social and Spatially-aware Information Access*. TAIA '15. URL: https://cs.uwaterloo.ca/~jimmylin/publications/Paik_Lin_TAIA2015.pdf.

Pariser, Eli (Apr. 2012). *The Filter Bubble*. Penguin LCC US. ISBN: 0143121235.

Paul, Christopher and Miriam Matthews (2016). *The Russian "Firehose of Falsehood" Propaganda Model: Why It Might Work and Options to Counter It*. RAND Corporation. DOI: 10.7249/pe198.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay (Nov. 2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830. URL: http://dl.acm.org/citation.cfm?id=2078195.

Persily, Nathaniel and Joshua A. Tucker, eds. (Aug. 2020). *Social Media and Democracy*. Cambridge University Press. DOI: 10.1017/9781108890960.

Petitjean, François, Alain Ketterlin, and Pierre Gançarski (Mar. 2011). "A global averaging method for dynamic time warping, with applications to clustering". In: *Pattern Recognition* 44.3, pp. 678–693. DOI: 10.1016/j.patcog.2010.09.013.

Pfeffer, Jürgen, Katja Mayer, and Fred Morstatter (Dec. 2018). "Tampering with Twitter's Sample API". In: *EPJ Data Science* 7.1, pp. 50–70. DOI: 10.1140/epjds/s13688-018-0178-0.

Phillips, Whitney (May 2018). *The Oxygen of Amplification*. Report. Data & Society Research Institute. URL: http://datasociety.net/output/oxygen-of-amplification/.

Pitcavage, Mark (Feb. 2001). "Camouflage and Conspiracy: The Militia Movement From Ruby Ridge to Y2K". In: *American Behavioral Scientist* 44.6, pp. 957–981. DOI: 10.1177/00027640121956610.

Radicioni, Tommaso, Fabio Saracco, Elena Pavan, and Tiziano Squartini (June 2021). "Analysing Twitter Semantic Networks: The case of 2018 Italian Elections". In: *Scientific Reports* 11.1. DOI: 10.1038/s41598-021-92337-2.

Ram, Rohit, Quyu Kong, and Marian-Andrei Rizoiu (Mar. 2021). "Birdspotter: A Tool for Analyzing and Labeling Twitter Users". In: *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. WSDM '21. ACM, pp. 918–921. DOI: 10.1145/3437963.3441695.

Ratkiewicz, J., M.D. Conover, M.R. Meiss, B. Gonçalves, A. Flammini, and F. Menczer (2011). "Detecting and Tracking Political Abuse in Social Media". In: *Proceedings of the Fifth International Conference on Weblogs and Social Media*. ICWSM '11. The AAAI Press, pp. 297–304. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14127.

Rid, Thomas (2020). *Active measures: the secret history of disinformation and political warfare*. Profile Books. ISBN: 9781788164757.

Rizoiu, Marian-Andrei, Timothy Graham, Rui Zhang, Yifei Zhang, Robert Ackland, and Lexing Xie (2018). "#DebateNight: The Role and Influence of Socialbots on Twitter During the 1st 2016 U.S. Presidential Debate". In: *Proceedings of the Twelfth International Conference on Web and Social Media*. ICWSM '18. AAAI Press, pp. 300–309. URL: https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17886.

Rizoiu, Marian-Andrei, Lexing Xie, Scott Sanner, Manuel Cebrián, Honglin Yu, and Pascal Van Hentenryck (Apr. 2017). "Expecting to be HIP: Hawkes Intensity Processes for Social Media Popularity". In: *Proceedings of the 26th International Conference on World Wide Web*. WWW '17. ACM, pp. 735–744. DOI: 10.1145/3038912.3052650.

Robins, Garry L. (Jan. 2015). *Doing social network research: Network-based research design for social scientists*. SAGE Publications Ltd. ISBN: 1446276139.

Roccetti, Marco, Giovanni Delnevo, Luca Casini, and Paola Salomoni (Feb. 2020). "A Cautionary Tale for Machine Learning Design: why we Still Need Human-Assisted Big Data Analysis". In: *Mobile Networks and Applications* 25.3, pp. 1075–1083. DOI: 10.1007/s11036-020-01530-6.

Rogers, Everett M and Dilip K Bhowmik (1970). "Homophily-heterophily: Relational concepts for communication research". In: *Public opinion quarterly* 34.4, pp. 523–538. DOI: 10.1086/267838.

Romano, Aja (July 2016). "Milo Yiannopoulos's Twitter ban, explained". In: *Vox*. Accessed 2022-01-31. URL: https://www.vox.com/2016/7/20/12226070/milo-yiannopoulus-twitter-ban-explained.

Ruths, Derek and Jürgen Pfeffer (Nov. 2014). "Social media for large studies of behavior". In: *Science* 346.6213, pp. 1063–1064. DOI: 10.1126/science.346.6213.1063.

Saleem, Saima and Monica Mehrotra (Nov. 2021). "Emergent Use of Artificial Intelligence and Social Media for Disaster Management". In: *Proceedings of International Conference on Data Science and Applications*. ICDSA '21. Springer Singapore, pp. 195–210. DOI: 10.1007/978-981-16-5348-3_15.

Samuels, Elyse (Feb. 2020). "How misinformation on WhatsApp led to a mob killing in India". In: *The Washington Post*. Accessed 2022-01-31. URL: https://www.washingtonpost.com/politics/2020/02/21/how-misinformation-whatsapp-led-deathly-mob-lynching-india/.

Sarabadani, Amir, Aaron Halfaker, and Dario Taraborelli (2017). "Building Automated Vandalism Detection Tools for Wikidata". In: *Proceedings of the 26th International Conference on World Wide Web Companion*. WWW '17. ACM Press, pp. 1647–1654. DOI: 10.1145/3041021.3053366.

Scannapieco, Monica, Paolo Missier, and Carlo Batini (2005). "Data Quality at a Glance". In: *Datenbank-Spektrum* 14, pp. 6–14. URL: https://eprints.ncl.ac.uk/174218.

Schliebs, Marcel, Hannah Bailey, Jonathan Bright, and Philip N. Howard (May 2021). *China's Inauthentic UK Twitter diplomacy: A Coordinated Network Amplifying DRC Diplomats*. Working Paper 2021.2. The Programme on Democracy & Technology, Oxford University. URL: https://demtech.oii.ox.ac.uk/china-public-diplomacy-casestudy-uk.

Schroeder, Ralph (Jan. 2018). *Social Theory after the Internet*. UCL Press. DOI: 10.14324/111.9781787351226.

Scott, Mark (Jan. 2021). "Capitol Hill riot lays bare what's wrong with social media". In: *POLITICO*. Accessed 2021-02-08. URL: https://www.politico.eu/article/us-capitol-hill-riots-lay-bare-whats-wrong-social-media-donald-trump-facebook-twitter/.

Şen, Fatih, Rolf T. Wigand, Nitin Agarwal, Serpil Tokdemir Yuce, and Rafal Kasprzyk (Apr. 2016). "Focal structures analysis: Identifying influential sets of individuals in a social network". In: *Social Network Analysis and Mining* 6.1, 17:1–17:22. DOI: 10.1007/s13278-016-0319-z.

Serrano, M. A., M. Boguna, and A. Vespignani (Apr. 2009). "Extracting the multiscale backbone of complex weighted networks". In: *Proceedings of the National Academy of Sciences* 106.16, pp. 6483–6488. DOI: 10.1073/pnas.0808904106.

Sessions, Valerie and Marco Valtorta (Nov. 2006). "The Effects of Data Quality on Machine Learning Algorithms". In: *Proceedings of the 11th International Conference* *on Information Quality*. ICIQ '06. MIT, pp. 485–498. URL: http://mitiq.mit.edu/ICIQ/Documents/IQConference2006/papers/TheEffectsofDataQualityonMachineLearningAlgorithms.pdf.

Shah, Neil (Apr. 2017). "FLOCK: Combating Astroturfing on Livestreaming Platforms". In: *Proceedings of the 26th International Conference on World Wide Web*. WWW '17. ACM, pp. 1083–1091. DOI: 10.1145/3038912.3052617.

Shao, Chengcheng, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer (2016). "Hoaxy: A Platform for Tracking Online Misinformation". In: *Proceedings of the 25th International Conference Companion on World Wide Web*. WWW '16. ACM Press, pp. 745–750. DOI: 10.1145/2872518.2890098.

Shao, Chengcheng, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer (Nov. 2018a). "The spread of low-credibility content by social bots". In: *Nature Communications* 9.4787. DOI: 10.1038/s41467-018-06930-7.

Shao, Chengcheng, Pik-Mai Hui, Lei Wang, Xinwen Jiang, Alessandro Flammini, Filippo Menczer, and Giovanni Luca Ciampaglia (Apr. 2018b). "Anatomy of an online misinformation network". In: *PLOS ONE* 13.4, e0196087. DOI: 10.1371/journal.pone.0196087.

Sharma, Karishma, Yizhou Zhang, Emilio Ferrara, and Yan Liu (Aug. 2021). "Identifying Coordinated Accounts on Social Media through Hidden Influence and Group Behaviours". In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. KDD '21. ACM, pp. 1441–1451. DOI: 10.1145/3447548.3467391.

Shearer, Elisa and Elizabeth Grieco (Oct. 2019). *Americans are wary of the role social media sites play in delivering the news*. Report. Pew Research Center. URL: https://www.journalism.org/2019/10/02/americans-are-wary-of-the-role-social-media-sites-play-in-delivering-the-news/.

Shorey, Samantha and Philip N Howard (2016). "Automation, Algorithms, and Politics| Automation, Big Data and Politics: A Research Review". In: *International Journal of Communication* 10, pp. 5032–5055. URL: http://ijoc.org/index.php/ijoc/article/view/6233/1812.

Simmel, Georg (1908). "Das Geheimnis und die geheime Gesellschaft". In: *Soziologie. Untersuchungen über die Formen der Vergesellschaftung*, pp. 256–304.

Simon, Herbert A. (1971). "Designing organizations for an information rich world". In: *Computers, communications, and the public interest*. Ed. by Martin Greenberger. Baltimore, pp. 37–72. ISBN: 0-8018-1135-X.

Singer, P. W. and Emerson T. Brooking (Oct. 2019). *Likewar: The Weaponization of Social Media*. MARINER BOOKS. ISBN: 0358108470.

Skitka, Linda J. and Christopher W. Bauman (Jan. 2008). "Moral Conviction and Political Engagement". In: *Political Psychology* 29.1, pp. 29–54. DOI: 10.1111/j.1467-9221.2007.00611.x.

Smith, Rory, Seb Cubbon, and Claire Wardle (Nov. 2020). *Under the surface: Covid-19 vaccine narratives, misinformation and data deficits on social media.* Report. First News. URL: https://firstdraftnews.org/vaccine-narratives-full-report-november-2020.

Starbird, Kate (July 2019). "Disinformation's spread: bots, trolls and all of us". In: *Nature* 571.7766, pp. 449–449. DOI: 10.1038/d41586-019-02235-x.

Starbird, Kate, Ahmer Arif, and Tom Wilson (Nov. 2019). "Disinformation as Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operations". In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW, 127:1–127:26. DOI: 10.1145/3359229.

Starbird, Kate and Tom Wilson (Jan. 2020). "Cross-Platform Disinformation Campaigns: Lessons Learned and Next Steps". In: *Harvard Kennedy School Misinformation Review.* DOI: 10.37016/mr-2020-002.

Stilgherrian (Jan. 2020). "Twitter bots and trolls promote conspiracy theories about Australian bushfires". In: *ZDNet.* Accessed 2022-01-31. URL: https://www.zdnet.com/article/twitter-bots-and-trolls-promote-conspiracy-theories-about-australian-bushfires/.

Strick, Benjamin (Aug. 2021). *Analysis of the Pro-China Propaganda Network Targeting International Narratives.* Research Report. Centre for Information Resilience. URL: https://www.info-res.org/post/revealed-coordinated-attempt-to-push-pro-china-anti-western-narratives-on-social-media.

Subramanian, Samanth (Feb. 2017). "Inside the Macedonian Fake-News Complex". In: *Wired.* Accessed 2022-01-31. URL: https://www.wired.com/2017/02/veles-macedonia-fake-news/.

Sun, Yanmin, Andrew K. C. Wong, and Mohamed S. Kamel (June 2009). "Classification of imbalanced data: A review". In: *International Journal of Pattern Recognition and Artificial Intelligence* 23.04, pp. 687–719. DOI: 10.1142/s0218001409007326.

Sunstein, Cass R. (June 2002). "The Law of Group Polarization". In: *Journal of Political Philosophy* 10.2, pp. 175–195. DOI: 10.1111/1467-9760.00148.

Sunstein, Cass R. and Adrian Vermeule (June 2009). "Conspiracy Theories: Causes and Cures". In: *Journal of Political Philosophy* 17.2, pp. 202–227. DOI: 10.1111/j.1467-9760.2008.00325.x.

Sylwester, Karolina and Matthew Purver (Sept. 2015). "Twitter Language Use Reflects Psychological Differences between Democrats and Republicans". In: *PLOS ONE* 10.9, e0137422. DOI: 10.1371/journal.pone.0137422.

Tamine, Lynda, Laure Soulier, Lamjed Ben Jabeur, Frederic Amblard, Chihab Hanachi, Gilles Hubert, and Camille Roth (July 2016). "Social Media-Based Collaborative Information Access: Analysis of Online Crisis-Related Twitter Conversations". In: *Proceedings of the 27th ACM Conference on Hypertext and Social Media.* HT '16. ACM, pp. 159–168. DOI: 10.1145/2914586.2914589.

Tasnim, Samia, Md Mahbub Hossain, and Hoimonty Mazumder (May 2020). “Impact of Rumors and Misinformation on COVID-19 in Social Media”. In: *Journal of Preventive Medicine and Public Health* 53.3, pp. 171–174. DOI: 10.3961/jpmph.20.094.

The Soufan Center (Apr. 2021a). *QAnon – A U.S. National Security Threat Amplified by Foreign-Based Actors*. IntelBrief. The Soufan Center. URL: https://thesoufancenter.org/intelbrief-2021-april-20/.

The Soufan Center (Apr. 2021b). *Quantifying the Q Conspiracy: A Data-Driven Approach to Understanding the Threat Posed by QAnon*. Special Report. The Soufan Center. URL: https://thesoufancenter.org/research/quantifying-the-q-conspiracy-a-data-driven-approach-to-understanding-the-threat-posed-by-qanon/.

Timberg, Craig, Elizabeth Dwoskin, and Reed Albergotti (Oct. 2021). “How Facebook played a role in the Jan. 6 Capitol riot”. In: *The Washington Post*. Accessed 2022-01-31. URL: https://www.washingtonpost.com/technology/2021/10/22/jan-6-capitol-riot-facebook/.

Tollis, Ioannis G., Giuseppe Di Battista, Peter Eades, and Roberto Tamassia (July 1999). *Graph drawing : Algorithms for the visualization of graphs*. Prentice Hall. ISBN: 9780133016154.

Tromble, Rebekah, Andreas Storz, and Daniela Stockmann (Dec. 2017). “We don’t know what we don’t know: When and how the use of Twitter’s public APIs biases scientific inference”. In: *SSRN Electronic Journal*, pp. 1–26. DOI: 10.2139/ssrn.3079927.

Truong, Bao Tran, Oliver Melbourne Allen, and Filippo Menczer (Jan. 2022). “News Sharing Networks Expose Information Polluters on Social Media”. In: *arXiv preprint*. arXiv: 2202.00094 `[cs.SI]`.

Tsvetkova, Milena, Ruth García-Gavilanes, Luciano Floridi, and Taha Yasseri (Feb. 2017). “Even good bots fight: The case of Wikipedia”. In: *PLOS ONE* 12.2, e0171774. DOI: 10.1371/journal.pone.0171774.

Tufekci, Zeynep (June 2014). “Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls”. In: *Proceedings of the Eighth International Conference on Weblogs and Social Media*. ICWSM ’14. The AAAI Press, pp. 505–514. URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8062.

Tuke, Jonathan, Andrew Nguyen, Mehwish Nasim, Drew Mellor, Asanga Wickramasinghe, Nigel Bean, and Lewis Mitchell (Mar. 2020). “Pachinko Prediction: A Bayesian method for event prediction from social media data”. In: *Information Processing & Management* 57.2, p. 102147. DOI: 10.1016/j.ipm.2019.102147.

Tversky, Amos and Daniel Kahneman (Sept. 1973). “Availability: A heuristic for judging frequency and probability”. In: *Cognitive Psychology* 5.2, pp. 207–232. DOI: 10.1016/0010-0285(73)90033-9.

Understanding Mass Influence (Aug. 2021). Commissioned Report. Produced for the Australian Department of Defence. The University of Adelaide, The University of Melbourne,University of New South Wales, Edith Cowan University and Macquarie

University. URL: https://documentcloud.adobe.com/link/review?uri=urn:aaid:scds:US:dcbca90e-72e8-469d-98a6-605b8d97421b#pageNum=1.

Vargas, Luis, Patrick Emami, and Patrick Traynor (Nov. 2020). "On the Detection of Disinformation Campaign Activity with Network Analysis". In: *Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop*. CCS '20. ACM, pp. 133–146. DOI: 10.1145/3411495.3421363.

Varol, Onur, Emilio Ferrara, Clayton A. Davis, Filippo Menczer, and Alessandro Flammini (May 2017a). "Online Human-Bot Interactions: Detection, Estimation, and Characterization". In: *Proceedings of the Eleventh International Conference on Web and Social Media*. ICWSM '17. AAAI Press, pp. 280–289. URL: https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15587.

Varol, Onur, Emilio Ferrara, Filippo Menczer, and Alessandro Flammini (July 2017b). "Early detection of promoted campaigns on social media". In: *EPJ Data Science* 6.1, 13:1–13:19. DOI: 10.1140/epjds/s13688-017-0111-y.

Venturini, Tommaso, Anders Munk, and Mathieu Jacomy (Dec. 2019). "Actor-Network versus Network Analysis versus Digital Networks Are We Talking About the Same Networks?" In: *Digital STS: A Handbook and Fieldguide*. Princeton University Press, pp. 510–524. DOI: 10.1515/9780691190600-034.

Verma, Vijay and Rajesh Kumar Aggarwal (June 2020). "A comparative analysis of similarity measures akin to the Jaccard index in collaborative recommendations: empirical and theoretical perspective". In: *Social Network Analysis and Mining* 10.1, 43:1–43:16. DOI: 10.1007/s13278-020-00660-9.

Villa, Giacomo, Gabriella Pasi, and Marco Viviani (Aug. 2021). "Echo chamber detection and analysis". In: *Social Network Analysis and Mining* 11.1, 78:1–78:17. DOI: 10.1007/s13278-021-00779-3.

Vo, Nguyen, Kyumin Lee, Cheng Cao, Thanh Tran, and Hongkyu Choi (July 2017). "Revealing and Detecting Malicious Retweeter Groups". In: *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ASONAM '17. ACM, pp. 363–368. DOI: 10.1145/3110025.3110068.

Vosoughi, Soroush, Deb Roy, and Sinan Aral (2018). "The spread of true and false news online". In: *Science* 359.6380, pp. 1146–1151. DOI: 10.1126/science.aap9559.

Waldek, Lise, Brian Ballsun-Stanton, and Julian Droogan (Nov. 2020). "After Christchurch: Mapping online right-wing extremists". In: *The Interpreter*. Accessed 2022-01-31. URL: https://www.lowyinstitute.org/the-interpreter/after-christchurch-mapping-online-right-wing-extremists.

Wang, Gang, Christo Wilson, Xiaohan Zhao, Yibo Zhu, Manish Mohanlal, Haitao Zheng, and Ben Y. Zhao (2012). "Serf and Turf: Crowdturfing for Fun and Profit". In: *Proceedings of the 21st World Wide Web Conference*. WWW '12. ACM, pp. 679–688. DOI: 10.1145/2187836.2187928.

Wang, Sze-Yuh Nina and Yoel Inbar (Dec. 2020). "Moral-Language Use by U.S. Political Elites". In: *Psychological Science* 32.1, pp. 14–26. DOI: 10.1177/0956797620960397.

Wardle, Claire (Jan. 2017). “Fake news. It’s complicated”. In: *First Draft*. Accessed 2021-11-24. URL: https://firstdraftnews.org/articles/fake-news-complicated/.

Wardle, Claire (Sept. 2019a). “A New World Disorder”. In: *Scientific American* 321.3, pp. 88–93. DOI: 10.1038/scientificamerican0919-88.

Wardle, Claire (Oct. 2019b). *Understanding Information Disorder*. Essential Guide. First Draft News. URL: https://firstdraftnews.org/wp-content/uploads/2019/10/Information_Disorder_Digital_AW.pdf.

Wardle, Claire and Hossein Derakhshan (Sept. 2017). *Information Disorder: Toward an interdisciplinary framework for research and policy making*. Report DGI(2017)09. Council of Europe. URL: https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c.

Wasserman, Stanley and Katherine Faust (Nov. 1994). *Social network analysis: Methods and applications*. Vol. 8. Cambridge University Press. DOI: 10.1017/cbo9780511815478.

Waugh, Benjamin, Maldini Abdipanah, Omid Hashemi, Shaquille A. Rahman, and David M. Cook (Dec. 2013). “The Influence and Deception of Twitter: The Authenticity of the Narrative and Slacktivism in the Australian Electoral Process”. In: *14th Australian Information Warfare Conference*. AIWC ’13. DOI: 10.4225/75/57a849a9befb7.

Weber, Derek (Nov. 2019). *On Coordinated Online Behaviour*. Poster presented at the Australian Social Network Analysis Conference, ASNAC ’19. URL: https://www.slideshare.net/derekweber/on-coordinated-online-behaviour.

Weber, Derek and Lucia Falzon (July 2021). “Temporal Nuances of Coordination Network Semantics”. In: *arXiv preprint*, pp. 1–14. arXiv: 2107.02588v2 [cs.SI].

Weber, Derek, Lucia Falzon, Lewis Mitchell, and Mehwish Nasim (June 2022). “Promoting and countering misinformation during Australia’s 2019–2020 bushfires: A case study of polarisation”. In: *Social Network Analysis and Mining* 12.1, 64:1–64:26. DOI: 10.1007/s13278-022-00892-x.

Weber, Derek, Mehwish Nasim, Lucia Falzon, and Lewis Mitchell (Apr. 2020a). “#ArsonEmergency and Australia’s “Black Summer”: Polarisation and Misinformation on Social Media”. In: *Disinformation in Open Online Media*. MISDOOM ’20. Springer, pp. 159–173. DOI: 10.1007/978-3-030-61841-4_11.

Weber, Derek, Mehwish Nasim, Lucia Falzon, and Lewis Mitchell (Nov. 2020b). *Revealing social bot communities through coordinated behaviour during the 2020 US Democratic and Republican National Conventions*. Talk presented at the Australian Social Network Analysis Conference, ASNAC ’20. URL: https://www.slideshare.net/derekweber/revealing-social-bot-communities-through-coordinated-behaviour.

Weber, Derek, Mehwish Nasim, Lewis Mitchell, and Lucia Falzon (Dec. 2020c). “A method to evaluate the reliability of social media data for social network analysis”. In: *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ASONAM ’20. IEEE, pp. 317–321. DOI: 10.1109/asonam49781.2020.9381461.

Weber, Derek, Mehwish Nasim, Lewis Mitchell, and Lucia Falzon (July 2021a). "Exploring the effect of streamed social media data variations on social network analysis". In: *Social Network Analysis and Mining* 11.1, 62:1–62:38. DOI: 10.1007/s13278-021-00770-y.

Weber, Derek and Frank Neumann (Dec. 2020). "Who's in the Gang? Revealing Coordinating Communities in Social Media". In: *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ASONAM '20. IEEE, pp. 89–93. DOI: 10.1109/asonam49781.2020.9381418.

Weber, Derek and Frank Neumann (Oct. 2021). "Amplifying influence through coordinated behaviour in social networks". In: *Social Network Analysis and Mining* 11.1, 111:1–111:42. DOI: 10.1007/s13278-021-00815-2.

Weber, T.J., Chris Hydock, William Ding, Meryl Gardner, Pradeep Jacob, Naomi Mandel, David E. Sprott, and Eric Van Steenburg (Mar. 2021b). "Political Polarization: Challenges, Opportunities, and Hope for Consumer Welfare, Marketers, and Public Policy". In: *Journal of Public Policy & Marketing* 40.2, pp. 184–205. DOI: 10.1177/0743915621991103.

Williams, Hywel T.P., James R. McMurray, Tim Kurz, and F. Hugo Lambert (May 2015). "Network analysis reveals open forums and echo chambers in social media discussions of climate change". In: *Global Environmental Change* 32, pp. 126–138. DOI: 10.1016/j.gloenvcha.2015.03.006.

Woolley, S. C. and D. R. Guilbeault (2018). "United States: Manufacturing Consensus Online". In: *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media*. Oxford University Press. Chap. 8, pp. 185–211. DOI: 10.1093/oso/9780190931407.001.0001.

Woolley, Samuel C. (Mar. 2016). "Automating power: Social bot interference in global politics". In: *First Monday* 21.4. DOI: 10.5210/fm.v21i4.6161.

Woolley, Samuel C. and Philip N. Howard, eds. (Nov. 2018). *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media*. Oxford University Press. DOI: 10.1093/oso/9780190931407.001.0001.

Wu, Liang, Fred Morstatter, Xia Hu, and Huan Liu (Dec. 2016). "Mining misinformation in social media". In: *Big Data in Complex and Social Networks*. CRC Press. Chap. 5, pp. 125–152. ISBN: 9781498726849. DOI: 10.1201/9781315396705.

Wu, Tingmin, Sheng Wen, Yang Xiang, and Wanlei Zhou (July 2018). "Twitter spam detection: Survey of new approaches and comparative study". In: *Computers & Security* 76, pp. 265–284. DOI: 10.1016/j.cose.2017.11.013.

Yang, Kai-Cheng, Onur Varol, Clayton A. Davis, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer (Jan. 2019). "Arming the public with artificial intelligence to counter social bots". In: *Human Behavior and Emerging Technologies* 1.1, pp. 48–61. DOI: 10.1002/hbe2.115.

Yang, Zhao, René Algesheimer, and Claudio J. Tessone (Aug. 2016). "A Comparative Analysis of Community Detection Algorithms on Artificial Networks". In: *Scientific Reports* 6.30750, pp. 1–16. DOI: 10.1038/srep30750.

Yap, Andy J. (Mar. 2020). “Coronavirus: Why people are panic buying loo rolls and how to stop it”. In: *The Conversation.* Accessed 2020-03-10. URL: https://theconversation.com/coronavirus-why-people-are-panic-buying-loo-roll-and-how-to-stop-it-133115.

Youyou, Wu, Michal Kosinski, and David Stillwell (Jan. 2015). “Computer-based personality judgments are more accurate than those made by humans”. In: *Proceedings of the National Academy of Sciences* 112.4, pp. 1036–1040. DOI: 10.1073/pnas.1418680112.

Yu, Rose, Xinran He, and Yan Liu (Oct. 2015). “GLAD: Group Anomaly Detection in Social Media Analysis”. In: *ACM Transactions on Knowledge Discovery from Data* 10.2, pp. 1–22. DOI: 10.1145/2811268.

Yu, William (July 2021). “A Framework for Studying Coordinated Behaviour Applied to the 2019 Philippine Midterm Elections”. In: *Proceedings of the 6th International Congress on Information and Communication Technology.* ICICT ’21. URL: https://archium.ateneo.edu/discs-faculty-pubs/207/.

Zannettou, Savvas, Tristan Caulfield, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn (May 2019). “Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web”. In: *Companion Proceedings of The 2019 World Wide Web Conference.* WWW ’19. ACM, pp. 218–226. DOI: 10.1145/3308560.3316495.

Zannettou, Savvas, Tristan Caulfield, Emiliano De Cristofaro, Nicolas Kourtelris, Ilias Leontiadis, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn (2017). “The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources”. In: *Proceedings of the 2017 Internet Measurement Conference.* IMC ’17. ACM, pp. 405–417. DOI: 10.1145/3131365.3131390.

Zhang, Albert (July 2021). “#StopAsianHate: Chinese diaspora targeted by CCP disinformation campaign”. In: *The Strategist.* URL: https://www.aspistrategist.org.au/stopasianhate-chinese-diaspora-targeted-by-ccp-disinformation-campaign/.

Zhang, Yizhou, Karishma Sharma, and Yan Liu (Oct. 2021). “VigDet: Knowledge Informed Neural Temporal Point Process for Coordination Detection on Social Media”. In: *Thirty-fifth Conference on Neural Information Processing Systems.* NeurIPS ’21. URL: https://proceedings.neurips.cc/paper/2021/file/1a344877f11195aaf947ccfe48ee9c89-Paper.pdf.

Zhao, Qingyuan, Murat A. Erdogdu, Hera Y. He, Anand Rajaraman, and Jure Leskovec (Aug. 2015). “SEISMIC: A Self-Exciting Point Process Model for Predicting Tweet Popularity”. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* KDD ’15. ACM, pp. 1513–1522. DOI: 10.1145/2783258.2783401.