A Network Science Framework for Detecting Disruptive Behaviour on Social Media

A thesis submitted in partial fulfilment

of the requirement for the degree of Doctor of Philosophy

James R. Ashford

December 2022

Cardiff University
School of Computer Science & Informatics

Declaration

This work has not previously been accepted in substance for any degree and is not concurrently submitted in candidature for any degree.

Signed ..... (candidate)

Date .....

Statement 1

This thesis is being submitted in partial fulfilment of the requirements for the degree of PhD.

Signed ..... (candidate)

Date .....

Statement 2

This thesis is the result of my own independent work/investigation, except where otherwise stated. Other sources are acknowledged by explicit references.

Signed ..... (candidate)

Date .....

Statement 3

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed ..... (candidate)

Date .....

Copyright © 2022 James Ashford.

Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled “GNU Free Documentation License”.

To my wife, Ella,
for her patience and support throughout the PhD.

Abstract

Social networking platforms enable individuals to interact with others in a public forum by creating and/or consuming both written and visual content. Due to the popularity and wide-spread adoption of social media, this has led to unforeseen negative consequences where actors use social media to intentionally disrupt normal discourse to subversively influence individuals or groups. As a result, this leads to the problem of detecting anomalous activity, which is challenging due to large quantities of textual information combined with multimedia. Furthermore, this is compounded by issues such as foreign languages. This motivates research into techniques that can detect anomalies in social media activity through language-agnostic approaches.

This thesis examines ways in which this can be achieved through network science, using different forms of networks to represent the behaviour of actors in social media, rather than the specific content they have produced. However, diverse affordances on alternative social media platforms make this a complex problem. This thesis examines three alternative classes of network representation with respect to detecting disruption in social media. We examine these representations using techniques from complex network theory. Using a range of social media systems, this thesis provides evidence that network-based signals aligning to disruptive behaviours can be detected for alternative forms of social media engagement (e.g., collaboration, message, community and feed-based interactions). Through this approach, this thesis determines prospects for assessing social media in complex and dynamic scenarios without recourse to processing natural language.

Contents

List of Figures

List of Tables

Acknowledgements

I want to take this opportunity to thank my close friends and family for their support in helping me complete this thesis. In particular, I would like to thank my wife, Ella, for her constant support, encouragement and patience in helping me get through the challenging moments of my PhD and for always being prepared to listen. I would also like to thank the members of Emmanuel Baptist Church, Cardiff, for always showing an interest in my work and looking out for my well-being when I needed support.

Throughout my time as a PhD student and a member of the Security, Crime and Intelligence Innovation Institute, I've got to know some incredibly talented people. I want to thank members of the DAIS-ITA program for providing me with many opportunities to collaborate with people both nationally and internationally. I'm grateful to have been part of a program that has provided me with more options than most PhD students, and without the members of the DAIS-ITA program, there is no way I would've gained the confidence and support to pursue a career in research. I'd also like to thank my colleagues within the SCIII team. It has been a real privilege getting to know and work with them over the past few years under the OSCAR project.

Finally, and most importantly, I thank my supervisor, Professor Roger Whitaker, for his support and encouragement in helping me complete this PhD. My journey as a PhD student hasn't always been as straightforward as I thought it would be, but Professor Whitaker has always been incredibly supportive and has helped me every step of the way to complete this thesis. I'd also like to thank Professor Alun Preece and Dr Liam

Turner for their much-appreciated wisdom, insights and contributions throughout this work. I could not have asked for a better supervisor and supervision team.

Chapter 1

Introduction

Online social media platforms provide multiple mechanisms that support alternative forms of user interaction. Since their introduction in less than two decades, their adoption has been viral, fuelling a global revolution in social communication for all sectors of society. As of 2021, it is estimated that a total of 3.7 billion users worldwide are using some form of social media and is expected to increase to 4.41 billion users by the year 2025 [3]. Key platforms include Facebook (+2,960M active users[1], as of 2021), Instagram (+1,000M active users), Twitter (+330M active users) and Reddit (+430M active users).

Typically, the primary form of communication involves exchanging text-based messages or multimedia such as images and GIFs. Beyond this, platforms such as Facebook and Twitter also provide a means for users to share content produced by another user, expression of support for content (e.g., 'likes') while other platforms such as Reddit and YouTube allow users to comment on a particular topic or item. This results in a range of affordances for the user, dependent on the platform.

Unlike traditional communication networks such as email and instant messaging, social media platforms are also public-facing, where users interactions are often public, in a shared online space when behaviour, arguments and ideas can be openly observed. This has the power to influence large audiences who can also participate in the conversation. Social media platforms can also be used to create and support informal communities,

[1] https://investor.fb.com/investor-news/press-release-details/2022/Meta-Reports-Third-Quarter-2022-Results/default.aspx

where often like-minded people or those with common views on an issue can participate and interact with each other, aligned to forces of homophily, resulting in in-group formation.

While social media platforms are clearly positive in connecting people, the affordances they provide, combined with massive participation, can readily be used to cause disruption or potential harm on a large scale. This occurs through cultural influence - the social transmission of content that affects human behaviour in the offline world.

There have been numerous significant instances of this in recent years. For example, the 'Stop the Steal' movement is a recent example of this where communities mobilised online to coordinate the disruption of the processing of the 2020 presidential election results, the consequences of which resulted in terrorism and insurrection. This took place over several online social networks using a combination of Facebook groups [143] and Twitter [363]. Similar comments can be made regarding the use of disinformation during the 2016 presidential election [14] and Brexit (The UK's withdrawal from the European Union) [182]. Consequently, there is an increasing need to identify such forms of technology misuse in a new context - something for which the world has been ill-prepared[2][3].

1.1 Problem Definition and Approach

The focus of this thesis is on *the problem of detecting disruptive behaviour in social media*. In the context of this thesis, "disruptive behaviour" can be defined as activity, which takes place on social media, in which a user (or group of users) seek to interrupt the normal proceeding of events which goes against the norms of the platform. Within

[2]How Were Social Media Platforms So Unprepared For 'Fake News' And Foreign Influence? : `https://www.forbes.com/sites/kalevleetaru/2019/07/09/how-were-social-media-platforms-so-unprepared-for-fake-news-and-foreign-influence/?sh=3e48797345cb`

[3]When Social Media Is Really Problematic for Adolescents: `https://www.nytimes.com/2019/06/03/well/family/when-social-media-is-really-problematic-for-adolescents.html`

this thesis, a few examples of disruptive behaviour (see Chapter 2 for more) include trolling and disinformation.

Detecting disruptive behaviour is complex for several reasons. Firstly, assessing the context for disruption (this includes the user's language, and use of images and multimedia etc.) is a challenge due to the scale of social media platforms and the volume of data that they produce. For example, directly analysing content could require subjective assessments, such as for images, combined with advanced natural language processing (also known as NLP) to extract meaning. Secondly, it is important to recognise that content is not restricted to a particular language. Solutions that are language-agnostic will mean that effective techniques for international deployment can be achieved with relative ease. Thirdly, different social media platforms offer different affordances for users. This means that actors can disrupt behaviour in different ways depending on the type of social media used [394, 53, 364]. While metadata (such as timestamps, keywords, etc.) and raw content can be used as features for detecting disruptive behaviour, relatively little work has considered the role of users, their relationships and choice of affordances.

These problems motivate new and novel techniques to support the detection of disruptive behaviour across different social media platforms in a language-agnostic manner. Accordingly, this thesis adopts a general strategy of *assessing the behaviour of social media users as opposed to directly analysing the content they produce*. Although similar approaches to achieving this have been introduced in a limited context, there remains a considerable gap within the literature (see Chapter 2). This is because additional social media platforms have emerged with new affordances. These offer different functionality though which interactions can take place. Consequently, this thesis introduces alternative representations for user behaviour based on different user affordances provided by alternative social media platforms, as presented in Chapter 3.

To conveniently represent different forms of user behaviour, this thesis focuses on network-based representations, which we call *behavioural networks*. These are derived

from the social media activity of users (see Chapter 3). The edges and nodes of behavioural networks can take on many forms and can represent different types of interactions depending upon the affordances of the underlying social media platform. Based on the functionality of mainstream social media platforms, this thesis proposes three types of behavioural network representation that capture significant user affordances: *transitional* (see Chapter 4), *user-to-user* (see Chapter 5) and *user association* (see Chapter 6). In doing so, these network-based representations provide a principled approach to test the hypothesis concerning the identification of disruption through the actions of users, rather than the specific content of social media.

While behavioural networks offer representations to capture different aspects of user interaction with each other and social media content, methods are also required to assess and compare behavioural networks to establish anomalous activity. Consequently, some form of social network analysis techniques are required to capture the characteristics of behavioural networks. However, given the mass usage of social media [1, 2, 179] and its dynamic nature, it is useful to consider approaches that are aligned to complex networks. [54, 124].

Accordingly, using techniques from complex networks, this thesis investigates how the under and over representation of induced substructures in behavioural networks can potentially signal anomalous user behaviour aligned to disruption. This includes a combination of both motif [263, 264] and subgraph analysis. This approach builds on considerable theory from complex systems and is well-suited because the induced substructures represent signatures aligned to patterns of small-scale human interaction (i.e., consideration of a user's neighbourhood and their surrounding interactions).

1.2 Hypothesis and Research Questions

The challenges outlined in Section 1.1 highlight the challenges associated with the detection of disruptive activity. These motivate the following hypothesis:

***Hypothesis**: Anomalous activity related to conflict or disruption in social media can be detected through the construction and analysis of networks representing different types of user behaviour and interaction, based on alternative affordances provided by social media.*

This thesis is dependent upon the computational methods that are required to extract, model and evaluate networks derived from social media platforms. To explore the hypothesis, a series of research questions are produced:

Research Question 1. *How can behavioural networks be defined from activity on social media platforms?*

Online social networks can be represented many different ways. For example, platforms such as Twitter allow users to interact with other users by means of sharing, liking, mentioning and replying. Although these interactions can be represented in the same way, each type exhibits different network structures, which in turn may result in different types of affordances to appear.

Chapter 3 addresses this research question by outlining the ways in which social media platforms structure their environments and provides an overview for capturing user affordances on as many platforms as possible.

Research Question 2. *Is it possible to produce a concise framework of alternative network-based representations capturing alternative affordances provided across social media platforms?*

Social media platforms encourage different types of interaction, which result in different reactions [309, 139, 168, 310]. For example, the differences between Facebook and Twitter can affect the way in which conversations are preformed [310]. Similar observations can be made, ranging from the way entrepreneurs and businesses engage with the larger audiences on Twitter [139] to users of online marketplaces discussing and recommending products [309].

The work produced in Chapters 4, 5 and 6 help address this research question by using real-world social and behavioural networks extracted from various social media platforms. In doing so, these chapters provide unique insights into scenario-specific situations where many network representations are used to capture different variations of conflict or disruption.

Research Question 3. *How can diverse affordances and the interactions they facilitate be represented?*

Different social media platforms can be used in different ways. Consequently, different user affordances can be extracted from a platform using different representations. This issue is mostly addressed in Chapter 3 by considering the ways in which individuals use social media. For example (see Chapter 4) Wikipedia is widely used for collaborating with others indirectly to improve the quality of articles. This is different from Reddit or Twitter (see Chapter 5) where the platforms are designed to encourage direct user-to-user interaction.

Research Question 4. *To what extent can local features (i.e., subgraphs) provide signalling, on which prediction of disruptive behaviour be can be made?*

The ability to detect disruptive activity on social media using language-agnostic network representations with little computational overhead is highly desirable. To investigate this, Chapters 4, 5 and 6 address this research question by using well-established binary classification algorithms to assess predictive utility. These classifiers include: support vector machine (SVM), binary logistic regression (BLR) and random forest classifier (RFC).

Local network features are used to better understand structural properties by examining smaller subcomponents, which are used to form a detailed profile of the network. In addition to this, it may be necessary to explore the role of global features such in/out degree, transitivity, density and reciprocity where possible (as seen in Chapters 5 and 6) as part of a subtask to examine which network features perform best.

1.3 Thesis Structure

This thesis is structured around the ways in which user activity can be modelled through online social networking platforms. This can be broken down into broken down into the following chapters.

- **Chapter 2: Related Literature**: This provides a broad overview of key areas and disciplines surrounding the problem space and presents a gap within the literature regarding the utility of behavioural networks and complex approaches for detecting disruptive activity.
- **Chapter 3: Characterising Diverse Functionality in Social Media**: This chapter explains the many ways in which social media platforms allow users to interact and engage with their service and identifies the need for multiple network representations for capturing user affordances, which addresses issues relating to **Research Questions 1, 2 and 3**. The network representations introduced are examined in the following three chapters.
- **Chapter 4: Transitional Networks**: This chapter introduces the concept of transitional networks as a means to identify disruptive behaviour by using data collected from Wikipedia and Reddit. This chapter introduces methods for generating a network representation from time-series data. This results in the ability to detect controversial and non-controversial articles based upon the structure of the revision history using network motif analysis. As a result, this contributes to **Research Question 2** by using a network-based approach for modelling user affordances though switching behaviour and **Research Question 4** by classifying controversial and non-controversial articles using network-based features.
- **Chapter 5: User-To-Users Networks**: This chapter studies multiple ways in which two users can interact with each other. Both Twitter and Reddit are used to demonstrate different types of message-based interactions. In addition to

this, induced subgraphs are used as feature vectors for classifying disruptive and non-disruptive activity from the perspective of individual user accounts and their local neighbourhood (egocentric networks) using Reddit and larger networks (composed of retweets, replies and mentions) derived from various topics on Twitter. By considering these user-to-user interactions, this supports **Research Question 2** by contributing alternative representations for particular types of affordance and **Research Question 4** by detecting the presence of disruptive behaviour though massage-based network interactions.

- **Chapter 6: User Association Networks**: This chapter uses bipartite networks to model the association between a user and a topic or activity. Association networks and substructures capture two features; a user's varied interests and similarity with other users. This chapter uses Reddit to generate association models based upon the relationship between user and subreddits they post in. Induced graphlets are used to characterise communities associated with potential for misinformation relating to the COVID-19 pandemic and other topics. This also contributes to **Research Question 2** by providing a novel approach for embedding user affordances based upon simple bipartite connections and **Research Question 4** by predicting whether a community has the potential for misinformation to emerge according to the presence of bipartite graphlets.

- **Chapter 7: Conclusions and Future Work**: This concludes the thesis by reflecting on the contributions made with respect to the research questions defined in the introduction and providing additional insights. This is a followed by some additional thoughts with respect to future work and real-world applications.

1.4 Thesis Contributions

By addressing the research questions defined above, this thesis offers three unique contributions: a *modelling framework*, analysis of *behavioural networks* and functionality

to support possible applications for *content moderation*.

Modelling Framework: This thesis presents a framework for three alternative representations of user behaviour for different social media platform affordances. These representations are referred to in the thesis as *transitional*, *user-to-user* and *association*. These representations are tested using data collected from Reddit, Wikipedia and Twitter.

As far as we can establish, little research has considered an overall framework for analysing different social media interactions through a network representation. As a result, this thesis attempts to contribute to this issue by demonstrating the utility of network representations on multiple platforms.

Behavioural Networks: Capturing and understanding user behaviour online is of great significance to this thesis. As mentioned previously, disruptive behaviour is exposed in multiple forms and appropriate methods are required for capturing such behaviour. This thesis addresses the value of using a network-based approach for analysing human behaviour.

Content Moderation: The results of this thesis are particularly valuable as a precursor for content moderation in an online environment to combat disruptive activity. For example, issues such as groups of users producing disruptive behaviour can be hard to identify, as moderators have a very limited view of their interactions. This thesis attempts to resolve this issue through methods to process data and present results in such a way that it can help moderators to reduce (or even ban) such interactions in the future.

1.5 List of Publications

1.5.1 Substantial Contributions

This thesis is based upon the following peer-reviewed publications:

- [22]: James Ashford, Liam Turner, Roger Whitaker, Alun Preece, Diane Felmlee, and Don Towsley. Understanding the signature of controversial Wikipedia articles through motifs in editor revision networks. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 11801187, 2019.

- [23]: James R Ashford, Liam D Turner, Roger M Whitaker, Alun Preece, and Diane Felmlee. Assessing temporal and spatial features in detecting disruptive users on Reddit. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 892896. IEEE, 2020.

- [24]: James R Ashford, Liam D Turner, Roger M Whitaker, Alun Preece, and Diane Felmlee. Understanding the characteristics of COVID-19 misinformation communities through graphlet analysis. *Online Social Networks and Media*, page100178, 2021.

1.5.2 Collaborative Contributions

The results from this thesis contributed to the research of the following publications:

- [110]: Cai Davies, James Ashford, Luis Espinosa-Anke, Alun Preece, Liam Turner, Roger Whitaker, Mudhakar Srivatsa, Diane Felmlee. Multi-scale user migration on Reddit, *AAAI*, 2021

  A paper which utilises data influenced by the findings of the investigation on COVID-19 using subreddits with the potential for containing misinformation (see

Chapter 6, Section 6.4). The data is used to observe multiple levels of migration patterns across different subreddits.

- [254]: Cassie McMillan, Diane Felmlee and James R Ashford. Reciprocity, transitivity, and skew: Comparing local structure in 40 positive and negative social networks. *Plos one*, 17(5):e0267886, 2022.

  Used as part of an investigation for assessing positive and negative ties in social networks which makes use of the same methodology for generating Wikipedia revision networks using techniques derived from transitional networks (see Chapter 4, Section 4.4).

- [235]: Eunjin Lee, James Ashford, Malgorzata Turalska, Liam Turner, Vera Liao, Rachel Bellamy, Geeth de Mel, and Roger Whitaker. An exploratory analysis of suspicious Reddit user accounts based on sentiment and interactions, 2019.

  A study which makes use of the same dataset used to investigate suspicious user accounts on Reddit (see Chapter 4, Section 4.5).

Chapter 2

Related Literature

This thesis brings together a number of academic areas that are used to help develop new approaches and functionality to support social media analysis. It is therefore useful to identify related literature and key concepts in these areas. We begin with the concept of the internet.

The use of the internet has revolutionised the way people communicate with each other on a large scale. The introduction of the Web 2.0 and social media platforms made it possible to establish contact with others using different forms of communication (e.g. instant messaging, microblogging and bulletin boards). As a result, social media has become a ubiquitous tool to facilitate the spread of information and has become a fundamental component to the way society communicates to the point of becoming dependent on it [382, 184].

While social media has provided many positive contributions to society, as of recent developments, one of the unforeseen consequences of this is to cause disruption. Disruption can be defined as *the action of preventing something, especially a system, process, or event, from continuing as usual or as expected*[1]. In the context of this thesis, this definition can be extended to include interactions taking place on social media.

Issues such as "fake news" are a form of disruptive behaviour whereby groups of users seek to (deliberately or inadvertently) deceive others through by spreading misleading or false narratives which is largely driven by the inattention of others, resulting polarisation

[1] `https://dictionary.cambridge.org/dictionary/english/disruption`

and confusion [305]. Data from Google Trends[2] (see Figure 2.1) reveals how the search term "fake news" began to make an appearance towards the end of 2016 due to the vast quantity of false stories shared during the 2016 US Presidential Election [14].

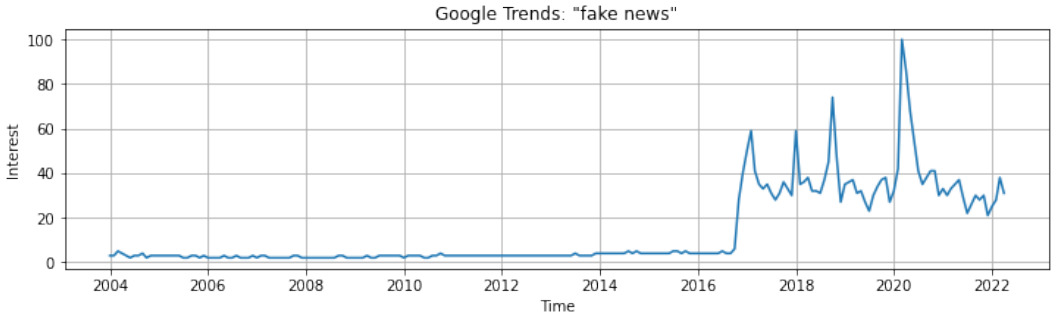

Figure 2.1: Appearance of the search term "fake news" used on Google over time between 2004 and 2022 shows a distinct peak in activity towards the end of 2016.

Fake news has since accelerated rapidly due to events such as the 2016 US Election 2016 and Brexit (explained further in Sections 2.1.2 and 2.1.3 respectively) as explained by the rapid increase in interested in Figure 2.1 around the end of 2016 [328, 305].

These results indicate that public perceptions of fake news and their disruptive consequences have increased over time. As a result, this is also reflected in considerable upsurge in academic study and motivates a need to study this issue further. Most studies focus on using textual/language-based solutions, where major research has gone into advancing techniques such NLP [295, 112], video/image processing [7, 313, 79] and other techniques using multi-modal data [414, 215].

Disruptive behaviour has become a widespread issue and was largely unforeseen by the creators of social media platforms, as their intentions were to provide communication functionality. As a result of these consequences, this has led to more research into possible computational solutions (such as detection methods) to prevent (or, at least, minimise) disruption from taking place in future events.

This gives rise to new and unique challenges for research with respect to analysing user activity in an online environment. Woolley and Howard of the Oxford Internet Insti-

[2]Google Trends: `https://trends.google.com/`

tute have raised awareness of this issue in their book *Computation Propaganda* where they have documented many instances where disruptive behaviour has been observed worldwide [396]. This includes some of their more theoretical research surrounding misinformation [181, 219], trolls [312, 60], social bots [187] and echo chambers [121, 123] (see Section 2.2). In addition to the theoretical work, research perform by Pennycook and Rand seek to understand the psychological aspects of how and why some people fall to issues such as fake news [305, 307, 306]. Consequently, these issue have led to practical implications whereby researchers at OSoMe (The Observatory on Social Media, Indiana University) have developed algorithms for detecting social bots [356, 111], hoaxes [336] and misinformation [115] on Twitter.

Due to the scale of social media and the internet, disruptive behaviour had become a global issue which needs addressing [188, 318]. As a result, computational solutions are needed to detect instances of disruptive behaviour appearing in near real-time. Given the vast geographical distribution of social media users, issues surrounding spoken language and location begin to emerge.

This chapter considers the use of social network analysis as an alternative solution to help aid our understanding of user behaviour from the perspective of their interactions and relationships with other users and entities. Network representations of online social networks are extremely useful because once the representation is established, network-based techniques can be exploited as successfully harnessed in many other areas of research [176].

This review provides an overview of key literature surrounding the applications of disruptive behaviour analysis using multiple social media platforms. In addition, this related literature considers the role of complex networks and their characteristics for extracting predictive signals within a large social network.

In order to understand the literature surrounding these issues, this review is broken down into the following sections:

- **Section 2.1: Social media being used for disruption**. A high-level overview of how social media usage has been exploited in recent real-world events.
- **Section 2.2: Types of disruptive behaviour on social media**. Important forms of disruptive activity within the literature are presented alongside proposed computational solutions.
- **Section 2.3: Modelling interactions of user activity**. Consideration of the different variations of social network configurations used within the context of behavioural analysis.
- **Section 2.4: Understanding the role of complex networks**. A summary of the characteristics of complex networks and their applications for detecting disruptive behaviour on social media.
- **Section 2.5: Conclusions**. An overview of the literature presented with reference to gaps within the research.

2.1 Social Media Being Used for Disruption

In recent years, there have been numerous situations in which social media platforms have had an impact on the outcome of a particular event or in the way in which people behave in an online and offline setting. This is primarily caused by a combination of disruptive behaviour (see Section 2.2) and the way in which users make use of social media (see Chapter 3). In almost all of these instances, it was simply unknown to the public the magnitude of how disruptive these events were at the time, which were unforeseen by the creators of these platforms.

The purpose of this section is to highlight how disruptive behaviour emerged in recent events, where social media had a significant contrition in the development of the events. These include **Arab Spring**, **Brexit**, **2016 US Presidential Election** and **COVID-19** - some of which have been featured in greater detail throughout this thesis. The incidents

surrounding these events have demonstrated how coordinated influence operations and disruptive behaviour can be used to manipulate groups of users.

2.1.1 Arab Spring

One of the first occasions in which social media was used to cause disruption was during the developments of the "Arab Spring". Social media significantly contributed to what eventually led to multiple demonstrations and protests taking place. The role of social media in this context was twofold - to communicate and mobilise.

Social media played a significant role in allowing actors to communicate both locally and globally. This was primarily achieved through the use of Twitter, Facebook and various online blogs to vacillate a national and international conversation [71, 186, 174]. Furthermore, this allowed journalists and politicians alike to use social media as a means for communicating to protesters, as well as establishing a global audience [10, 395, 100]. Secondly, the use of social media allowed actors to organise and mobilise both anti-Government and pro-government protests offline at scale [353, 287]. The global connectivity of social media meant that activists could coordinate protests globally in an attempt to gain worldwide support and coverage [345].

The Arab Spring stands as an example of how social media was used as a means to seek positive outcomes (in this case, freedom and democracy) through mass communication. As a result, issues such as misinformation were overshadowed by individuals using social media as a mechanism to promote a clear message. While this event may not necessarily fit this chapter's definition of disruption, it does however serve as an example of how social media supports mobilisation which can have a negative impact (as seen in future events).

2.1.2 Brexit

The Brexit (United Kingdom's withdrawal from the European Union) referendum saw wide-spread disruption on multiple social media platforms. The most notable instance of this took place on Twitter, using a network of highly-coordinated fake user accounts used to spread disinformation believed to have been performed by Russian state actors [155]. This process involved using fake accounts to influence public opinion and drive division though the use of elaborate echo chambers [155, 172] combined with microtargeting users to enhance political polarisation [333].

This was achieved by provoking users with content designed to trigger certain reactions in favour of a particular cause. One way to achieve this was through the use of trolling by creating a hostile environment for others, disrupt public debate and to damage the reputation of others [218, 38]. Research has shown the impact of these coordinated campaigns on trends as they develop over time [25]. As a consequence, Brändle et al. observed how Brexit (as well as the after effects) resulted in the "politics of division" where individuals are heavily conformed to either a "Remain" or "Leave" position [63].

2.1.3 2016 US Presidential Election

The 2016 US Presidential election saw similar disruptive effects playing out during Brexit, however, operations were more widespread covering dominant platforms including Twitter, Facebook and Reddit to list a few [56]. Much like Brexit, the use of mass coordination of fake accounts and disinformation played a significant role in the election which resulted in heavily polarised "filter bubbles" emerging [161], however, this was further amplified by the use of exclusive microtargeting to reach specific groups of people based upon their personality traits [385, 36].

To begin, Twitter was primarily used as a communication tool to allow campaigners to reach a large audience. This was a fundamental component to support the Trump campaign and was used to generate a reaction out of their audience [142, 250]. Using

social media in such a way was considered rather unconventional and "amateurish" compared to other political campaigns [127]. This behaviour was compounded by a small minority of users (around 1%) who were responsible for sharing 80% of fake news, which amounted to around 6% of all news articles consumed on Twitter [160]. Consequently, this resulted in heavily polarised groups of users among those who identify as Trump supporters [383, 145].

Secondly, Facebook further contributed to disruption through the spreading of fake news and microtargeting. During and prior to the election, fake news was circulating around Facebook, having a powerful network effect where favourable pro-Trump narratives were shared as many as 30 millions times [163, 14]. Furthermore, Facebook's ad services were abused in such a way that their algorithms were exploited in an attempt to maliciously microtarget individuals who are likely to engage politically [319]. A small proportion of these ads were later identified to be of Russian origin [216]. This had the effect of reinforcing a user's existing beliefs making them less likely to change their voting intentions [241]. In response, Facebook identified that their platform and services were being used to form complex influence operations to propagate fake news [390].

Finally, Reddit was used to cause disruption, although it had far less of an effect in comparison to Twitter and Facebook. Given that Reddit as a platform is more community oriented, this lead to the formation of echo chambers emerging within subreddits. This was evident through the r/the_donald subreddit where pro-Trump supporters gathered. Research has shown that members of r/the_donald didn't participate as much in other communities, thus forming echo chambers [164] which, in turn, was further amplified by the presence of bots [191]. These echo chambers were identified by highly interconnected groups of users within more right-leaning subreddits, which are typically characterised with ties that are strong in homophily [346, 252].

2.1.4 COVID-19

As of this writing, the ongoing COVID-19 pandemic has been the subject of multiple forms of disruption. This typically comes in the form of misinformation (e.g. regarding vaccinations) and hateful content (e.g. racism in relation to the Chinese origin of the virus).

The issue of misinformation and conspiracies has had a significant role surrounding the developments of the COVID-19 pandemic [298, 8, 294]. This had a widespread presence on many social media platforms including Twitter, Instagram, YouTube, Reddit and Gab in which the spread of misinformation rapidly accelerated over time [96, 69, 67]. As a consequence, this can have a negative impact on both mental and physical well-being [362].

In particular, misinformation surrounding anti-vaccination narratives have dominated conversations on Twitter [148]. These are often directed towards a community of influential accounts such as health workers and policy-makers, using persuasive and emotive language in an attempt to shift opinion [266].

In addition to misinformation, social media has also been used to promote trolling and hurtful content commonly described as hate speech [77, 177]. Platforms such as Twitter have seen a rise in anti-Asian hate speech, using terms such as “Chinese virus” and other derogatory terms [177, 134, 415].

2.1.5 Summary

To summarise, the events introduced in this section document important examples regarding how social media can facilitate disruption surrounding real world events. It is evident that disruption on social media has evolved to become more aggressive over time. For example, the Arab Spring movement demonstrated how social media can be used to as a communication channel for politicians, journalist and protesters alike. Since

then, recent events such as the 2016 election and Brexit have revealed how users seek to spread false narratives and cause division among other users. Brexit, 2016 Election and COVID-19 all serve as examples for how misinformation and "fake news" have become some of the most widely used methods for causing disruption and having an effect on offline events.

2.2 Types of Disruptive Behaviour on Social Media

In view of the events studied in Section 2.1, disruptive behaviour can occur in numerous ways across different social media platforms. This section presents particularly significant forms of disruptive activity found within the surrounding literature, which are often found within many social media platforms.

2.2.1 Social Bots

Social bots are automated social media accounts designed to mimic real users at a much faster pace. A social bot has been described as "*a computer algorithm that automatically produces content and interacts with humans on social media, trying to emulate and possibly alter their behaviour.*" [136]. These automated accounts are often used to divert discussions and the opinions of users who use social media as a platform for news digest [136] in an autonomous or semi-autonomous manner [159].

Research has been performed to predict automated accounts though a combination of temporal and network-based features [137, 111, 191, 166]. Furthermore, a network-based solution provides a robust framework for modelling agent behaviour [146]. For example, research performed by Hurtado et al. reveal significant evidence that weighed edge ties accurately characterise bot-like activity within a political discussion used by labelled bot accounts [191]. Furthermore, these bot accounts heavily occupy social networks such that they produce highly connected clique-like interactions with other

users which are identified through the use of k-core decomposition [396, 35]. Research by Barberá et al. reveals how a directed network of retweets on Twitter reproduces this behaviour, where the central clique serves as a highly-connected and influential component of the network [35]. Each layer of the core demonstrates how interactions have transitioned from a central clique which cascades and branches out to other users, depicting the process of political recruitment.

2.2.2 Trolling

The act of trolling evolves using highly emotive responses to provoke users in an online social setting. Trolling can be observed within a social network though the use of discussion threads, to get a user or group to respond emotionally, either out of amusement or to obtain a desired reaction [86]. Furthermore, trolling can be used as a strategy for causing political upheaval between groups of users [141]. This form of behaviour is a major cause for concern and has the potential for spreading misinformation [113].

Motivation surrounding the understanding of troll-like behaviour has been an active area of research. Existing work has attempted to address this issue by considering the online environment combined with mood and discussion context [88, 297]. A case study performed on Reddit alludes to the utility of applying textual-based analysis in an attempt to understand troll-like behaviour through conversation and comment dialogues [260]. A network-based solution has been developed using Twitter data by modelling a network of user interactions through retweeting behaviour, composed of users who retweet another user's content [354]. These results concluded that trolls are characterised by large bursts of activity which are positioned within the top percentiles of retweets.

2.2.3 Brigading

Brigading (otherwise known as vote brigading or web brigading) is a form of disruptive activity where a group of users deliberately attempt to distort a result or outcome by means of mass coordinated participation. One of the leading and most effective strategies used in brigading involves voting with the intention of making content more popular or less popular, contrary to the opinion of the community. This form behaviour can be used to determine the appearance of a particular outcome, making the final result unreliable and false.

Platforms such as Reddit and various other Q and A platforms make use of various voting mechanisms as a metric for crowdsourced opinions and popularity to the norms of a given community or topic. Research has observed how political and apolitical discussion threads online have been subject to voter manipulation, which, in turn, affects the visibility of certain items [82, 200]. Work performed by Jeong et al. revealed how sequential patterns of user activities provide a strong characteristic of coordinated behaviour and vote manipulation [200].

As well as exploiting voting features, similar reactions have been observed through the use of cross-community interactions with the intention of causing conflict [226]. Conflicts are modelled by generating a network of interactions composed of negative ties, where one community negatively describes another based upon sentiment and opinion. Their networks can be partitioned with relative ease, as users are likely to reveal characteristics of loyalty [170]. This solution makes the task of projecting the possibility of future conflict much easier to predict.

2.2.4 Echo Chambers

An echo chamber is an environment where a user is likely to encounter information which supports and reinforces their common beliefs. It has been observed that echo

chambers reveal signs of preferential attachment where users form a small highly-connected inward group of isolated users where information can be exchanged freely between members with little or no external influence [164, 70]. Using social network analysis, this behaviour is characterised by a combination of high reciprocity [57] and homophily [97]. Research has revealed how echo chambers have a negative impact on diversity of opinion, with the potential of increased political polarisation [34, 199].

Echo chambers have been studied in the context of being a leading contribution of misinformation and false ideas spreading online [331]. Significant evidence reveals how weak network ties connected to echo chambers facilitate the propagation of misinformation. [188, 371]. This effect is observed through the notion of a "wildfire" where the echo chamber initialises the wildfire which rapidly expands as more users participate, which produces a cascading network effect [389, 371].

2.3 Modelling Interactions of User Activity

Navigating though social networks using online environments are often complex and can be represented in many forms depending upon the setting. In its simplest form, social networks are thought of as a collection of user nodes who share interactions with each others through an associated edge. Depending upon the setting, the semantics of edges and nodes can vary significantly based upon the behaviour of users and the surrounding application. This section is devoted to outlining a few of these networks which have a significant presence within the supporting literature.

2.3.1 Collaboration

Collaboration networks are composed of users who work and co-operate together to achieve a goal or complete a task collectively [281]. Furthermore, applying a network-based approach can be used to gain significant insights towards understanding how users

behave in certain environments. The study of collaborative networks has a significant presence within the literature [278, 279, 281]. However, with the introduction of the internet, the use of modern technologies has helped to facilitate co-operation of large groups of people. Using widely available resources, collaboration networks provide significant value for allowing researchers to discover new collaborators based upon their position in the network [375, 299, 245].

Although collaboration networks are best represented as a bipartite network of authors and items [400, 398, 314] alternative network representations can be produced though projected mapping [279]. Figure 2.2 presents an example bipartite network composed of items (A-E) and authors (1-6) with three network projects.

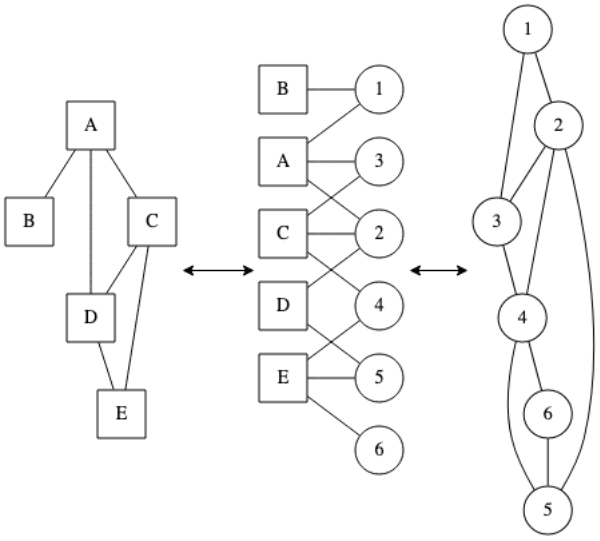

Figure 2.2: Bipartite networks can be used to project interactions between different types of entity and infer connections. From left to right, a network of linked items based upon mutual authors, a bipartite network mapping items to authors and a network of co-authors based upon mutual items in which they have collaborated.

In an academic setting, simple collaboration networks are constructed by modelling papers and authors as nodes, where a paper has many authors and an author has many papers. Although this produces a bipartite representation, this can be used to infer connections to link users together based upon mutual papers [299, 245]. Research produced by Newman et al. used a similar approach by mapping co-authors of papers

together [279].

A considerable body of research is dedicated towards using online crowdsourcing as a mechanism for collaboration. This is behaviour is observed though simple Question and Answer (Q&A) forums [5] and with more-sophisticated collaborative services such as Wikipedia[3] [315]. These platforms expose the idea of the wisdom of the crowd [359] where users collectively work to contribute towards building knowledge and developing a solution. Research performed by Wu et al. make use of bipartite networks to model collaborations by linking editors and Wikipedia articles where an edit has occurred [400, 398, 314]. It is revealed that distinct network structures correlate directly with article quality.

2.3.2 Informal / Social Networks

Online social media platforms provide wide-spread access to friendship groups and offers basic functionality where users can share, comment and exchange messages with other users. These interactions are explicitly considered however less obvious interactions such as "liking" another user's content are less obvious and often implicit in network representations, or not considered.

Platforms such as Facebook, Twitter and Reddit facilitate these interactions between friends and other users alike. These platforms allow users to produce clear interactions with other users and serves as the basis for research surrounding behavioural analysis. A basic example can be produced by modelling discussion threads over a particular item or topic [256, 391, 392]. Figures 2.3 and 2.4 present an example of two network representations for modelling discussion using a tree (see Figure 2.3) with time-constrained interactions and a reply network (see Figure 2.4) composed of directed replies between users. A discussion tree may contain multiple appearances of a single user but provides discussion depth whereas a reply network flattens interactions to focus on the dynamics

[3]Wikipedia: `https://wikipedia.org`

between users as opposed to their position in a discussion.

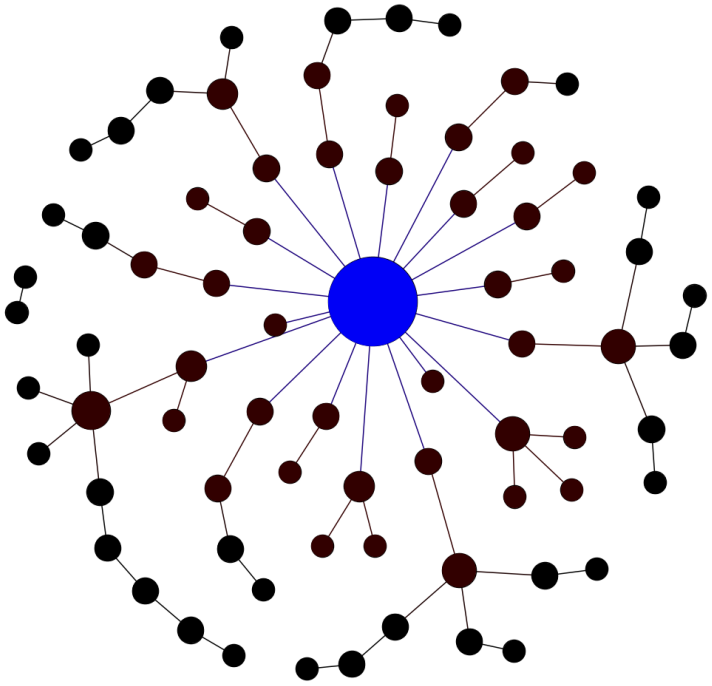

Figure 2.3: Example of an informal undirected discussion tree taken from Reddit of comments and their replies. The node shown in blue represents the root, top-level comment.

Social network analysis provides significant insights for studying users on an individual or community level. Small user indications produce ties which contribute towards discovering valuable connections to other users and groups within the network at large. This is observed through the notion of "the strength of weak ties" phenomena where connections beyond one's peer group support the mobility of information and behaviour from one community to another [157, 30, 64]. Furthermore, networks can be isolated to observe the spread of information using weak ties around a specific user in the form of an egocentric network [300, 175, 83] which closely resembles behaviour of humans in offline networks [20]. Similar behaviour can be measured through reciprocity where pairs of users share a stable mutual connection maintained through communities [283], social capital [126] and indirectly through other users [288].

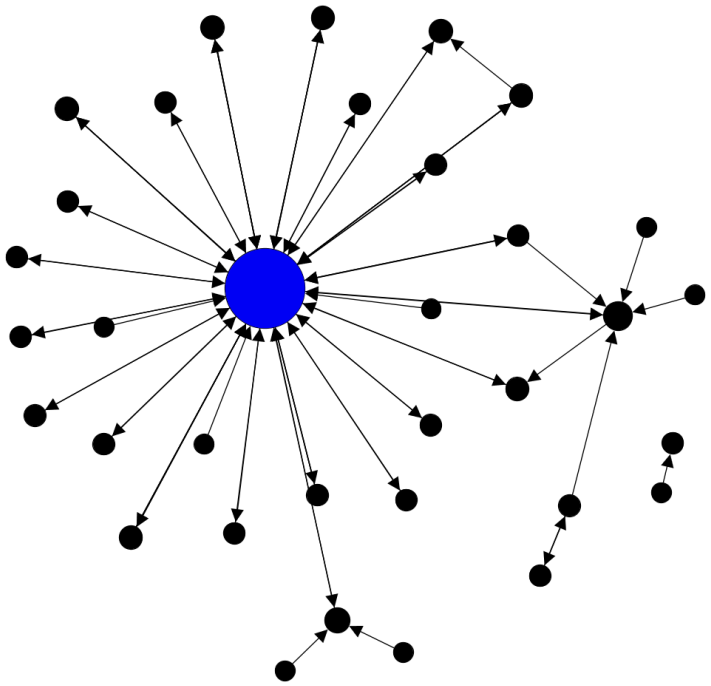

Figure 2.4: Example of an informal directed reply network taken from Reddit of users replying to other users based upon the original discussion tree shown in Figure 2.3. The node shown in blue represents the user who initiated the conversation.

On a larger scale, network analysis can be used to cluster and partition groups of users based upon repeated interaction and weak ties [150]. Community detection can provide a useful metric for understanding the large spread of behaviour through a combination of clustering and positive correlations [283, 72]. In addition to this, higher-level semantics such as individual roles can be extracted to understand how a specific user behaves within a network using clustered embeddings and node structure [158, 75]. Using these features, user reactions are much easier to study with applications such as detecting fake accounts [80], anti-social behaviour [87] and controversy [153, 30].

2.3.3 Communication

Similar to informal networks presented in sub-subsection 2.3.2, communication networks are composed of users who exchange messages (such as email and SMS) between each other which are specifically directed at a user or a group of users.

Although the semantic value of interactions are of high value, such traces of communication are typically hidden and only specific parts of the network are exposed to a user at one particular time. These networks are centred around a user in the form of an egocentric network in which only a central user is aware of their interactions with other users [140].

The rise of modern technology has resulted in a rapid increase in research surrounding the study of communication networks though multiple formations [341, 101]. One significant area of research focuses on the impact of spam. Network analysis is applied to understand the distinct properties of a network propagating spam combined with temporal features [202, 45, 114, 92]. In addition to spam detection, general behavioural characteristics can be extracted to infer more meaningful connections such as trading networks [323, 338] and organisational networks [271].

2.4 Understanding the Role of Complex Networks

In view of the network structures described in the previous sections, it is evident that modelling interactions on social media using network representations leads to the emergence of *complex networks*. Complex networks refer to a subset of networks which do not feature simple or regular structures (e.g. a lattice) and often reflect structures which are representative of real world systems (e.g. communication, biological, computer and, most importantly, social networks) [12].

Complex networks are typically distinguished by their scale and/or dynamism, and therefore are useful in representing social media. In doing so, this opens up more

opportunities to examine social media in greater detail.

2.4.1 Properties of Complex Networks

As mentioned previously, the properties of complex networks are relevant to social media due to the wealth of user interactions and the scale of which these networks grow and change. Social media networks, and their associated representations, have natural characteristics that lend themselves to being considered complex networks (e.g. scale and dynamism). The collection of links that build up in particular ways lead to patterns that can be identified at scale. For example, two of the most well-known and established classes of complex networks include **small-world** and **scale-free** networks. These networks have properties that emerge at scale and affect how the network functions.

For example, a small-world network has low average node degree, where most nodes are not neighbours of one another, but paths can reach every other node with a short length [388]. Typically, these networks are sparsely populated and have a very low global clustering coefficient but feature small cliques such that a node can be accessed directly from any neighbouring node. Small-world networks have a wide-spread presence on the web [135] but also apply specifically to social media platforms as they are due to the high connectivity of interactions [286]. Small-world properties also have an impact on information diffusion and conversational dynamics [208, 93].

A scale-free network is a further class of complex network, characterised by a degree distribution (degree being the number of edges connected to a node) which follows a power law distribution [32]. These are typically characterised by the presence of high-degree nodes otherwise known as “hubs”, with their spokes having a small degree. These properties are often driven by preferential attachment in network growth processes [33]. The properties of preferential attachment are observed in this thesis by taking observations from metrics such degree distribution, transitivity and reciprocity. These are observed in greater detail in Chapter 5 using user-to-user networks by examining

high-degree egocentric subgraphs centred around a single user.

Scale-free networks are fundamental to social network structure and appear naturally according to simple edge formation rules as observed through preferential attachment [49, 33].For example, these phenomena have been observed on Twitter by studying how follower connections contribute to certain users rapidly gaining popularity though the "rich gets richer" philosophy [18]. Furthermore, scale-free networks also make it possible to find communities structures [201] and model the spread of information (including rumours) [302, 276].

2.4.2 Complex Networks in Social Media

As identified by Guliciuc et al., social media can be treated a type of complex system as "*complex systems are composed of a very large number of different elements with non-linear interactions; furthermore the interaction structure, a network, comprises many entangled loops*" [165, 99].

Both social media and complex networks are characterised by the essence of scale and dynamics. For this reason, complex networks are well-suited for studying social media activity by using specific techniques have been developed over time. These techniques provide methods to interrogate platforms in new ways by studying networks derived from social media. As a result, this opens up more possibilities to analyse behaviour to a greater extent though the use of social network analysis and is explored further in Chapter 3 as the basis for understanding the behaviour of users on social media

2.5 Conclusions

To conclude, this overview of literature from related academic areas provides context to the problem addressed in this these and outlines the existing research surrounding

disruptive behaviour on social media. This was first explored by considering the real-world relevance of the impact of disruptive behaviour based upon recent events. This was then followed by an overview of the types of disruptive behaviour that takes place on social media and the state-of-the-art methods used to identify and combat them. Finally, this literature review provides an overview of some of the most fundamental network structures present within any social setting and examines the ways in which behaviour can be modelled using principals taken from complex networks.

As a result of the research presented throughout this review, three key observations can be mode which motivates the work set out in this thesis. These are summarised and explained as follows:

1. **Disruptive activity has evolved and become more mainstream over time.**
2. **Frameworks for studying social media platforms are valuable and needed.**
3. **Complex networks provide potential for modelling disruptive behaviour.**

Firstly, as observed in Section 2.1, it is important to acknowledge that as social media platforms have evolved over time (in terms of features and registered users), so too has disruptive activity become more mainstream and politically motivated over time. For example, fake news on Twitter and Facebook can resurface on multiple occasions, gradually becoming more intense and extreme over time [340, 15]. Similar comments can be made with trolling with respect to how it has become the norm on social media with the potential to shape politics and legislation [171].

Secondly, the work featured in Section 2.2 outlines the many ways in which it is possible to discoverer and predict different forms of disruptive activity using different types of information. While there are a few examples which use network-based solutions to detect disruptive activity (e.g. [111, 191, 226, 57, 97, 371]) it has become clear that there is no coherent framework for modelling and predicting disruptive behaviour using a single, cross-platform framework which is suitable for most, if not, all circumstances.

The work featured in Sections 2.3 and 2.4 suggest that it is possible to create a framework using different network representations for capturing different types of behaviour.

Thirdly, given that many of the proposed solutions in Section 2.2 make use of non-network-based approaches (e.g. [88, 297, 260, 354, 34, 199, 331]) such as NLP, it is important to state that there are several disadvantages to these techniques which need to be considered before use. For example, by considering NLP as a possible solution to detect disruptive behaviour, it is import to acknowledge potential problems such as irony [381], multiple word meanings [230] and general ambiguity [205]. These issues are exacerbated when multiple languages are considered meaning that separate models are needed for each language. Consequently, this solution simply does not scale well according to the growing demand of the internet and social media usage [1, 2, 179]. It is clear that a language-agnostic solution is needed.

Fourthly and finally, by treating social networks as "complex networks" (see Section 2.4) this opens up more possibilities with respect to analysing disruptive behaviour from the perspective of social complexity. Furthermore, complex network provide many characteristics include community structure, reciprocity and induced other substructures such as triads, which make it easier to study diverse user interactions across multiple social networks / platforms.

In view of these observations, it is evident that a suitable framework is needed to understand social media platforms and disruptive behaviour based upon social network representation derived from user activity. Social network analysis serves as an ideal solution due to simplicity, effectiveness, versatility and are language-agnostic. This supports the need for further investigations to demonstrate the utility of network-based approaches for disruptive behaviour which is explored further in the next chapter, Chapter 3. Given the scale of the problem, additional literature is provided within each of the subsequent chapters addressing on the issues relevant to the work of each chapter.

Chapter 3

Characterising Diverse Functionality in Social Media

3.1 Introduction

The use of social media has transformed the way we communicate in a public setting online. By using social media platforms, people can communicate with others using a range of different ways. Because of these differences, it is important to acknowledge that interactions using different platforms are not always equivalent.

Additionally, it is also important to consider future social media platforms, which may depend upon affordances that are different to those used today. Research has shown how the future of social media is used in many areas of our lives (e.g. LinkedIn for Work, Spotify for music) and will continue to grow and expand in multiple domains (such as online/offline integration and the role of bots) with an emphasis on platforms which encourage user generated content [19, 210].

The concept of affordances was introduced by James Gibson to describe the relationship between a human (or other organisms) and its environment [149]. Since its introduction, the notion of an affordance has been expanded to describe the relationship between users and the functionality of online environments by means of a computer, tablet or smartphone. This is highly relevant to social media platforms, in which users interact with different content through alternative social media platform interfaces [74].

Research performed by Majchrzak et al. introduced the concept of "network-informed associating" from the perspective of knowledge sharing, where users of a platform establish ties between other users and content [246]. These ties are designed to have a positive effect by expanding social capital and link building. Given that there are many social media platforms available, often designed for different purposes, Treem et al. [364] determined that there are at least four known affordances (visibility, persistence, editability, and association) which can be found on almost all social media. While this is interesting to note, these affordances, in terms of purpose and usage provided to the user, are so high level they are not particularly helpful in distinguishing between types of individual social media platforms. Instead, they represent how social media may compare to other general classes of media. Therefore, we progress by proposing affordances that reflect how different social media maybe used at a more detailed level. Specifically, we consider ways in which an *individual* may extract value from interaction with alternative possible types of social media platform.

It is clear that when it comes to understanding social media, affordances are relevant and have an important impact on the way in which individuals use these platforms. Using principles taken from database management, this thesis considers the *create*, *read*, *update* and *delete* operations (otherwise known as **CRUD**) as affordances due to the way in which social platforms provide functionality allowing users to interact with and manipulate activity in a certain way. Consider the following example, where these CRUD affordances apply to tweets on Twitter.

- **Create**: A user can create a new tweet.
- **Read**: A tweet can be found on a user's timeline news feed.
- **Update**: A tweet can be edited to remove or add new content.
- **Delete**: A tweet can be removed.

As shown, these generic affordances encompass a range of different activities and apply

to many other types of social media platform, including both Wikipedia and Reddit - the focus of this thesis.

For example, on Reddit, a user can **create** a post in subreddit, which can be **read** by others who are a part of said community. This post can also be edited (**update**) or removed altogether (**delete**).

Likewise with Wikipedia, a user can **create** an article which is **read** by others. An article can be revised (**updated**) by the community or can be removed (**deleted**) altogether.

3.2 Categorisation of Social Media Through Data Structures

In view of the many different types and variations of social media available, this thesis considers how users (collectively and individually) may engage with potential content via generic affordances (such as CRUD) acting on user activity. In this section, data structures relating to social media content are introduced as a mechanism to categorise different social media platforms. These platforms can be classified based upon the presence (or absence) of certain data structures which can be observed on the platform.

To understand the role of data structures within this thesis, Figure 3.1 describes the relationship between - **Data Structures**, **Platforms** and **Networks**.

Using the diagram presented in Figure 3.1, the role of a data structure concerning social media content is twofold. Firstly, data structures can be used to *categorise* platforms based upon the presence of certain features provided on a platform and secondly, they can be used to *describe* the basic components of a network representation of a social media platform. As a result, the network representation can be used to *model* the activity on a social media platform.

There are very few instances from the surrounding literature where a classification is

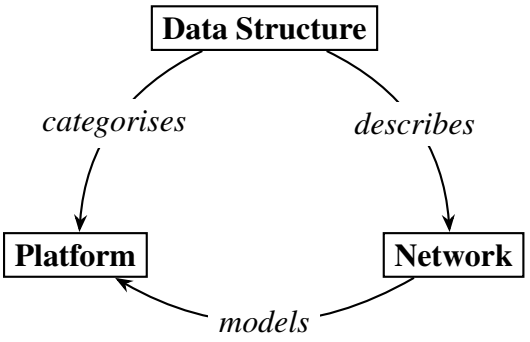

Figure 3.1: The relationship between data structures of social media activity, networks and platforms. The data structures categorise the platform and describes the networks. The networks are used to model activity on the platform.

used as a means of describing and modelling user activity on social media. Research by Vuori et al. [380] attempts to classify social media platforms with respect to an "innovation and business-to-business relationship context". They do so using the "5C's", (Communicating, Collaborating, Connecting, Completing and Combining) as a way to evaluate and compare different tools provided by social media platforms.

However, research by Koukaras et al. identified that social media platforms are experiencing a rapid stage of evolution and that more taxonomies are needed [220]. They introduce 14 high-level utilities which include "connecting", "multimedia" and "professional" as some of the most popular utilities based upon existing research. In addition to this, research produced by Ouirdi et al. [296] considers three dimensions for classification including "who" (Micro-level, Meso-level and Marco-level), "what" (Images, Text, Video, Audio, Games) and "why" (Networking, Sharing, Collaboration, Geo-location) as a method for building a taxonomy.

In view of these findings, this chapter proposes a novel framework for categorising different social media platforms according to the presence of certain data structures. Currently, there are no existing cross-platform frameworks which attempt to classify social media platforms in this manner.

3.2.1 Data Structures for Social Media Content

In the context of this thesis, a data structure is defined as a collection of fields (e.g. a username, timestamp, action, etc) based upon attributes which necessary to represent a user's activity or action on a social media platform which can be manipulated by the basic CRUD affordances (see section 3.1 for the description of these).

This thesis introduces four high-level abstract data structures which we postulate are sufficient to encompass almost all interactions on social media in its current manifestation. We call these data structures **Community**, **Message**, **Collaborative** and **Feed**. These and are described as follows:

- **Community** (*See Section 3.2.2*): Community data structures are needed to focus on user interactions surrounding community / group engagement. For example, posting a submission to a community.
- **Message** (*See Section 3.2.3*): Message data structures focus on interactions directed from one user to another. These conversations can be contained within a *thread* where users can *reply* to one another or through short expressions such as "liking" or "favouriting".
- **Collaborative** (*See Section 3.2.4*): Collaborative data structures are used to capture a user's collaborative activity with respect to an editable repository (e.g. a wiki or git project).
- **Feed** (*See Section 3.2.5*): Feed-based data structures feature time series activity created by a user (e.g. posts or new stories) and arranged according to an algorithm.

The four data structures described above are based upon a novel concept which do not appear within the supporting literature and are created to support the work of this thesis. These data structures are developed using a total of $N = 8$ unique fields which are *User*,

Timestamp, *Interaction*, *Community*, *Recipient*, *Repository*, *Action* and *Content*. Of these fields, only the *User* and *Timestamp* fields are required for all four data structures. The remaining fields are determined by the type of platform that is being considered. This is elaborated further within each of the following sections.

3.2.2 Community Data Structure

Field	Description
User	The user associated with the community
Community	A clearly defined community of interest
Interaction	A type of interaction taking place in the community
Timestamp	The timestamp of when interaction took place

Table 3.1: Proposed necessary fields required for the community data structure

Community structures and interactions are a fundamental component to many social media platforms [84] and are widely used to support homophily (attraction through commonality) and assimilation of like-minded individuals with respect to certain issues [47, 379]. Additionally, communities can be used to help others to reach a target audience for the purposes of self-promotion (e.g. marketing a brand [120]). As a result, community structures have an import role on building social capital and improving information quality with respect to both bonding and bridging [81].

The data structure presented in Table 3.1 seeks to represent social media content within a community by identifying instances where a *User* engages with a *Community* based upon some *Interaction* (e.g. posts to, subscribes to) at some *Timestamp* in time. These are essential fields in order to classify a platform as supporting activity in an online **community**. The data structure provides the basis for modelling a many-to-many mapping between multiple users and communities.

Field	Description
User	The user initiating the message
Recipient	The user receiving the message
Interaction	The type of interaction being made (e.g. a reply)
Timestamp	The timestamp of when interaction took place

Table 3.2: Proposed necessary fields required for the message data structure.

3.2.3 Message Data Structure

Message data structures are essential for allowing users to communicate with one another publicly through social media. In particular, the use of messaging features (e.g. comments and replies) have an impact on driving user engagement [239, 26, 122]. Messages can be used to extract predictive signals for making recommendations [13] and sharing content [349]. As well as driving engagement, the use of message-based features can also be used to offer feedback [173] thorough meaningful conversations based upon shared emotional intensity [40, 152].

Using the proposed data structure in Table 3.2, the presence of a **Message** data structure is conditioned on a pair of users where a *User* initiates an *Interaction* (e.g. a reply, like or comment) on a *Recipient* at a point in time *Timestamp*. The relationship between a *User* and *Recipient* is important for understanding engagement based upon a message being directed to a specific user. The *Interaction* is used to represent explicit actions such as replies and mentions, although this can be extended to include less-explicate actions (such as "likes" and "favourites") as they too can be interpreted as messages or expressions (e.g. a "like" being used to support or agree).

3.2.4 Collaborative Data Structure

As discussed earlier, collaboration has become an important component of the modern Web 2.0 by exploiting the idea of the "Wisdom of Crowds" [359]. According to Hemsley and Mason, social media can be used to "*enable people to connect, communicate, and collaborate*" [180]. For this reason, collaboration is an important process of social

Field	Description
User	The user acting upon the repository
Repository	A block of information which can be manipulated by multiple users (e.g. an article)
Action	The action to manipulate the repository (e.g. create, edit, delete)
Timestamp	The timestamp of when the action took place

Table 3.3: Proposed fields for the collaborative data structure.

media interactions and is therefore relevant to this thesis in an attempt to understand user behaviour on social media. An example of collaboration can be found using various wiki-based platforms (such as Wikipedia) where users collaborate on articles using dynamic social interactions to build high quality content [248, 249].

Table 3.3 proposes necessary fields for the **Collaboration** data structure by making reference to a *User* acting on a *Repository* based upon some task *Action* at a point in time *Timestamp*. In the context of this thesis a *Repository* can be defined as a block of information (e.g. an article, file or code project, etc) which can be edited by multiple users at concurrently. Furthermore, this can also be composed of additional subcomponents (e.g. an article contains headings, images and paragraphs as subcomponents. Much like the **Community** data structure, the **Collaborative** data structure can also be used to model a many-to-many relationship between multiple users and repositories they have manipulated.

3.2.5 Feed Data Structure

Field	Description
User	The creator / author of the content
Content	The main body of the content for the feed
Timestamp	The timestamp of when the content was created

Table 3.4: Fields required for the feed data structure.

Feed-based data structures are widely used across almost all social media platforms as a way of organising and presenting temporally relevant information to users. On

most platforms, a user has a commutable, time-ordered sequence of messages adjusted according to their interests. News feeds have been the subject of discussion in recent years due to the complexity of the algorithms which social media platforms use to present information to users [116, 361] and to remain informed on current affairs [51]. This has resulted in the need for interpretable and explainable news feed algorithms [269]. Others have tried to find alternative ways of adjusting their news feed by comparing alternative algorithms [129] and by unsubscribing from topics or users they no longer find interesting [52]. Consequently, this has the potential to contribute to the formation of echo chambers and filter bubbles [350, 46].

The **Feed** data structure in Table 3.4 proposes a simple construct for identifying the presence of a feed. In this case, a *User* is the creator (or author) of a piece of *Content* (e.g. a blog post, submission, image or video) which was created at a given point in time, determined by a *Timestamp*. The *Timestamp* field is essential as this is used to determine where the item is positioned in a feed according to a given algorithm.

3.3 Platforms of Interest

With respect to the proposed data structures (see Section 3.2) it is important to acknowledge that social media platforms provide a wealth of information for understanding how users behave online.

According to Obar and Wildman, "social media" can be defined by four key characteristics. Firstly, "*social media services are (currently) Web 2.0 Internet-based applications*". Secondly, "*user-generated content is the lifeblood of social media*". Thirdly, "*individuals and groups create user-specific profiles for a site or app designed and maintained by a social media service*". And finally, "*social media services facilitate the development of social networks online by connecting a profile with those of other individuals and/or groups*" [289]. Furthermore, as mentioned previously, Hemsley and Mason describe social media as a set of tools that "*enable people to connect, communicate, and*

collaborate" [180].

Using definitions of social media, the following services are considered examples of widely adopted platforms throughout society among English-based speakers.

- **Facebook** allows users to create profile pages, create posts, 'like' pages, join groups, share photos and videos with 'friends' - a bidirectional connection between two profiles.
- **Instagram** is a multimedia sharing service where users can upload and share photos and short videos. Instagram allows users to follow other user accounts to interact and follow their content.
- **Twitter** is a microblogging platform where users post short pieces of text (280 characters) where users have the ability to retweet (share), mention and reply to other users' tweets based upon users accounts they follow.
- **YouTube** is a video sharing platform where users can upload videos, subscribe to channels and engage with other users' videos by means of leaving comments, 'liking' and 'disliking'.
- **Reddit** is a social news aggregation site, allowing users to submit links and text posts to communities known as subreddits where others can leave replies and "upvote" and "downvote" submissions.
- **Wikipedia**[1] is an online encyclopedia and collaboration platform where users can help improve the quality of articles by collaborating and interacting with others through the use of 'talk' pages and revision logs both directly and indirectly.
- **TikTok** is a video sharing platform designed to enable users to upload and share short entertaining video clips. In return, users can react and collaborate with others though videos clips of their own.

[1]Considered as social media according to Treem et al. [364]

- **Snapchat** is a messaging service designed to allow users to send images and messages to each other which disappear after a short interval such that the original recipients can no longer access the content after this period.

- **Pinterest** is an image sharing platform where users create pins (images, videos, GIFs, etc) found across the web designed to help users create, share and contribute ideas with others. These pins can be grouped together using 'boards'.

- **Quora** is a Q&A-based platform in which users can submit questions and provide contributions in the form of answers in response. Other users can help improve the accuracy of answers by providing adjustments.

As of this writing, of the platforms mentioned above, Wikipedia, Reddit and Twitter are accessible using a freely available public API. Facebook, YouTube and Instagram provide limited or restricted access and TikTok, Snapshot and Quora have no public API.

For the purposes of this thesis, it is necessary to consider only a subset of these social media platforms by focusing on those that exemplify different behaviours and depend on the different data structures introduced in Section 3.2 and have accessible data collection. To establish suitable candidate platforms for research purposes, the following criterion is used to consider possible platforms.

- **Public API**: In order to obtain the necessary data used for analysis, it is important that the platform of interest supports an API where anyone can register for access with minimal rate restrictions. Collecting data via a public API is not straightforward for several reasons. Firstly, any automated implementation (e.g. a Python script) must take into consideration rate limits by ensuring that all requests made to the API are within the allowance window. Secondly, most APIs return results in blocks meaning that the implementation will need to page though each set of results in order to receive all the data needed for the analysis. Finally, all

APIs vary depending on the platform meaning that data retrieval strategies will need to be adjusted accordingly. For example, using the Twitter API, tweets are "hydrated" meaning that the full tweet metadata is retrieved using its ID. Without a public API, the task of collecting data (e.g. via web scrapping) may prove challenging at scale and/or may be forbidden.

- **Time series**: A platform of interest must store user activities (e.g. a post or reply) with a timestamp such that the data can be arranged in a time-series based format. This provides a representation for understanding patterns and trends over time and is a requirement of the proposed data structures that we are seeking to consider.

- **Lookup Query**: The platform must support the ability to "lookup" the details of a specific user account or other item of interest. This is particularly important when considering a community or topic.

- **Traceable**: In addition to **Lookup Query**, the platform must support the ability to trace activity. For example, the ability to look up a user account and to find a list of recently created posts. This can be reversed to look up the post and find the original author. This is relevant to the community and collaborative data structures.

- **Mode of Media**: Understanding the primary mode in which a platform operates (e.g. users posting text, images, videos, etc.) is important for information retrieval. For the purposes of this thesis, text-based information is easier to process as the relevant information can be extracted in an automated fashion (e.g. extracting URL's in a text post).

A summary of popular social media with respect to the above criteria is presented in Table 3.5. While there are many other alternative social media, it is apparent that very few of them satisfy all the above criteria. In view of the results provided in Table 3.5, this thesis focuses on three specific social media platforms of interest; **Wikipedia**, **Reddit**

and **Twitter**. These three platforms are known throughout the thesis as "platforms of interest" and are summarised further in the following sections.

Platform	Public API	Time series	Lookup Q.	Traceable	Mode of Media
Wikipedia[2]	Yes	Yes	Yes	Yes	Text
Reddit[3]	Yes	Yes	Yes	Yes	Text
Twitter[4]	Yes	Yes	Yes	Yes	Text
Facebook	Limited	Yes	Limited	Yes	Text
YouTube	Limited	Yes	Limited	Yes	Video
Instagram	Limited	Yes	Yes	Yes	Multimedia
TikTok	No	N/A	N/A	N/A	Video
Snapchat	No	N/A	N/A	N/A	Multimedia
Pinterest	Restricted	No	Yes	Yes	Images
Quora	No	N/A	N/A	N/A	Text

Table 3.5: Matching popular social media platforms to the criterion reduces the number of platforms down to a manageable size.

3.3.1 Wikipedia

Wikipedia[5] is an online encyclopedia founded in 2001 by Jimmy Wales and Larry Sanger. Wikipedia's articles are created and maintained by a community of international volunteers and supports over 300 spoken languages.

MediaWiki[6] (the software which operates Wikipedia) provides the functionality for users to interact with each other in a collaborative fashion by allowing users to register for an account and make revisions to articles they are interested in. Furthermore, users can communicate with one another though the use of talk pages - a dedicated page attached to an article for discussing matters relating to the improvement of an article.

As a result, Wikipedia is predominantly a collaborative-based platform as it is used to facilitate the process of revising articles. Data such as the article revision history is preserved since the article's creation and can be publicly accessed without the need of creating a user account.

[5] `https://wikipedia.org`
[6] MediaWiki: `https://www.mediawiki.org/`

3.3.2 Reddit

Reddit[7] is a social news aggregation and discussion website founded in 2005 by Steve Huffman, Aaron Swartz and Alexis Ohanian. Reddit's users (also known as "Redditors") can submit content to different communities (known as "subreddits") in the form of links, text posts, images and videos. Users can determine the popularity of content through the process of "upvoting" and "downvoting" other user's submissions. Additionally, Reddit can be treated as a discussion forum by encouraging users to leave comments and replies in response to a submission.

As a platform, Reddit offers the most in terms of extracting user interactions. Upon creating a user account, users are encouraged to subscribe to topic-based subreddits where users can submit posts (either as a link or piece of text), leave comments, and reply to others. What makes this platform unique is that users have an incentive to produce meaningful contributions using "karma". Karma can be collected through people "upvoting" and "downvoting" submissions and comments. These voting patterns can lead to some rather interesting data structures.

3.3.3 Twitter

Twitter[8] is a popular microblogging platform founded in 2006 by Jack Dorsey, Noah Glass, Biz Stone and Evan Williams. Twitter allows registered users to post and interact with short 280-character long messages known as "tweets". As well as composing tweets, users can interact with others in a public setting.

Twitter is perhaps one of the most used and recognised platforms featured in this thesis. Unlike Reddit and Wikipedia, activity on Twitter mainly based on direct user-to-user interactions and does not rely upon topic orientated communities as Twitter's primary means of user interaction is initiated through Tweets. As a result, publicly available

[7] `https://www.reddit.com/`
[8] `https://www.twitter.com/`

user interactions can be achieved in one of three ways. A user can *retweet* a user (the process of sharing another user's tweet), *reply* to another user's tweet and *mention* a user (referencing a user in their own tweet).

3.4 Data Structures in Social Media

The platforms of interest listed in Section 3.3 correspond to different data structures as introduced in Section 3.2). Table 3.6 identifies each of these platforms with respect to alternative data structures. Cells marked with "N/A" indicate that the platform and data structure do not align and are therefore beyond the scope of this thesis.

		Data Structure			
		Community	Message	Collaborative	Feed
Platform	Wikipedia	N/A	N/A	*See 3.4.1*	N/A
	Reddit	*See 3.4.2*	*See 3.4.2*	N/A	*See 3.4.2*
	Twitter	N/A	*See 3.4.3*	N/A	*See 3.4.3*

Table 3.6: Relationship between platforms of interest and all data structures.

As observed in Table 3.6, Reddit is one of the most versatile platforms of interest as it spans all data structures except collaborative. Twitter and Reddit share similar functionality where they are both considered message and feed-based platforms.

For example, message-based data structures can be observed on Twitter through three known user interactions (mention, reply and quote retweet) and through Reddit with replies to submissions within community-centric subreddits.

As a result, similar network representations can be used to compare data structures in other platforms. This thesis considers the three platforms of interest further with respect to data structures in the following sections.

3.4.1 Wikipedia

As a platform, Wikipedia provides multiple ways for capturing user interactions. Using the API, all article revisions can be exported as well as the contributions that other users have made. The exported data provides many useful attributes which can be modelled into multiple network representations. A list of valuable attributes can be found in Table 3.7.

Attribute	**Data type**
User	*String, A unique username identifying the user*
Article	*String, The title of the article of interest*
Action	*String, The type of action performed on the article (delete, append, revert etc.)*
Timestamp	*DateTime, The exact time of the revision taking place*

Table 3.7: Raw attributes extracted from the Wikipedia API.

Collaboration

Platform Attribute		**Collaborative**
User		User
Article	$\longrightarrow$	Repository
Action		Action
Timestamp		Timestamp

Table 3.8: One-to-one mapping between platform attributes and the collaborative data structure fields using data provided from the Wikipedia API.

As demonstrated in Table 3.8, the platform attributes presented in Table 3.7, best align with the collaborative data structure.

In doing so, this approach can be used to understand mutual ties between articles and users. The main benefit of this approach is to understand how users collaborate on multiple articles with other users. This is particularly important for identifying disruptive activity such as brigading where a collection of articles are vandalised by multiple users in a coordinated effort to cause mass disruption. As a result of using this approach, groups of users can be identified with ease meaning that such activity can be minimised for future use.

In addition to this, the *Timestamp* attribute can be used to observe possible switches between different revisions of articles. This has applications for identifying revisions of an article of interest where vandalism or "fighting" has taken place between two individuals projected over time. Using this model, the use of reciprocated ties can be used to identify the individuals responsible for manipulating another user's unique contribution.

3.4.2 Reddit

Much like Wikipedia, Reddit also provides a central API to access the data needed to capture user interactions. With the API, all user activity and the most recent posts of a subreddit can be exported. In addition to this, conversations in the form of comments and replies can be extracted and reassembled in its original hierarchical formation.

Attribute	**Data type**
User	*String, A unique username identifying the user who made the post*
Post URL	*String, The URL of the post (if link)*
Post Body	*String, The body of text associated with the post*
Subreddit	*String, The subreddit the post was submitted to*
Karma	*Integer, The net-upvotes score (upvotes minus downvotes)*
Timestamp	*DateTime, The exact time the post was created*

Table 3.9: Raw attributes extracted from the Reddit API for processing submissions.

Attribute	**Data type**
User	*String, A unique username identifying the user who commented*
Parent	*String, The unique username of the parent user (if reply)*
Post	*String, The URL of the submission*
Comment Body	*String, The raw text of the comment*
Subreddit	*String, The subreddit the post was submitted to*
Karma	*Integer, The net-upvotes score (upvotes minus downvotes)*
Timestamp	*DateTime, The exact time the comment was created*

Table 3.10: Raw attributes extracted from the Reddit of a API for processing comments.

Communities

Platform Attribute
User
Subreddit
"*Post*"
Timestamp

$\longrightarrow$

Community
User
Community
Interaction
Timestamp

Table 3.11: One-to-one mapping between platform attributes and the community data structure fields using data provided from Reddit.

Table 3.11 demonstrations how the attributes presented in Table 3.9 for processing Reddit submissions align with the community data structure.

The attributes presented in Tables 3.9 and 3.10 can be used to model the interactions between *User* and *Subreddit* to capture different types of activity. In the case of Reddit, this includes a user submitting a post to a subreddit or leaving a reply on a post within another subreddit. This approach can be used to find important connections which can contribute towards the detection of brigading through coordinated groups.

Messages

Platform Attribute
User
Parent
"Comment" / "Reply"
Timestamp

$\longrightarrow$

Message
User
Recipient
Interaction
Timestamp

Table 3.12: One-to-one mapping between platform attributes and the message data structure fields using data provided from Reddit using comments.

As shown in Table 3.12 the raw attributes extracted from Reddit (see Table 3.10) for processing comments align with message data structure.

The comment attributes (see Table 3.10) can be used to model the temporal conversational dynamics between pairs of users within a comment thread. This approach can be used to find who produces and receive the most replies as well as discovering reciprocated ties. As a result, this approach is important for understanding direct user-to-user

interactions. This is particularly valuable for identifying disruptive users (e.g. trolls [354, 297]) to see how users react in response to provocative content.

Feeds

Platform Attribute		Feed
User		User
Post Body	$\longrightarrow$	Content
Timestamp		Timestamp

Table 3.13: One-to-one mapping between platform attributes and the message data structure fields using data provided from Reddit using submissions.

Finally, Table 3.13 demonstrates that Reddit also aligns with the feed data structure for processing submissions (see Table 3.9).

The *Timestamp* attribute of Table 3.9 can be used to study the temporal ordering of information in the way that it is presented to different users. This can be achieved in two ways. Firstly, the *User* attribute can be used to monitor the flow of submissions from the perspective of a user's home feed or a subreddit. In doing so, this considers the potential for coordination to emerge based upon how information is posted at different intervals by a group of actors (e.g. brigading [200]). Secondly the *Subreddit* attribute can be used to observe how a single user transitions between different communities and topics over time. This has the potential to reveal an agenda in the form of switching patterns and to understand basic patterns of migration [110].

3.4.3 Twitter

All content produced on Twitter can be accessed via the central API which can be collected using different techniques. Tweets can be collected based upon a user's profile where all tweets produced from a specific account are collected. Alternatively, data can be collected based upon a set of keywords, hashtags, or a phrase. In doing so, this

returns a set of tweets round a particular topic or theme which isn't biased to a certain user.

Attribute	**Data type**
User	String, A unique username identifying the user who commented
Tweet Text	String, The raw text extracted from the original tweet
Hashtags	String, A list of hashtags featured in the tweet
URLs	String, A list of URLs featured in the tweet
Mentions	String, A list of usernames mentioned in the tweet
Retweet	String, The username of the retweeted user
Reply	String, The username of the parent tweet
Timestamp	DateTime, The exact time the tweet was created

Table 3.14: Raw attributes extracted from the Twitter API for processing tweets.

Message

Platform Attribute		**Feed**
User		User
Mention, Retweet or Reply	$\longrightarrow$	Recipient
"Mention", "Retweet" or "Reply"		Interaction
Timestamp		Timestamp

Table 3.15: One-to-one mapping between platform attributes and the message data structure fields using data provided from Twitter.

Table 3.15 demonstrates how attributes for processing tweets aligns with the message data structure.

As mentioned previously, most interactions on Twitter are primarily represented by user-to-user interactions. Using the *User* attributes, interactions are formed based upon *Replies*, *Quote retweet*, and *Mentions*. These interactions can be modelled using a network where a directed edge between the two users can be used to indicate the direction of the interaction. For example, a network representation could be used to capture reciprocated ties for replies (do they reply to each other?). By using this approach, simple data structures can be extracted to capture behaviour such as trolling - where users are likely to provoke others by means of replies or mentioning.

Feeds

Platform Attribute		Feed
User		User
Tweet Text, Hashtags or URLs	$\longrightarrow$	Content
Timestamp		Timestamp

Table 3.16: One-to-one mapping between platform attributes and the feed data structure fields using data provided from Twitter.

In addition to the message data structure, Table 3.16 demonstrates how attributes for processing tweets also align with the feed data structure.

Metadata tags (such as Tweet *hashtags* and *URLs*) can be represented as an association model to analyse the relationship between a *User* and an entity (e.g. a hashtag or hyperlink featured in a Tweet. As a result, this approach considers how groups of users use certain hashtags or hyperlinks to promote their tweets. Furthermore, by using this bipartite relationship, it's possible to discover communities of users based upon shared mutual interests. This is particularly important for discovering how disinformation is propagated across a network and how users can amplify certain topics.

To summarise, the platforms of interest introduced in this chapter provide multiple ways for accessing and modelling social media platforms using network representations derived from data structures. It is important to consider that there are many alternative platforms that could have been included (e.g. Instagram, Facebook, YouTube, e.t.c), the platforms of interest featured in this investigation support API's in which data can be retrieved with ease and provide an endpoint in which networks can be directly observed (as shown) with minimal alterations.

Each of the platforms of interest are members of at least one data structure and collectively span all data structures of interest. While it is theoretically possible for for a platform to span other data structure described in this investigation which are not currently aligned (e.g. Reddit and Collaboration, Wikipedia and Feeds) it is important to state that there is no strict criterion for data structure compatibility.

The combination of platforms and aligned data structures described in this investigation best represent behaviours which can be adequately captured with network representations. Therefore, the chosen data structures and platforms sufficiently capture all social media characteristics of interest for the purposes of this thesis therefore, the remaining chapters of this thesis focuses on these three platforms exclusively.

3.5 Behavioural Networks

The data structures introduced so far in this chapter provide the basic elements from which the behaviour of users in the social media platforms can be assessed. Analysing user behaviour in social media platforms can be achieved in one of two ways. Firstly, the **content** of the platform can be used to understand *what users are saying* and secondly, the **users** tell us *what they are doing* and *how they interact with others* through network analysis. Analysing user behaviour can be achieved using network-based representations which are referred to in this thesis as "behavioural networks". For example, throughout the literature, the notion of behavioural networks have had an important part in recommendation systems (for understanding consumer activity) [130] and detecting leaders and influential users [131].

As a result, by constructing network representations based around user behaviour, this approach can be used to understand how users behave on different social media platforms with respect to their relationships and interactions. Contrary to a content-based approach (such as NLP), networks are language-agnostic meaning that they are universally interpreted and can be used to reveal more about how a user behaves and/or is likely to behave. This is an important component for the purposes of analysing disruption on social media which is largely driven through behaviour.

Behavioural networks are introduced to provide a modelling framework for the platforms of interest using the data structures described in Section 3.4. Through the use of network analysis, it is possible to build and analyse relationships surrounding user interactions

such that individual and, through neighbouring connections, collective behaviour can be observed. The data structures we have introduced can be used to construct a behavioural network (see Figure 3.1).

In an attempt to exploit data structures in different ways, three network representations are introduced and are examined throughout this thesis to establish their capacity to usefully model behaviours using network structures; **Transitional**, **User-To-User** and **User Association**. These network representations are unique to this thesis and aid the modelling of user behaviour on the platforms mentioned in Section 3.3. These representations are defined further using the following three naming conventions:

- **Transitional Networks**: We introduce these to capture patterns and anomalies within time-ordered online activity based upon rapid (within quick succession) sequential behaviour such as interaction or editing. As a result, this approach seeks to find important switches between pairs of users or other entities (e.g. articles, communities, groups, etc) which occur immediately after each other. This employs directed edges. Within the literature, the concept of transitional networks are derived from process mining [373, 347] for observing the exchange of information between different processes. This approach has been extended to focus on user behaviour such as smartphone app migration habits [370] and personal spending on credit cards [117].

- **User-To-User Networks**: We introduce these to model interactions that are performed between any pair of users where one user's actions are directed to another. For example, user-to-user networks are frequently used for modelling conversational dynamics between users (e.g. User A *replies* to User B) such that an edge is formed when a user replies to another user [412, 256]. Additionally, this network representation is fundamental for studying large-scale communication networks between users [228, 27, 238] and serves as the basis for detecting groups through techniques such as community detection based on highly interconnected users (e.g. friendships) [352, 282].

- **User Association Networks**: These can be used for modelling associative connections between users and other entities such as a community, group or article. For example, this approach can be used to model users' memberships to various groups where an edge is used to form the association such that User A *is a member of* Group X. This network representation uses a bipartite structure which has the capabilities of capturing the distribution of items (e.g. groups, communities, etc.) associated with a user and to discover intersecting associations with others (e.g. both User A and B are members of Group X and Y). Further examples of this representation can be observed through Wikipedia collaboration [397, 400, 398], online learning environments [227] and Reddit communities [231].

The three network representations (as above) can be built using data contained in the data structures introduced in Section 3.2 and can be used to model activity on the three platforms of interest.

The **Transitional**, **User-to-user** and **User association** network representations are used throughout this thesis for capturing user behaviour. In the context of this thesis, all network representations are formally defined as a graph G where $G = (V, E)$ with a set of vertices (also known as "nodes") V and edges E. Edges are represented as a pair of nodes $e = \{x, y\}$, $e \in E$, (an undirected edge) or $e = (x, y)$, $e \in E$ (a directed edge). Each of the network representations has a specific definition for vertices and edges, described as follows.

3.5.1 Transitional

Transitional networks are used to model data that has been arranged according to time. They provide a representation for understanding valuable connections (also known as "switches" or "hops") between two nodes. For this reason, they are particularly important for capturing feed-based data structures due to the way in which the data is presented using the timestamp field. This thesis explores transitional networks from

two perspectives: user-orientated and content-orientated (see Figures 4.4 and 4.5 for examples in Chapter 4). This approach supports multiple degrees of freedom as nodes can be represented as either users or pieces of content (e.g. a post or tweet associated with a user) however, edges are strictly bound to represent transitions.

A transitional network is defined by switching behaviours between two entities. In this case, the temporal component t (a timestamp) is used to determine the order and position of these entities. This approach can be used to model a change between states or users. A directed edge is formed where an item at t_i precedes another item at t_{i+1} such that an edge $e = (t_i, t_{i+1})$ forms $t_i \longrightarrow t_{i+1}$. Items here represent online social media.

Example: Wikipedia user contributions

Using Wikipedia as an example, a transitional network can be used to model a user's migration patterns between different articles they have made revisions to based upon sequential activity. For example, Table 3.17 below demonstrates a simple example of a Wikipedia contribution log where a user has edited articles A, B and C on different occasions marked by timestamp t_i. In this example, The log starts from A_{t_i}

Timestamp	Article	Action
t_1	A	EDIT
t_2	B	REMOVE
t_3	C	CREATE
t_4	A	EDIT
t_5	C	EDIT

Table 3.17: Example list of contributions made by a single Wikipedia editor.

Using the example given in Table 3.17, this sequential activity can be reproduced as a transitional network. Duplicate appearances of an article (e.g. C_{t_3} and C_{t_5}) are represented as a single node and edges represent the direction of the movement at the time the vision was made. This can be observed in Figure 3.2.

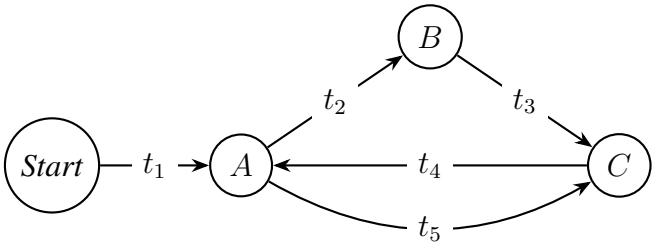

Figure 3.2: A simple transitional network based upon activity in the form of contributions to Wikipedia articles in relation to the activity of a single user. The timestamps t_i mark the order of interactions between nodes A, B and C originating from the $Start$ node.

Example: Twitter news feed

For a further example, suppose a user reads through their news feed (populated by the users they follow) or tracks a certain search term on Twitter (e.g. #coffee). In this example, a transitional network can be used to observe how a topic can change between different users. In doing so, this approach could be used to follow the context of a topic over time based upon when a user composes the tweet. Consider the example set of "tweets" in Table 3.18.

Timestamp	Username	Tweet Text
t_1	@A	*"Words cannot express how much I love #coffee."*
t_2	@B	*"I usually have my #coffee in the morning."*
t_3	@C	*"My favourite type of #coffee bean must be roasted in Italy."*
t_4	@A	*"The Italians know how to make great #coffee."*
t_5	@D	*"I want to go on holiday to #Italy to try some amazing #coffee"*

Table 3.18: An example of a hypothetical Twitter timeline using #coffee as a search query.

The network representation of the "tweets" featured in Table 3.18 can be found in Figure 3.3. Much like the previous example, nodes represent multiple instances where a user produces a tweet which matches the example search term. This approach is versatile as nodes are interchangeable and can represent either users or another variable or state.

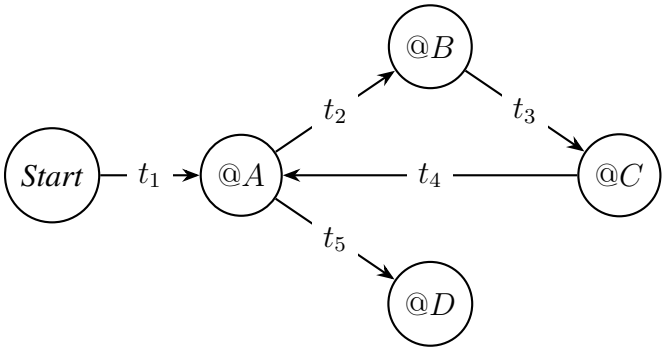

Figure 3.3: A simple transitional network modelling behaviour on Twitter in relation to a single topic (e.g. #coffee). The timestamps t_i mark the order of interactions between nodes $@A$, $@B$, $@C$ and $@D$ originating from the $Start$ node.

3.5.2 User-To-User

User-to-user networks model direct explicit interactions between any pair of users. In the context of this thesis, user-to-user networks and message-based data structures are used to capture interactions in a content driven manner (e.g. replying and sharing). However, this can also be extended to capture intersection derived from feed-based data structures too. For example, whether a user decides to reply to a tweet or not can be determined by how it was presented to the user (if at all) via their "timeline" - a news feed generated by an algorithm based upon a user's interested and who they follow. As a result, this network representation supports many degrees of freedom as edges can represent different types of interaction and hold important information (such as for classification purposes).

By definition, the vertices V of a user-to-user network $G = (V, E)$ are composed exclusively by a set of users U such that $V = U$. An edge $e \in E$ is defined based upon the presence of an interaction directed at a single user such that $e = \{u_i, u_j\}$ represents an interaction directed at u_j from u_i. All user-to-user networks are directed.

For example, a user-to-user network can be used to model users replying to each other in a discussion thread (message) or retweeting another user's content (feed). In these examples, an edge $e = \{u_i, u_j\}$ represents u_i *replies to* u_j or u_i *retweets from* u_j. In

addition to this, edges can hold different types of metadata to produce more accurate network representations. Additional data may include a timestamp t_i and the number of times an interaction occurred w_i forming $e = \{u_i, u_j, t_i, w_i\}$.

User-to-user networks can be represented in one of two ways. Firstly, a network can be *aggregated* meaning that every instance of an interaction over time is combined into a single network without taking into account when the interactions occur. This is ideal for studying behaviours surrounding user interactions. Secondly, a network can be represented as a *tree* such that a node represents an instance of an interaction rather than a collection. This is beneficial for understanding the spread of information between users. Unlike an aggregated representation, a tree contains a root node marking the start of the interactions with the remainder being derived from this point.

Example: Reaction network

Platforms such as Facebook and Twitter, allow users to "react" or "like" another user's content as opposed to leaving a written message. An example of a reaction may include " laughing", " angry" or "sad". This feature allows a user to respond with a simple gesture without leaving a comment. Furthermore, the process of a user "reacting" to another user's piece of content can be modelled as an aggregated user-to-user network. For example, when u_i reacts r to another user u_j at timestamp t_i, this forms a directed edge where $e = \{u_i, u_j, t_i, r\}$. An example network can be found in Figure 3.4.

3.5.3 User Association

User association networks are based upon a bipartite relationship between a user and community or topic. User association networks are used to represent community-centric data structures in which a user has an implicit or explicit connection with the community. Examples include, a user submitting a post to a subreddit or a user posting a Tweet featuring a hashtag. The community or topic of interest (in this case, a subreddit or

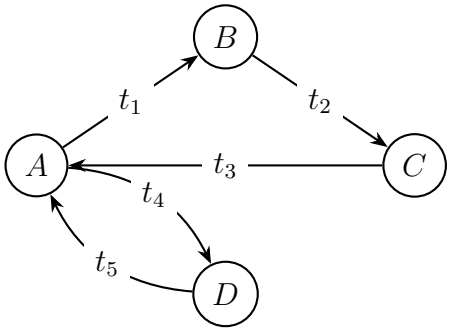

Figure 3.4: A simple example of a reaction network using an aggregated user-to-user network where a directed edge represents a user reacting to another user. The timestamps t_i mark the order of interactions between nodes A, B, C and D.

hashtag). By using this approach, community structures can be inferred based upon mutual connections with communities or topics.

User association networks are constructed using two mutually exclusive sets of nodes containing users U and communities or topics C such that $V = U \cup C$ and $U \cap C = \varnothing$ with every node of U joined to every node of C. An edge e is formed when a user U_i can be associated with a community or topic C_i such that $e = \{U_i, C_i\}$. This can also be extended to include a timestamp t_i as a label on the edge.

Example: Reddit Submission Interaction

As a platform, Reddit allows users to interact with another user's submission by directly commenting to their submission or via a reply to a another user's comment within the same submission. In this example, a user association network can be used to model the bipartite relationships between different users and the submissions that they have engaged with. As a result, a user u_i forms an edge with a submission s_i such that $e = \{u_i, s_i\}$ if they have left a comment on S_i. This relationship can be described as u_i *engaged with* s_i. This can be observed in Figure 3.5 where $U = \{A, B, C, D\}$ and $S = \{1, 2, 3, 4, 5\}$.

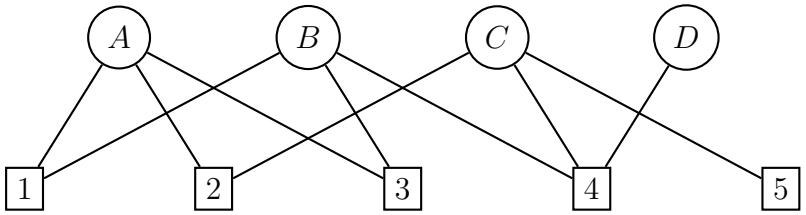

Figure 3.5: Example of a user association network used to model user engagement with subreddit submission. An edge in the bipartite network represents a user (A, B, C or D) posting in a subreddit (1, 2, 3, 4, or 5).

Example: Wikipedia Collaboration

In order to facilitate article collaboration, Wikipedia provides users with the ability to create, modify and delete articles. As a result, this process can be implemented as a user association network to model the different types of relations a user can have with an article. For example, not only can a user modify an article, they can also create and delete articles too. Taking this type of interaction into consideration can greatly improve the accuracy and representation of the network. In this example, a user u_i, can edit an article a_i such that $e = \{U_i, a_i, act_i, t_i\}$ where act_i represents an "action" from the set $\{CREATE, EDIT, DELETE\}$ and t_i represents the timestamp. An example network can be presented in the form of a network and edge list as observed in Figure 3.6 and Table 3.19 respectively where **Source** represents a user and **Target** represents the article.

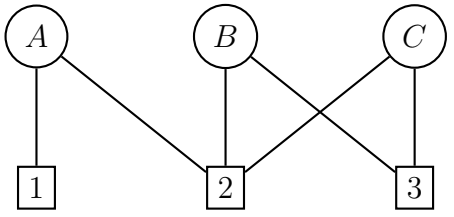

Figure 3.6: Example network modelling Wikipedia collaboration interactions. An edge in the bipartite network represents a user (A, B or C) editing an article (1, 2 or 3).

To summarise, the three network representations introduced in this thesis encompass common data structures which are present across many social media platforms. For this

Source	Target	Action	Timestamp
A	1	CREATE	t_1
A	2	CREATE	t_2
A	2	EDIT	t_3
B	2	EDIT	t_4
B	3	CREATE	t_5
C	2	REMOVE	t_6
C	3	EDIT	t_7

Table 3.19: Corresponding edge list for the network presented in Figure 3.6.

reason, they capture important user behaviour for subsequent analysis using complex network techniques that are designed to expose latent characteristics.

3.6 Capturing Data Structures Through Behavioural Networks

The network representations introduced and described in Section 3.5 demonstrate the potential for capturing behaviour through networks that are derived from simple data structures representing social media. Based upon the network representations described in Section 3.5, Table 3.20 outlines the behavioural networks in their relationship to data structures. This Table indicates how behavioural networks will be explored across all data structures (community, message, collaborative and feed).

		Data Structure			
		Community	Message	Collaborative	Feed
Network	Transitional	-	-	See 4.4 (Wikipedia)	See 4.5 (Reddit)
	User-to-user	-	See 5.4, 5.5 (Twitter, Reddit)	-	-
	User Association	See 6.4 (Reddit)	-	-	-

Table 3.20: Cells indicate the sections where network representations will be used to explore the data structure functionality with specific social media platforms.

3.7 Exploiting Techniques From Complex Networks

Behavioural network representations allow techniques from complex networks to be used to potentially identify characteristics of behaviour in online activity. In this section we introduce key techniques from complex networks that are the focus of our investigation.

As mentioned in Section 2.4, complex networks have an important role in understanding the substructures present within social networks and their corresponding representations. Both micro (local) and macro (global) phenomena are important features when studying complex networks [132]. In the social world, micro-based phenomena concern everyday human social interactions such as face-to-face interactions [344] which lead to macro phenomena, such as the small world effect.

This thesis makes use of various complex network methodologies in an attempt to understand global user behaviour by examining smaller, "local" interactions. This aligns with the hypothesis and research questions of this thesis (see Chapter 1, Section 1.2) by defining behavioural networks (see Research Question 1), modelling and capturing basic HCI affordances on social media (see Research Question 2 and Research Question 3).

In view of Research Question 4, one fundamental way to observe these local interactions is through the use of subgraph analysis using techniques such as network motif analysis [263, 264] and subgraph counting for discovering statistically significant induced subgraphs structures embedded within complex networks.

By definition, an induced subgraph H of a graph G is a graph formed by a subset of nodes V and edges E from the original graph G such that $V(H) \subseteq V(G)$, and $E(H) = \{xy : x \in H, y \in H, xy \in E(G)\}$. These induced substructures are used to consider localised interactions which take place within small groups of users.

Induced subgraphs can vary in size but are often used to focus on interactions in the form of dyads (two nodes), triads (three nodes), tetrads (four nodes) or more. However, due

to the subgraph isomorphism problem many of the algorithms for processing subgraphs scale exponentially with size [102] making it infeasible to detect large subgraphs in near real-time. For this reason, it is common to focus on smaller graphs typically of size two to five.

With respect to analysing behaviour, the use of subgraph analysis can be applied to non-trivial subgraphs. For example, research conducted on Twitter demonstrates how the presence of transitive ties in the form of a feed-forward motif (see Figure 3.7) can be used to influence information diffusion and to predict new followers among users [247].

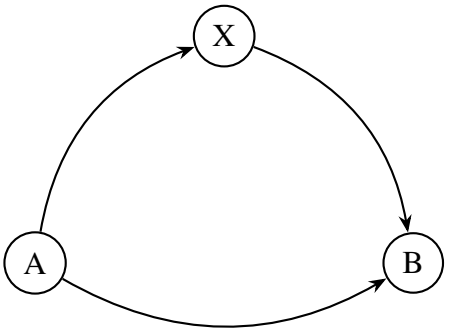

Figure 3.7: Example of a feed-forward motif used to demonstrate how a new connection forms between users A and B via an existing tie through user X. The feed-forward motif also represents the presence of indirect reciprocity in a social network.

As a result, simple triads similar to the one shown in Figure 3.7, can be used to better understand social network interactions and motivates the need to consider local network substructures through techniques such as network motif analysis.

Network motifs are statistically significant subgraphs or patterns which are part of larger graphs and are considered "building blocks" which are used to characterised a complex network by understanding frequent patterns of interaction [339, 263]. Research performed by Milo et al. defines network motifs as frequently occurring n-node subgraphs which are "*patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomized networks*" [263]. Dyads, triads and tetrads are often considered as induced subgraphs such over and under-represented induced subgraphs are called network *motifs* or *anti-motifs*.

Network motifs can consist of subgraphs of any size [393, 290, 398], although it is quite common in the literature to characterises motifs by triadic subgraphs in groups of three nodes [263, 316, 206]. A few more examples can be observed in Figure 3.8.

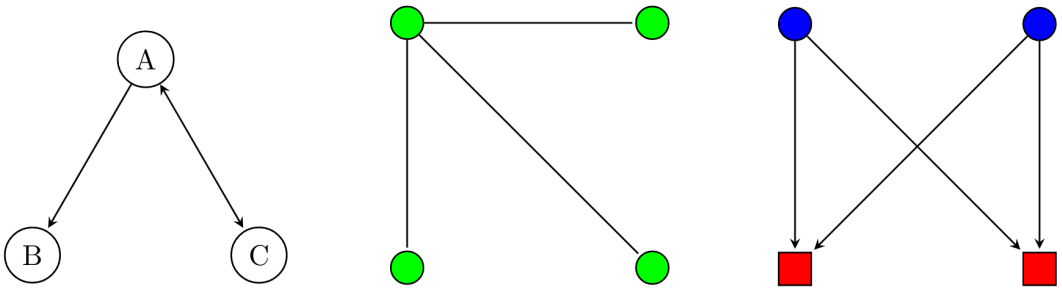

Figure 3.8: Example motifs taking on multiple forms. From left to right, the first subgraph represents a directed triad, second a non-directed tetrad and third a bipartite subgraph with different node types.

The identification of network motifs within complex networks is determined by the following procedures:

- **Subgraph counting** to determine how many induced subgraphs of n-nodes are present in the network (directed or non-directed) (see Section 3.7.1).
- **Significance profiles** are used to observe and compare the significance of each subgraph, based upon the quantity of each subgraph relative to a random network of similar properties (see Section 3.7.2).
- **Null models** govern the structure of the randomised network to suit the relative complexity of the target network in question (see Section 3.7.3).

3.7.1 Subgraph Counting

Counting subgraphs within a network is the fundamental first step. One is required to count the frequency of subgraphs appearing within a network. This process alone presents several key challenges which include the following:

- **Directed or non-directed edges** determines how the counting algorithm will scale based upon the possible permutations to count each distinct type of subgraph. Directed graphs would scale roughly $O(n^2)$ and non-directed would scale $O(\binom{n}{2})$[9].

- **Subgraph size** would determine the subset of possible subgraphs which are counted. As mentioned previously, it is quite common to count triads, although it is possible to count subgraphs of n-nodes including dyads (2-nodes) and tetrads (4-nodes).

- **Structure** governs formation of subgraphs within the network. Subgraphs may include different node types or could be bipartite [290, 398, 401].

- **Reciprocated edges** are edges which refer to the same node on more than one occasion.

In a typical case, subgraph counting is performed on static directed or undirected networks. There are many examples of subgraph counting algorithms available [405, 196, 39] with a few of which that concentrate on subgraphs with as many as four or five nodes each [393, 253, 66].

By convention, work featured by Batagelj et al. [39] is frequently used to produce a census of all distinct directed triads appearing within a network. A simple example of triadic subgraph counting in a small network can be found in Figure 3.9 based upon the cyclic triad. In this example, there are two cyclic triads among nodes $(0, 3, 2)$ and $(1, 8, 4)$.

Subgraph counting algorithms are typically categorised into one of two groups depending upon the scale of the task; subgraph-centric or network-centric.

- **Subgraphs-centric** algorithms focus on searching the entire network for a specific subset of pre-generated non-isomorphic subgraphs. This technique is fea-

[9] This is assuming the worst case scenario and not including rotation

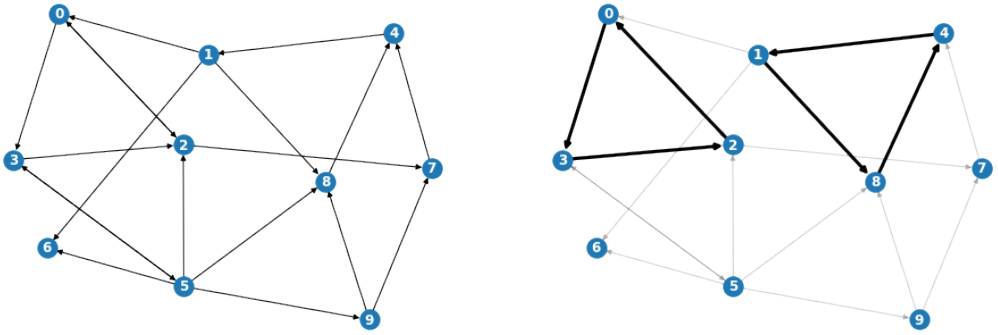

Figure 3.9: Simple example showing how subgraph counting is applied on a random network (left) using the cyclic triad formation as an example. Every instance of the cyclic triad are shown in bold where there are two present in the same random network (right).

tured in algorithms such as MAVisto [332].

- **Network-centric** algorithms use the target network to define a search space for all subgraphs of size n-nodes. This features in FANMOD [393], Kavosh [212] and G-Tries [320].

3.7.2 Significance Profiles

To determine the significance of certain subgraphs within a network, the notion of *significance profiles* are used to understand which subgraphs are more significant than others in a complex network. Researchers Milo et al. determined several solutions for identifying key motifs within networks [263, 264] which are widely adopted within the leading literature [398, 290, 291, 369]. To achieve this, Z-scores and subgraph abundance scores are primarily used to score the value of each triad relative to random networks. The Z-score of a network is characterised by the ratio of the frequency of real subgraph $Nreal_i$ subtracted from the mean of random subgraph $Nrand_i$ to the standard deviation of random subgraph $Nrand_i$. The calculation for determining the

Z-score can be found in Equation 3.1.

$$Z_i = \frac{Nreal_i - < Nrand_i >}{std(Nrand_i)} \tag{3.1}$$

Normalised Z-scores produce a *significance profile* (SP) which is used to understand the relative significance of subgraphs, rather than the absolute significance. This measure ensures that the size of the network doesn't affect the significance of a given subgraph. The calculation for normalising Z-scores is shown below in Equation 3.2.

$$SP_i = \frac{Z_i}{\sqrt{\sum Z_i}} \tag{3.2}$$

Unlike Z-scores, the subgraph abundance scores are used to profile a network which does not depend on the network size. A parameter ϵ is used to prevent a score Δ_i from being misleadingly high when a subgraph makes a rare appearance in the real and random networks. The value of ϵ is usually set to 4 [264, 369] to prevent the result from being misleadingly large when a subgraph rarely appears. The calculation for subgraph abundance scores can be found below in Equation 3.3.

$$\Delta_i = \frac{Nreal_i - < Nrand_i >}{Nreal_i + < Nrand_i > + \epsilon} \tag{3.3}$$

This solution is ideal for studying fewer subgraphs within a network such as non-directed connected tetrads (4-node subgraphs). A normalised abundance score produces a *subgraph ratio profile* (SRP) as shown in Equation 3.4.

$$SRP_i = \frac{\Delta_i}{\sqrt{\sum \Delta_i}} \tag{3.4}$$

These random networks $Nrand_i$ are generated according to a *null model* (discussed in Section 3.7.3). The use of "*cut off values*" are used to set upper and lower-bound thresholds to filter subgraphs into two groups. A significance value SP_i which exceeds

the threshold is known as a motif and a significance value SP_i which is below the threshold is known as an anti-motif. Any value which is not above or below the cut-off is rejected.

Distinct profiles are grouped into what are known as "*superfamilies*". A superfamily is used to describe a network's significance profile which is either distinct in formation or follows a structure which closely resembles a profile which is somewhat similar in type. In short, a network consists of a combination of triads resulting in the significance profile. Networks having common significance profiles are said to form "super-families" of networks. For example, Milo et al. proposed a superfamily through linguistical structures using the English, French, Spanish and Japanese languages. These profiles closely resemble each other when the text of each native language is processed as a network of word n-grams [264].

3.7.3 Null Models

As introduced in previous sections, random networks $Nrand_i$ act as a baseline comparison to understand if a subgraph appears more in a target network $Nreal_i$ than a network which is purely random. These random networks are generated according to a *null model* defined by certain properties and parameters. The choice of a null model is used in turn to highlight the significance of a particular subgraph for a given network as well as to suit computational requirements. Null models can take on many forms, however, the literature primarily focuses on three main solutions.

- NM_1: Same number of nodes and edges
- NM_2: Same distribution of dyads (bidirectional, single and null)
- NM_3: Same in-out degree distribution of nodes in the target network

The complexity of these null models can vary depending on size of the graph under analysis. Community classification problems with motifs are best performed under

MN_1 as this offers accuracy as well as performance for large scale complex graphs [368, 162]. This is typically achieved through the use of edge switching by producing random permutation of a network through random edge shuffling thus preserving the number of nodes and edges present within the original network. The second null model (NM_2) is ideal for graphs that concentrate on studying the reciprocation of interactions between nodes [351]. Links that are fully reciprocated are bidirectional and partially reciprocated are unidirectional. Furthermore, the use of NM_3 is used to preserve the specific relationships between nodes of different types [204] which can be achieved though the use of parametric models and the shuffling of links through random edge selection [222, 251, 387]. Modelling the degree distribution of networks, especially with high in-out degrees, have proven to be useful in examples such as the World Wide Web which contains nodes with relatively large connected components [284]. More complex examples of NM_3 have been used to model the strength of weighted edges. Alternatively, this can be broken down into multiple model to preserve the in-out degree, in-out degree strength, strength of each vertex and weighted average [351, 223].

The use of random null models has been the recipient of criticism within the literature based upon the central argument that they can distort results without preserving original detail [21, 108]. An example of this can be seen through the use of egocentric and bipartite networks, where random network permutations may create subgraphs which are impossible to recreate within the target network, thus providing misleading results. For example, a null model for an egocentric network might join two non-ego nodes together which draws attention away from the central ego. The choice of null model is subject to careful consideration as the random network must respect the topology of the original network.

3.8 Conclusion

To conclude, the contributions of this chapter are multifold. Firstly, this chapter introduces a novel framework for categorising social media platforms though the concept of data structures (see Section 3.2) that are manipulated through high-level HCI affordances (CRUD). Secondly, this chapter also introduces three network representations (see Section 3.5) as a framework intended to capture human behaviours on social media platforms. For the purposes of this thesis, these frameworks are closely examined on Wikipedia, Reddit and Twitter as platforms of interest (see Section 3.3) to explore the hypothesis of this thesis. Section 3.4 demonstrates how each of these platforms applies to each of the four data structures. Finally, Section 3.7 introduces subgraph analysis as methods for exporting complex networks and modelling user interactions on social media.

As a result, the framework provided in this chapter allows for new platforms to be rigorously considered for analysis using the language-agnostic network representations. These networks represent a general framework that is explored in the remaining chapters of this thesis. This is outlined in Section 3.6 (see Table 3.20) and justifies the choice of platforms for further consideration by using the two frameworks (data structures and networks) as dimensions to map out the investigations undertaken to explore the hypothesis. This, in turn, provides the foundations for Chapters 4, 5 and 6 using the three network representations introduced in this chapter.

Chapter 4

Transitional Networks

4.1 Introduction

The focus in this chapter is to assess the extent to which potential signatures of disruptive activity on social media can be detected using a simple temporal network representation of user activity, known as a *transitional network*. Two particular forms of transitional network are defined - one to capture the sequencing of different users' activity concerning particular content (as shown in Section 4.4 using Wikipedia), and one to capture the sequencing of different users' posting activity in response to each other (as shown in Section 4.5 using Reddit). These formulations allow the approach to be applied in a wide range of social media platforms.

As mentioned towards the end of Chapter 1, Research Question 2 addresses the need for multiple network-based representations as part of a framework for identifying and detecting disruptive activity on social media. This chapter contributes to Research Question 2 by modelling activity though the concept of transitional networks as a network-based representation for capturing user affordances (see Research Question 3) though switching behaviour, which can be used as part of as part of a wider collection of network-based representations.

Much like other network-based representations, the motivation and rationale for the transitional network comes from the observation that irrespective of language or content, the temporal behaviour of those motivated through the content, relative to other users,

may be sufficiently distinctive to leave a signature that is distinct from "average" users taking part in the normal discourse seen in social media. More specifically, the transitional network itself is designed to capture interactions in the form of "transitioning" or "switching" behaviour based upon interactions derived from the natural process of users navigating from one item to the next over time.

Overall, the methodology presented in this chapter provides valuable insights towards the modelling of user activity in the form of capturing the "switching" behaviour between two states. Additionally, this approach utilises the temporal component of user activity - a fundamental feature of modern social networking platforms. This demonstrates the versatility of this network representation, as it can be used across multiple platforms, similarly facilitating the comparison and differentiation of user activity and migration.

As mentioned in Chapter 3, Wikipedia and Reddit align to the collaborative and feed data structures respectively. These are shown in Table 4.1 (relative to the work of this thesis) using the collaborative and feed data structures.

		Data Structure			
		Community	Message	**Collaborative**	**Feed**
Platform	**Wikipedia**	N/A	N/A	***See 3.4.1***	N/A
	Reddit	*See 3.4.2*	*See 3.4.2*	N/A	***See 3.4.2***
	Twitter	N/A	*See 3.4.3*	N/A	*See 3.4.3*

Table 4.1: Relationship between platforms of interest and all data structures with the appropriate cells concerning the work of Chapter 4 highlighted in bold.

These data structures help inform the network structures (transitional networks) of this chapter and, in doing so, attempt to derive behavioural networks from different social media platforms (see Research Question 1) and represent diverse affordances (see Research Question 3).

In its most simplistic form, the *Timestamp* field of the collaborative and feed data structures can be used to represent a series of actions or queries which can be associated with a user's profile or any item. For example, a Twitter profile reveals a list of their

most recent tweets that they produced at a specific point in time. An example is shown in Figure 4.1.

Figure 4.1: Example screenshot demonstrating the order of activity in the form of tweets on Twitter. A transitional network can be formed by modelling activity around a user in time-series order.

The research presented in this chapter introduces the concept of the transitional network inspired by process mining - a method for utilising time-series event logs as a data source. Process mining is used to produce meaningful insights towards understanding the performance of complex operations with the aim of improving efficiency. Existing research has shown how data collected through process mining can be modelled as a social network to depict the flow of operations from one user to the next [373, 347]. For example, work performed by Van Der Aalst *et al.* uses social network analysis on event logs with the intention of understanding collaboration and user interactions within a complex process [373].

As a result, this chapter demonstrates, through the concept of *transitional networks*, how a technique similar to process mining applies to social media for detecting disruptive activity. This chapter adopts similar methods featured within [373, 347] by using transitional networks to model the flow or "switch" from one item to the next based upon their position in time-series order.

As shown in Chapter 3, platforms, such as Reddit, naturally exhibit a lot of switching behaviour due to being associated with the feed data structure. Alternatively, platforms such as Wikipedia are represented by the collaborative data structure, which feature a temporal component in the form of a timestamp. For this reason, it is necessary to distinguish between the two different types of transition behaviour by observing user activity from two perspectives; *users* (Reddit) and *content* (Wikipedia).

Wikipedia and Reddit have a collective goal to produce unique contributions which are valuable and insightful both to the localised community and general internet at large. In the context of Wikipedia, users (or editors) seek to offer their contributions to an article by modifying existing content as contributed by other user within the community. Reddit, by comparison, encourages their users or "Redditors" to contribute submissions to individual communities known as "subreddits" in which users receive "karma" (a crowdsourced metric using user votes based upon the popularity of their submission) in return.

4.1.1 Contributions

The overarching contribution of this chapter considers the effect of bad actors present within the network and how the process of collaboration (as a whole) is affected. In addition to this, two other minor contributions are produced. Firstly, this chapter considers the role of induced substructures in online process mining with the intention of simplifying network complexity by considering the role of local substructures. Secondly, this work applies transitional networks to two similar yet different social networking

platforms with the intention of understanding patterns of interactions in a collaborative setting. In doing so, this research demonstrates how the methods presented are transferable between multiple online platforms by using Reddit and Wikipedia as examples. Both platforms share similarities, which depend upon user contributions to an article or community.

As outlined in Chapter 1, this chapter also offers broader contributions which impacts this thesis as a whole. This chapter contributes to the work of this thesis by demonstrating the value of transitional networks as a **behavioural network** for analysing the behaviour of users of social media, which, in turn, satisfies the objectives of Research Question 2. Furthermore, we demonstrate how the methods developed in this chapter address Research Question 4 by detecting disruptive behaviour in the form of exchanges "switches" between users within controversial topics in Wikipedia articles.

4.1.2 Transitional Network Construction

As described in Chapter 3, Section 3.5.1, transitional networks are dependent on two fundamental components: *users* and *content* nodes. Users are defined as agents who are responsible for contributing new content within the platform. *Content* represents a collection of user contributed data in the form of text, images, videos or hyperlinks grouped under a unified topic. User contributed content is created and/or modified at a specific point in time.

As mentioned in Chapter 3, Section 3.3.1, using Wikipedia as an example, a *user* contributes to a given Wikipedia article by making modifications. Wikipedia's articles are considered as the *content*, as they can be created and revised by a community of editors. Likewise, Reddit *users* provide submissions in the form of hyperlinks to external content or text posts within a particular subreddit. Reddit also features *content* in the form of subreddits, as users can make contributions via novel submissions and comments.

With respect to Research Question 1, behavioural networks are generated by pairing associated items together using time-series data (e.g. events in an action log). The direction of the edge represents the order in which one entity appeared other its predecessor. For example, in the relation $A \rightarrow B$, A occurs after B. Examples of theses are shown in Figure 4.2 and 4.3.

Figure 4.2: A simple example demonstrating the utility of transitional networks used on Wikipedia where nodes serve as *content* and edges represent a transition "switch" from one article to the next.

The movement of users between content produces potential behavioural signals with many degrees of freedom. In this case, nodes within a network can represent transitional behaviour by means of switching between users or content within a temporal network. As show in Figures 4.2 and 4.3, the role of nodes are interchangeable, whereas edges are primarily used to denote the order of interactions that appear in order. This rule can be used to infer implicit interactions between the two nodes within the network (such as causality).

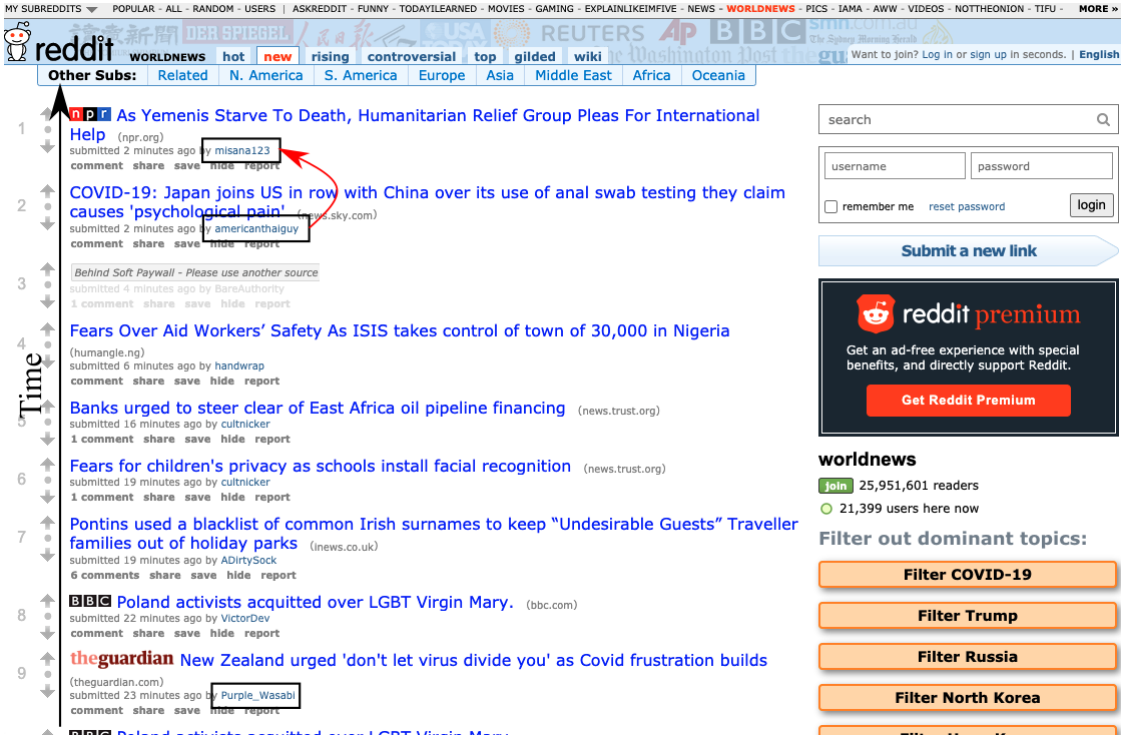

Figure 4.3: Similar to Figure 4.2, a transitional network is used on Reddit where nodes can represent *users* and edges represent a transition "switch" from one user to the next.

4.2 Motivation

The exploration of transitional networks in this chapter are motivated by real-world instances of disruptive behaviour (see Chapter 2, Section 2.1) which emerge on online social networks. Wikipedia and Reddit are used as example platforms where disruptive activity can occur as part of the community at large. Both platforms rely on various crowdsourcing mechanisms and community engagement for generating contributions to the platform.

Using Wikipedia as an example, many users adopt the platform as a point of reference to engage large quantities of information. As a consequence, the platforms are susceptible to vandalism as a bi-product of allowing anonymous users to edit content. This can cause major disruption. In a few recent examples, "vandalised" content as well as other

hoaxes have been known to make an appearance within Google's research results[1]. More specifically, well researched topics such as climate change have been the recipient of repeated attacks, such that the community are continuously required to revert content and to make amendments to ensure that the supported material is consistent and reliable[2]. Traces of vandalism and attacks are preserved through the use of the revision history log, such that attempts to remove vandalism are frequently marked with a "revert" so the latest content overrides the vandal's contribution.

Other social networks such as Reddit serve as a platform for aggregating and digesting news articles, such that registered users can share links and participate within community discussion. Trolling is a regular issue within certain communities, such that known trolls have been known for developing collectively elaborate posting strategies across the platform to produce convincing, yet fake, activity [42, 141]. To conclude, the heavy reliance of these platforms for everyday usage by ordinary internet users, motivates the need for solutions to combat disruptive activity across multiple channels.

4.3 Approach

		Data Structure			
		Community	Message	**Collaborative**	**Feed**
Network	**Transitional**	-	-	**See 4.4 (Wikipedia)**	**See 4.5 (Reddit)**
	User-to-user	-	See 5.4, 5.5 (Twitter, Reddit)	-	-
	User Association	See 6.4 (Reddit)	-	-	-

Table 4.2: A replica of Table 3.20 featured in Chapter 3 outlining the investigations of this thesis (with respect to data structures and network representations) with the appropriate cells highlighted in bold which refers to the problem space which Chapter 4 seeks to investigate.

[1] https://www.poynter.org/fact-checking/2018/wikipedia-vandalism-could-thwart-hoax-busting-on-google-youtube-and-facebook/

[2] https://mashable.com/feature/climate-change-wikipedia/?europe=true

As indicated in Table 4.2, this chapter utilises the transitional network for detecting disruptive activity in the form of switching behaviour using both Wikipedia and Reddit, which align with the collaborative and feed data structures respectively. As previously mentioned, this is investigated by observing two forms of transitional network; content-oriented (Wikipedia) and user-oriented (Reddit).

As recalled in Chapter 3, Section 3.5.1, transitional networks are designed to capture user behaviour in the form of switching behaviour from two perspectives using time-series ordered data; A single user switching between content and multiple users switching over a single piece of content. As a result, these two perspectives both make use of the transitional network and are used to help address Research Question 1 by defining behaviour networks based upon social media activity and Research Question 2 by contributing to a framework of alternative network-based representations. Furthermore, we address Research Question 4 by classifying disruptive and non-disruptive activity through the use of motif analysis.

These two perspectives involve inferring implicit connections based upon ordered time-series data. We consider this using two platforms as follows:

- **Section 4.4: Wikipedia - Content-oriented**: Detecting anomalies in the structure of revision networks for controversial and non-controversial Wikipedia articles with prediction. User switching behaviour is positioned around individual articles - the *content*.
- **Section 4.5: Reddit - User-oriented**: Navigating social media and understanding user flow / migration in the form of "switches" from one piece of content to the next. Interactions are based upon a single user's activity.

Figures 4.4 and 4.5 demonstrate the two perspectives in which switching behaviour can be modelled. In the case of Wikipedia, the revision history (as shown in Figures 4.4 and 4.6), uses the revision log of an article such that a network is formed using the rule where

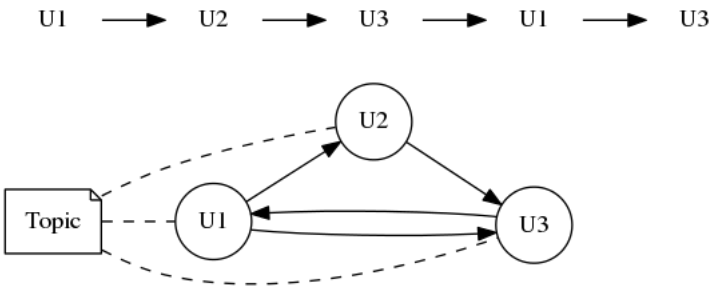

Figure 4.4: Example of a simple transitional network based upon user switching behaviour around a piece of content "Topic" derived from a sequence of activity (top).

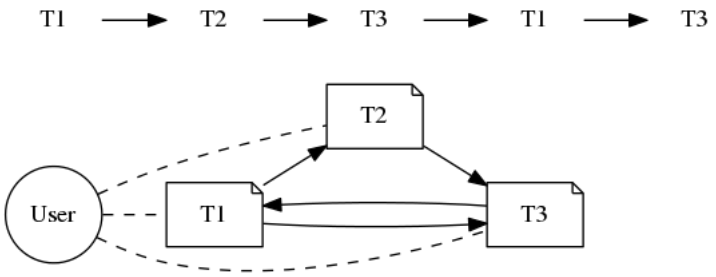

Figure 4.5: Example of a simple transitional network based upon the content switching behaviour of a user "User" derived from a sequence of activity (top).

user U_j edits after U_i producing $U_j \rightarrow U_i$. These are known as *revision networks*, which refers to transitional networks which represent activity derived from an article's revision log. Within larger networks, the task of discovering pairs of users becomes much easier as features such as edge weight, reciprocity and burst rate can reveal distinct behavioural types between the users. As well characterising the behaviours of individuals, these features also describe attributes belonging to the community and the network as a unit.

Reddit also lends its self to representing user behaviour concerning transitional networks, where users transition between any subset of communities or "subreddits" (see Figure 4.5). As a result, this approach considers the significance of content (irrespective of type) in relation to the order they appear. Furthermore, this method can be used to consider the importance of a subreddit with respect to its positioning relative to others. A formal definition of edge types and node values are featured in Table 4.3 for each

network.

Application	**Wikipedia Revision Network**	**Reddit Subreddit Switching**
Source Node	Editor (at t_i)	Subreddit (at t_i)
Edge (directed)	*edits after*	*posts in after*
Target Node	Editor (at t_{i+1})	Subreddit (at t_{i+1})

Table 4.3: Edge list definitions for each transitional network used in Chapter 4.

4.4 Content-oriented Transitional Networks on Wikipedia

Wikipedia has become a tremendous platform for crowdsourcing knowledge, representing a cornerstone of the World Wide Web [118]. It allows the "wisdom of the crowd" [360] to potentially emerge, providing intelligence on a vast range of topics [59]. However, complex dynamics support the emergence of content, since the formation of Wikipedia articles involves both human cooperation and human conflict, based on the extent of convergent and divergent views. Narrative and counter-narrative frequently jostle for presence in articles, representing a source of friction that is seen through editor interaction [334] and in the semantics of article content [316]. Wikipedia conveniently provides a list of controversial content that are labelled by the Wikipedia community themselves[3].

Our focus is to further understand the relationship between small groups of users, as induced by their editing sequences, by using a *revision network* - an example of a transitional network. This does not require information on the nature of the editing undertaken - it simply captures the ordering in which editing occurs and is therefore a simple metric to infer. Editors are represented by nodes, and a directed edge from node A to B indicates that "*Editor A edits the article after Editor B*" (see Figure 4.6). In doing

[3] `https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues`

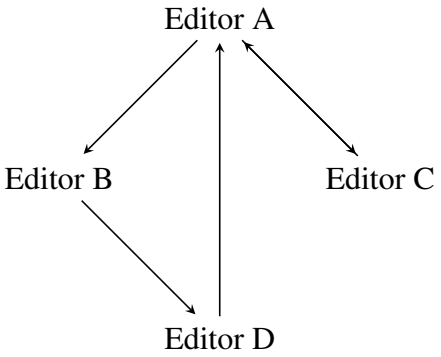

Figure 4.6: A network generated using the editor order A, B, D, A, C, A, from the newest edit to oldest. Each editor is characterised by a letter A to D, and each occurrence in the list marks a single revision of the article by the corresponding author. An edge is formed between the current editor and its adjacent neighbour in the sequence, forming the directed edges $(A, B), (B, D), (D, A), (A, C), (C, A)$.

so, this can determine the extent to which controversial articles exhibit a distinctive signature relative to those that are deemed non-controversial.

There has already been some consideration of revision networks in the literature [192, 214], where more recently the emphasis [397] has been to combine them with other network representations other than transitional networks. However, given the fundamental nature of revision networks, it is interesting to question the extent to which they hold sufficient information to characterise controversial Wikipedia articles - an environment of highly motivated users. Currently, this is not well-understood, and motivates this research.

4.4.1 Related Work

Understanding the content of crowdsourced platforms such as Wikipedia and the behaviour of their contributors is of wide research interest [360]. Wikipedia represents a dynamic network of articles with a structure resembling that of the World Wide Web [418], whereby dominant articles act as connectivity hubs. Dynamics also exist within the formation and maintenance of individual Wikipedia articles, through open and collaborative editing.

In an age of misinformation [4, 268, 377], understanding characteristics of controversial articles has increased in importance. Because of the controversial nature of some topics, the narrative in a Wikipedia article may contain misleading information that stops a neutral consensus emerging. Prior work in this area has established insights such as the predictability of controversy from editor behaviour [334], such as deletions, reversions, and statistics from the collaboration network, prediction of article quality taking insights from multiple models [400], and interactions between users, bots, admin and pages [204]. There have also been a number of different types of network developed to assess Wikipedia articles, including collaboration networks [62] that capture the positive or negative relationship between editors, edit networks that capture "undoing" of edits by a third party [217] and affiliation networks [207].

Interactions between editors range from positive to negative, where debates and arguments lead to different patterns of revision (e.g., [316, 233]), capturing behaviours such as vandalism [311] and the propagation of disinformation [225]. Characterising articles and contributors through revisions provides a means for Wikipedia to manage and review its content. This is potentially labour-intensive and has led to interest in creating and exploiting methods to detect issues (e.g., [6, 334]).

Controversial articles have become an increasing point of focus, and characterised as such by Wikipedia. Automated methods for classifying articles have received much attention (e.g., [316, 401, 397]). The associated revision log for Wikipedia articles has been shown to provide a basis to examine potential controversy through examining the collaborative behaviour of individual editors within an article [334] or across multiple articles [398]. An article's revision log identifies the structure underlying temporal interactions [397], and provides insight into how articles and contributors' habits may evolve over time [204]. Features from the aggregation of this, such as number of edits, revision, and previous version restorations have been shown to correlate (e.g., [334]).

Treating the revision log as a network between editors [334] has been shown to provide additional useful features using graph theory and social network analysis techniques

(e.g., [316]). This has ranged from global features such as the degree distribution (e.g., [334]), through to analysis of local sub-structures concerning the articles with which editors interact (e.g., [398]).

However, there has been little investigation of controversial articles based on the under or over representation of local-substructures. As introduced in Section 3.7, network motif analysis, originated from biology [264, 61], has been used to good effect in characterising other complex networks, including technology (e.g., [290]). In terms of Wikipedia, motif analysis has been used to determine how articles point to each other [418] and in assessing interactions between editors and different Wikipedia articles [398].

4.4.2 Hypotheses of Content-Oriented Transitional Networks

Using a transitional network, *interaction differences between small subgroups of Wikipedia editors is sufficient to distinguish between controversial articles and non-controversial articles*. This chapter considers the extent to which revision networks of Wikipedia articles have different local induced substructures based on their having controversial classifications. In doing so, this also address Research Questions 1, 2 and 4 of this thesis by using a *revision network* as a technique for building behavioural networks to capture disruption which is likely to take place among controversial topics. In line Research Question 4, this approach is based on techniques from complex networks [263, 264], that have been successful in classifying diverse and complex biological networks based on their latent induced subgraphs.

To achieve this at scale, and in contrast to previous literature [334, 316, 204], we adopt a relatively large sample of Wikipedia articles, involving over 21,000 Wikipedia articles, by determining their subgraph ratio profiles. Each such profile represents the under and over representation of induced triads in the revision network of a Wikipedia article using 13 dimensions of connected triads, while also normalising for differences in network

size.

4.4.3 Methods

Dataset

To consider transitional networks for Wikipedia, we focus on revision history logs and article meta-data of a sample of Wikipedia articles ($N = 21,631$) through Wikipedia's web API[4]. The revision logs contain time-series events and meta-data attributing the revision to a particular user at a given time. Within this set of articles, a subset ($N = 2,661$) are considered to be 'controversial' as they were listed in Wikipedia's "List of controversial issues"[3]. This provides a convenient, labelled sample of controversial articles. The remaining articles ($N = 18,970$) are random articles that are not contained in the controversial issue list to serve as a basis for comparison. These were taken from an original sample of 23,000 articles (20,000 non-controversial and 3,000 controversial), from which articles were removed if they did not contain sufficient information for motif analysis.

Network Construction

For each article, a revision network is constructed (in the same manner as [214, 397]) where nodes represent unique editors and directed edges show that an editor added a revision after another editor. The revision log list was traversed to build a network that spans the article's lifetime, adding nodes and edges as they appear in each event. Specifically, let the revision network of a Wikipedia article be defined by $G = (V, E)$, where each editor is represented by a node $v \in V$. An edge $(v_i, v_j) \in E$ indicates that editor v_i edits the article after editor v_j. This excludes self-loops, and editor v_i editing after editor v_j multiple times does not result in a multi-edge. A simple example can be found in Figure 4.6, which describes how the network is constructed.

[4] `https://en.wikipedia.org/w/api.php`

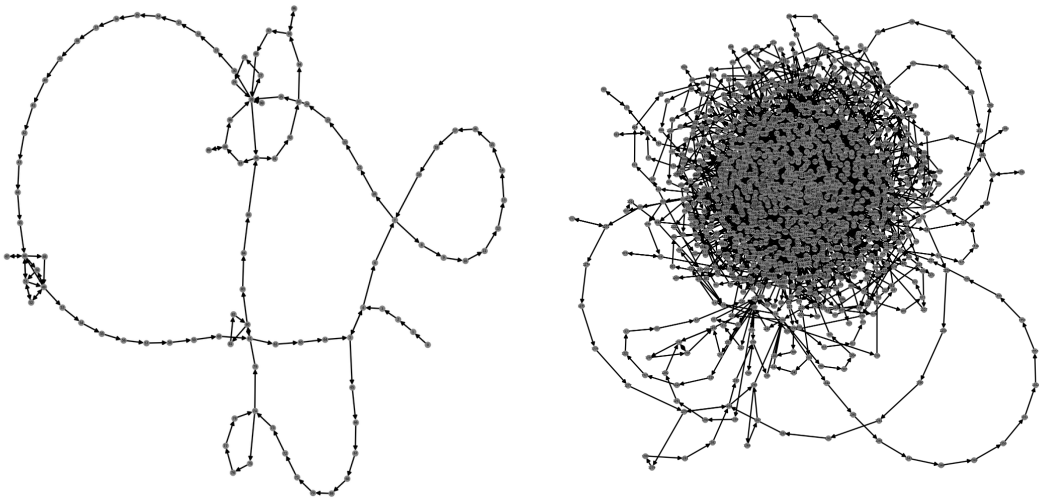

Figure 4.7: Revision network of two articles - A non-controversial article (left) of *The Web Conference* and a controversial article (right) of the *Brexit* Wikipedia article. The distinct contrast between the two reveals the complexity of revision networks as more information is aggregated over time.

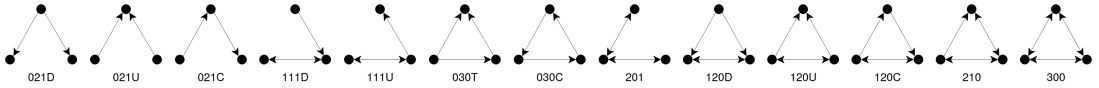

Figure 4.8: 13 possible combinations of connected triads in directed networks where each triad is assigned a unique code (bottom) according to the triadic census algorithm.

Two extreme examples from the dataset shown in Figure 4.7. This form of representation is potentially useful as large articles do not typically follow a linear or incremental structure. For example, it is highly likely that users will restore work back to an earlier revision should a revision become vandalised or irrelevant. Editors are likely to refer back to previous editor's work. These behaviours result in complex sequential patterns that are captured through revision networks.

Network Motif Analysis

As described in Section 3.7, network motif analysis focuses on determining the under or over representation of induced subgraphs [263, 61], as compared to an alternative sample of graphs (i.e., a null-model that acts a relevant baseline for comparison). Furthermore,

the use of motif analysis is used to address Research Question 4 by considering the extent to which local features (i.e. motif analysis) can be used to detect disruptive behaviour.

To perform motif analysis, an article's revision network is analysed using triads, representing how all possible triples of editors may sequentially interact. Triads are sufficiently large enough to capture both direct and indirect reciprocity between editors, while not being of a scale that is impeded by computational complexity - there are 13 possible connected triads, shown in Figure 4.8. The coded names listed in Figure 4.8 are provided as part of the convention used in the triad census algorithm [39].

For each article, the subgraph ratio profile (abbreviated as SRP) as defined by Milo *et al.* [264] is calculated (see Equations 3.3 and 3.4). This accounts for variations in network size. This is achieved by determining the relative abundance of each type of triad compared to random graphs generated by the null-model. For each type of triad i, Δ_i is calculated (see Equation 3.3). In this case, the null model uses 100 random graphs with the same number of nodes and edges as the graph under observation. This process is repeated for each triad i and normalised across triads to form the *subgraph ratio profile* (SRP) for a given network. The i^{th} SRP, denoted SRP_i, denotes the extent of under or over representation of the triad i, as defined in Equation 3.4.

The SRP composed for each article provides a 13-dimensional vector whose components indicate the extent of triad representation relative to networks in the null model. To assess these collectively, principal component analysis (PCA) is used, which allows for the SRPs to be considered in a lower dimensional space. In this case, both three and two-dimensional PCA is used. Finally, comparisons are made with a number of external variables (including number of editors/nodes, age of article, and revision rate) to understand potential correlations with motifs.

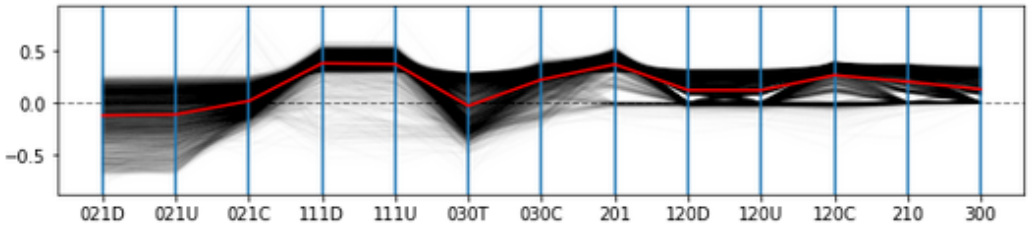

Figure 4.9: Subgraph ratio profiles of all controversial articles as an overlapping line plot to show the distinct formation of triads in each article. The average profile is displayed in red.

4.4.4 Results

Motif Analysis

The SRPs that arise from both controversial and non-controversial articles are first examined in isolation (Figures 4.9 and 4.10 respectively).

To find significant triads, a cut-off value of +0.3 and -0.3 was determined based upon a combination of findings informed from [263, 264, 229] and visual estimation. These results determine that controversial articles are strongly represented by triads `111D`, `111U` and `201`, which attain average *SRP* scores of 0.382, 0.375, 0.372, with relatively low dispersion (SDs of 0.136, 0.149 and 0.124 respectively). Together these represent a chain of three nodes, where one edge is reciprocated, with the other edge covering all possible directional types (i.e., reciprocated, directed in, directed out).

In contrast, the results for the non-controversial articles provide a different profile. Here `021D` and `021U` are significantly underrepresented (average *SRP* scores of -0.511, -0.485), albeit with higher standard deviations present (SDs of 0.192, 0.2). Interestingly, these anti-motifs (`021D` and `021U`) relate to a lack of subgraphs where directed edges either emanate from or are received by a single node in the triad. Such configurations relate to the role of a mediating editor that may be presented with or respond to the editing of others. In other words, such mediators have a reduced role in non-controversial articles.

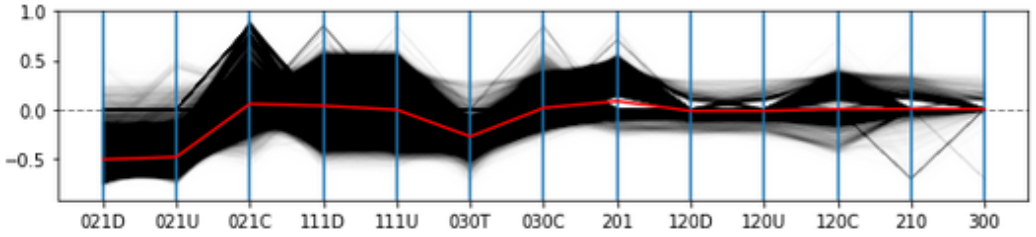

Figure 4.10: Subgraph ratio profiles of all non-controversial articles as an overlapping line plot to show the distinct formation of triads in each article. The average profile is displayed in red.

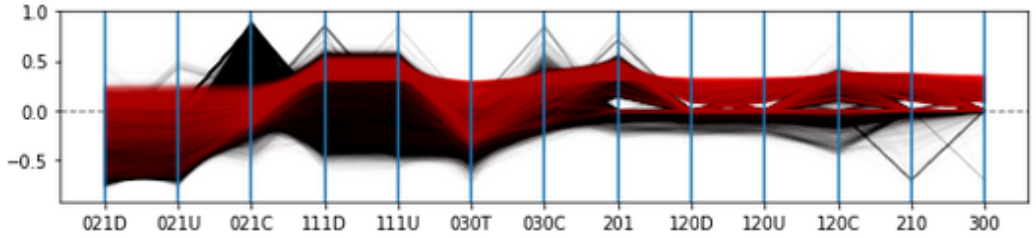

Figure 4.11: Subgraph ratio profiles of all articles, with controversial articles displayed in red and non-controversial in black revealing the distinct signature of controversial articles compared with non-controversial articles.

The comparison between these subgraph ratio profiles is shown in Figure 4.11. These profiles are quite distinct when compared. In addition to this, the Pearson correlation coefficient is determined for each distinct pair of articles in three groups - controversial articles, non-controversial articles and all articles. Controversial articles provide the greatest correlation to each other (M=0.41193, SD=0.41262). Non-controversial articles have a lower mean correlation (M=0.37498, SD=0.36749) which is similar to the result when considering all articles together (M=0.37569, SD=0.36844).

Principal Component Analysis using SRPs

The 21,631 subgraph ratio profiles are analysed to determine the relationship in terms of relative clustering. PCA is applied in order to reduce the 13 dimensions of the SRPs down to a lower dimensional space which is more easily interpret-able. The SRPs are projected into 3-dimensional space for clarity, as seen in Figure 4.12. This presents a

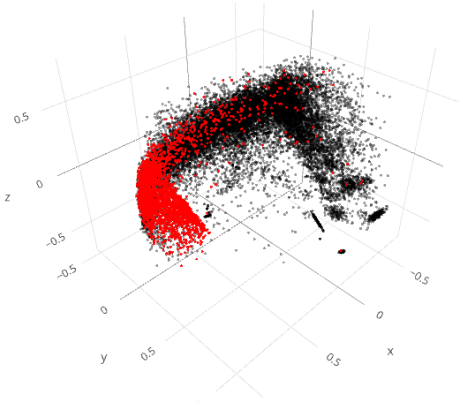

Figure 4.12: PCA scatter plot repression of the 13-point feature vector in 3D clustering space where red points represent controversial articles and black points represent non-controversial random articles. The PCA plot emphasises the unique clustering behaviour of controversial articles with respect to non-controversial articles.

distinctive region where controversial articles are dominant. This provides evidence for a distinction between the controversial and non-controversial articles, consistent with the variation in motifs identified in the previous section.

The PCA coefficients (Table 4.4) that define the three dimensions reveals that principal component one (x axis in Figure 4.12) primarily depends on triads `111D` and `111U`. Principal component two (y axis in Figure 4.12) primarily depends on triad `021C`. The third principal component primarily depends on triads `021C` and `030C`. Additional observations are made when represented in three-dimensional space, as the revision networks have limited dispersion in the third dimension (i.e., vertical dimension as plotted).

Calculating the percentage of explained variance by principal component confirms that the first principal component produces 53.7% of the shared variance, the second produces 22.3% and the third produces the least with 6.7%. This confirms that the third principal component provides a limited contribution to representation of the total variance across the significant ratio profiles. This supports representation through two principal components, as plotted in Figure 4.13, with the relative composition of each principal component being near identical to PC-1 and PC-2 in Table 4.4. As

	PC-1	PC-2	PC-3
021D	0.33	0.42	0.07
021U	0.321	0.43	0.12
021C	-0.12	0.64	-0.6
111D	0.5	-0.27	-0.39
111U	0.52	-0.24	-0.33
030T	0.14	0.2	0.13
030C	0.2	0.1	0.44
201	0.31	-0.05	0.11
120D	0.07	0.05	0.13
120U	0.07	0.06	0.13
120C	0.2	0.04	0.19
210	0.11	0.07	0.16
300	0.06	0.06	0.11

Table 4.4: PCA coefficients displaying the strongest triads.

anticipated, this is a similar dependency on the first and second primary components in the three-dimensional representation.

Representation in two dimensions further clarifies the distinction between controversial and non-controversial revision networks. In particular, from Figure 4.13 both classes of article exhibit a similar maximum and minimum range against principal component two, which is primarily defined by the linear path between three nodes (`021C`). However, it is the variation in the first principal component, dominated by `111D` and `111U`, which represent linear paths with reciprocation on one edge, that differentiate the non-controversial from controversial. High values in principal component one correlate with controversial articles - in other words, controversial articles exhibit more reciprocation on top of possible linear paths.

Combining Results With Additional Metadata

The relationship between revision networks and primary external variables (number of editors/nodes, age of article, and revision rate) using motifs is assessed. Specifically, using the dimensions of two-dimensional PCA analysis, the correlation with the external

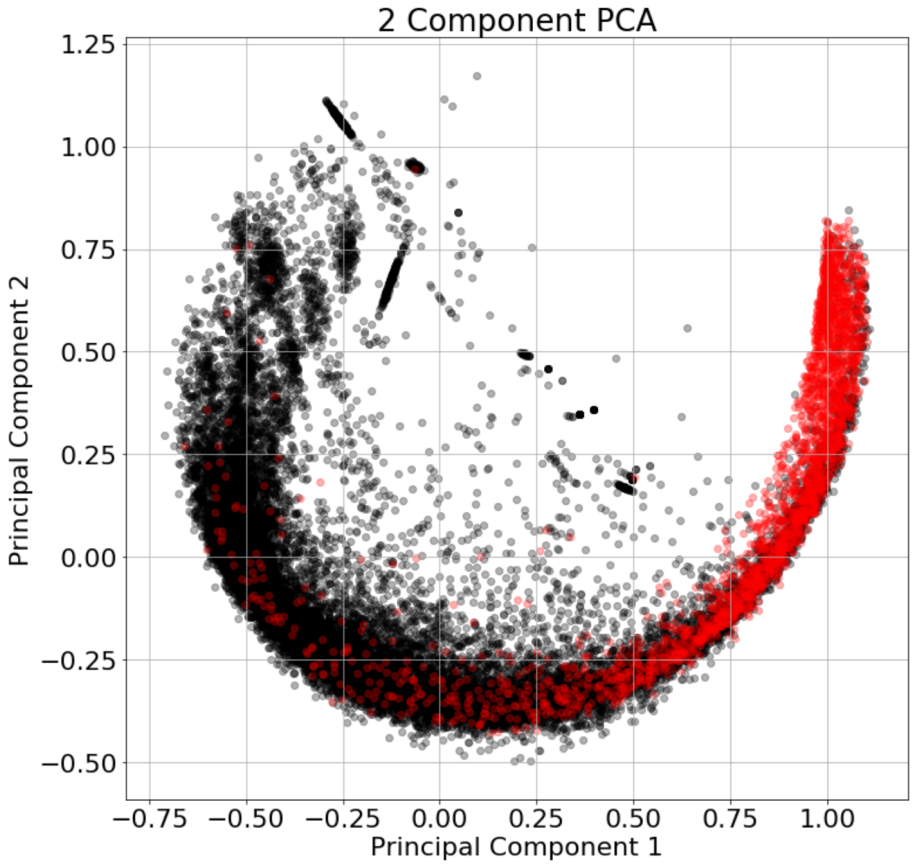

Figure 4.13: PCA scatter plot repression of the 13-point feature vectors in 2D clustering space. Similar to Figure 4.12, the PCA plot emphasises the unique clustering behaviour of controversial articles with respect to non-controversial articles when reduced to two dimensions.

variables is examined, and how this differs between controversial and non-controversial articles. The results are shown in Figures 4.14, 4.15 and 4.16.

The greatest differences between controversial and non-controversial articles occur with respect to article age (Figure 4.15). Here, controversial articles with high age cluster against high values of principal component one, and to some extent this occurs for principal component two. This contrasts against the clustering seen for non-controversial articles.

Figure 4.15 also shows that controversial articles have a tendency to be older. The high values in principal component one and two which align with dense clustering

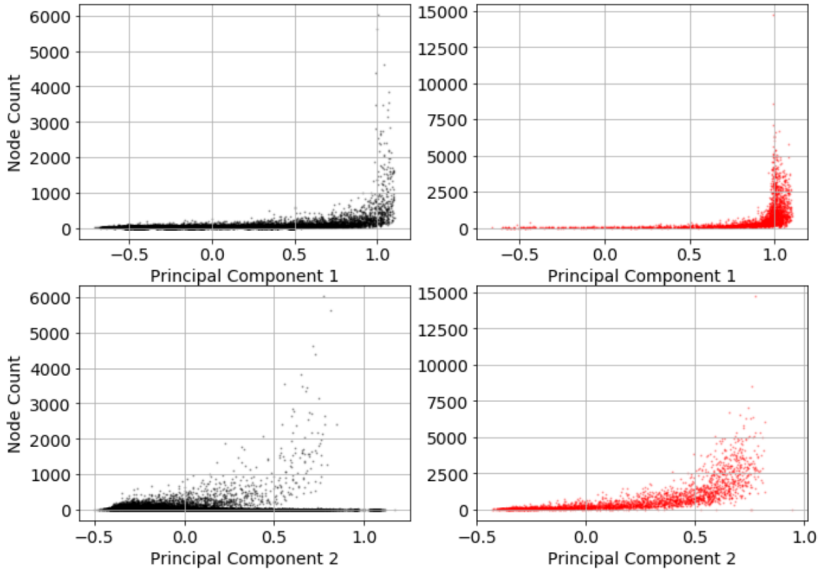

Figure 4.14: Scatter plot of each principal component combined with node count for both non-controversial and controversial articles. The figures reveal a correlation between the two variables for both controversial and non-controversial articles.

of controversial articles show that while such articles accumulate the linear revision path between authors (`021C` which dominates principal component two), controversial articles also accumulate instances of linear paths where one edge is reciprocated (i.e., `111D` and `111U` which dominate principal component one).

Dyadic Analysis

In addition to triad based analysis, we further include analysis of dyads in an attempt to understand the extent to which they help to explain the complex nature of interactions. The frequency of dyads is applied using a separate counting procedure from that of the triadic census algorithm [39]. This is to ensure that dyads are counted as pairs and not as triads.

The number of dyads is determined by counting the number of times a distinct bidirectional link $A \longleftrightarrow B$ (`102`) appears between a pair of nodes and the number of times a unidirectional link $A \longrightarrow B$ (`012`) appears separately.

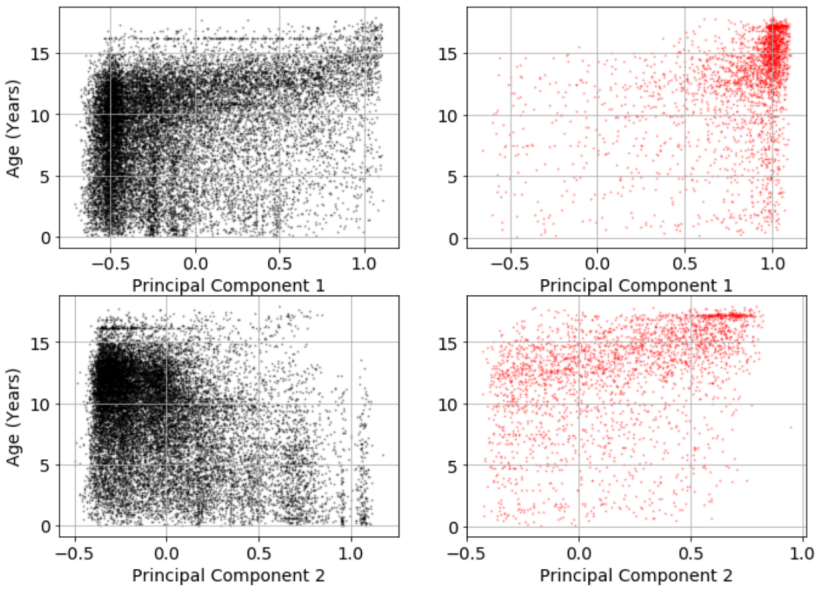

Figure 4.15: Scatter plot of each principal component combined with article age for both non-controversial and controversial articles. The figures reveal there is no correlation between the two variables for both controversial and non-controversial articles.

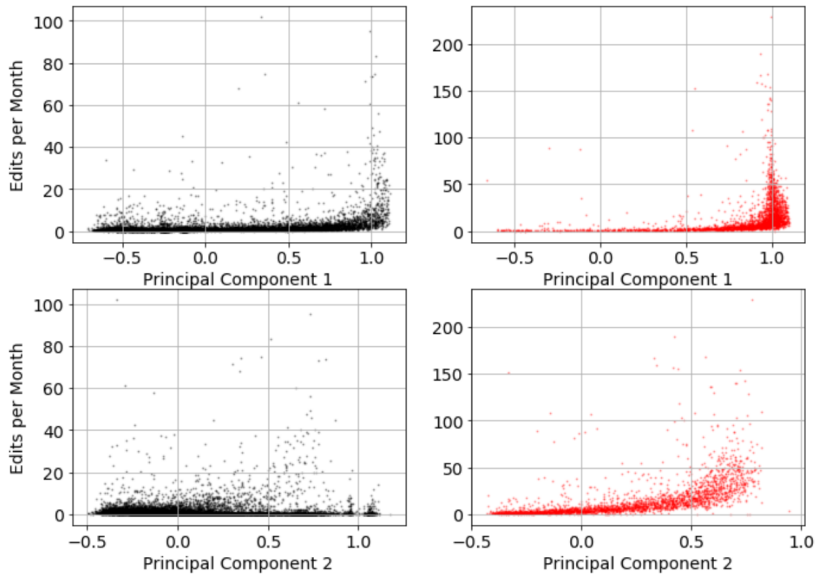

Figure 4.16: Scatter plot of each principal component combined with edit rate (mean number of edits per month) for both non-controversial and controversial articles. Much like, Figure 4.14, the figures reveal a correlation between the two variables for both controversial and non-controversial articles.

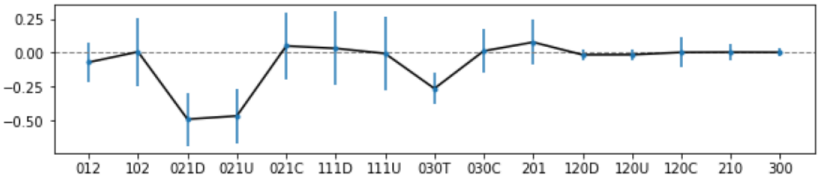

Figure 4.17: Using the original SRP's, the mean and standard deviation are taken for each triad (including dyads 012 and 102) for all non-controversial articles.

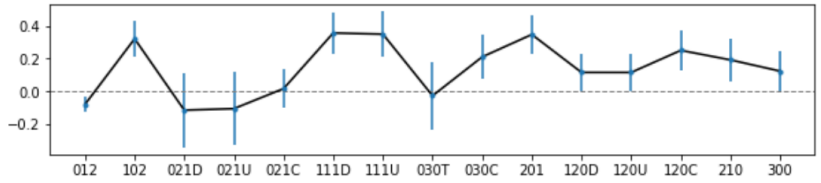

Figure 4.18: Using the original SRP's, the mean and standard deviation are taken for each triad (including dyads 012 and 102) for all controversial articles.

Reproducing the SRP to include dyads reveals the extent to which they describe the structure of the revision networks. Figure 4.17 indicates that both dyads have a mean value of 0 each with a relatively high standard deviation, with dyad `012` varying the least out of the two. Contrary to Figure 4.18, controversial profiles offer little variance with a high mean significance for dyad `102`.

Plotting each dyad as individual components in two-dimensional space reveals clustering behaviour similar to that of Figures 4.12 and 4.13. It is clear from Figure 4.19 that controversial content (marked in red) produces a distinct cluster, except for a few outliers.

Prediction

In view of the results provided from principal component analysis, these findings support the potential for clustering and classification which help address Research Question 4. As a result, a support vector machine (SVM) and binary logistic regression (BLR) are used in line with Research Question 4 to perform binary classification on the profile feature vectors for dyads, triads and both combined for all controversial and

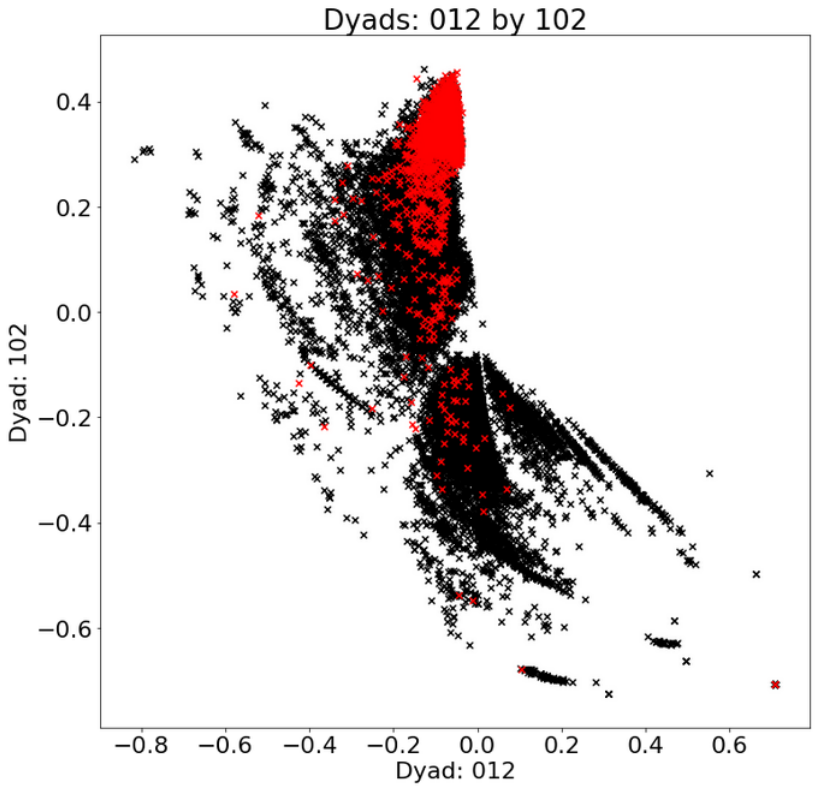

Figure 4.19: A simple two-dimensional scatter plot of dyads 012 and 102 significance values in isolation reveals similar clustering behaviour observed with triads.

non-controversial profiles. These results can be observed in Figures 4.20, 4.21 and 4.22 in the form of Receiver Operating Characteristic (ROC) curves.

Generally, each classifier performs relatively well with an accuracy of $P = 0.93735$ for BLR and $P = 0.93578$ for SVM using triads in isolation. The results presented in Figure 4.20 indicates that logistic regression has a consistent performance compared to a SVM where performance progresses slowly. Similar comments can be made about the performance of each classifier when both triads and dyads are used in combination (see Figure 4.22) with an accuracy of $P = 0.9374$ for BLR and $P = 0.9366$ for SVM. Using dyads, independent of triads, for prediction fails to perform as well as the previous two in Figures 4.20 and 4.22) resulting in an accuracy of $P = 0.88072$ for BLR and $P = 0.89555$ for SVM. A more detailed analysis of the binary classification performance can be observed in Table 4.5 for each classifier.

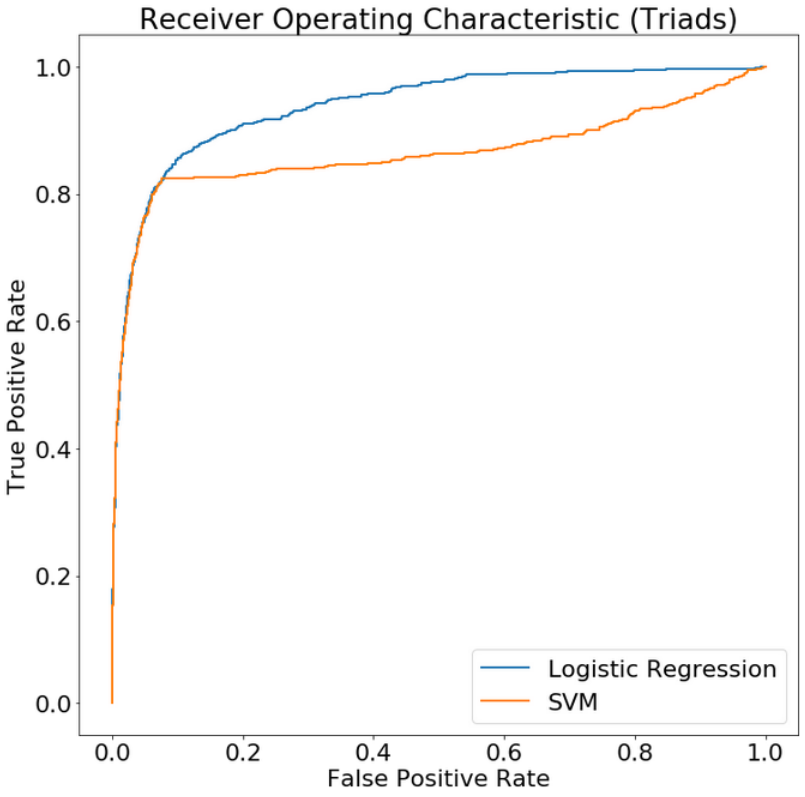

Figure 4.20: The ROC curve used to measure classification performance of each binary classifier using triads as feature vectors.

4.4.5 Discussion

Motif analysis of revision networks gives insight into how the temporal editing relationship between small groups of Wikipedia authors create signatures that allow controversial articles to be distinguished which helps answer Research Question 4. In contrast to previous work, this work is performed using a relatively large sample of Wikipedia articles, where distinct patterns emerge. This provides strong support for the hypothesis of this thesis and addresses Research Question 1 and Research Question 2 by demonstrating that the revision network can be used to capture behavioural signals on social media in the form of "switches" between article revisions based upon collaborative affordances such as "reverting" and "deleting". This also reaffirms the importance of the revision network as a simple but fundamental element in editing Wikipedia.

Through motif analysis, reciprocation on linear paths can be identified among triads of

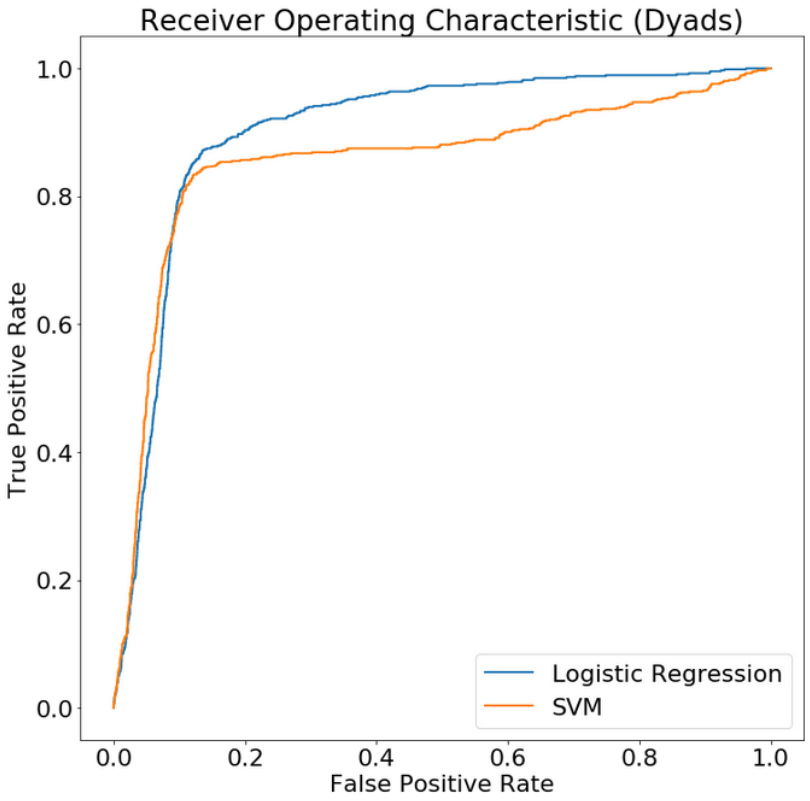

Figure 4.21: The ROC curve used to measure classification performance of each binary classifier using dyads as feature vectors.

editors in the revision network is over-represented in controversial Wikipedia articles. These motifs are defined by the triads `111D`, `111U` and `201`. In contrast, the revision networks from non-controversial articles exhibit two anti-motifs, involving the under-representation of triads involving two directed edges either arriving at or emanating from a mediating node (triads `021D` and `021U`). These motifs and the underlying subgraph ratio profiles represent an unusual and distinctive profile which represent distinctive "super-families" beyond those seen in other technologically related networks, such as the World Wide Web [418].

Performing dimensionality reduction upon the subgraph ratio profiles from each revision network provides an understanding of the relationships between Wikipedia articles. This analysis shows that the structure of the data is amenable to reduction to two dimensions, where the principal components are dominated by triads `111D` and `111U` in the first

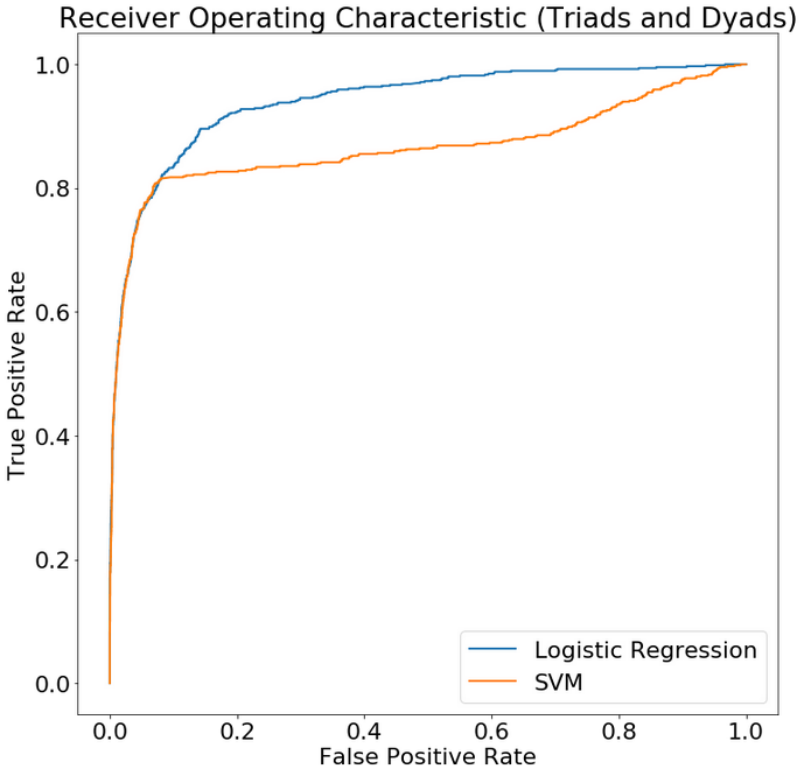

Figure 4.22: The ROC curve used to measure classification performance of each binary classifier using both triads and dyads as feature vectors.

component, and mainly `021C` in the second component, but with lesser contributions from triads `021D` and `021U`.

The results from two-dimensional principal component analysis show that the dominant triads in both components, as listed above, correspond to linear paths, i.e., open triads which represent sequences of editing without indirect reciprocity. The extent and format of reciprocated (i.e., bidirectional) edges on these open triads is sufficient to define the two principal components. The dominant triads in the first principal component each involve reciprocation on one edge, whereas interestingly, in the second principal component, the dominant triads are open triads with no reciprocated edges. From this it is suggested that short paths, rather than short loops of editing that represent indirect reciprocity, are important features in characterising Wikipedia revision networks.

In addition to this, using two-dimensional principal component analysis, the first prin-

			Precision	Recall	F1-Score	Support
Dyads	BLR	NC	0.9	0.97	0.93	4799
		C	0.48	0.23	0.31	663
	SVM	NC	0.92	0.95	0.94	4799
		C	0.56	0.41	0.47	663
Triads	BLR	NC	0.95	0.98	0.96	4752
		C	0.79	0.64	0.71	656
	SVM	NC	0.95	0.98	0.96	4752
		C	0.8	0.6	0.69	656
Both	BLR	NC	0.95	0.98	0.96	4799
		C	0.79	0.64	0.71	663
	SVM	NC	0.95	0.98	0.96	4799
		C	0.8	0.61	0.7	663

Table 4.5: Classification report for every possible combination of data type, classifier and label where "NC" is non-controversial and "C" is controversial.

cipal component strongly distinguishes between the revision networks of controversial and non-controversial articles. The dominant triads defining this capture the extent of direct reciprocation being present. Finally, through consideration of the principal components against additional external variables, it can be observed that article age plays a role in distinguishing the controversial articles. High values in both principal components aligns with strongest clustering of controversial articles, which is not the case for non-controversial articles.

With respect to dyadic analysis, thesis results provide additional support into understanding the clustering behaviour of controversial articles. Although it is clear that clustering can be performed by combining each dyad in a two-dimensional feature space, we can conclude that controversial articles do not vary greatly in dyad `012` ($A \longrightarrow B$) which supports the hypothesis that controversial articles do not typically follow an incremental structure as reverts and “undoing actions” are frequently made.

With respect to prediction, the data extracted can be used as feature vectors to solve many classification and prediction related problems. As demonstrated, the use of binary classifiers (in this case, BLR and SVM) can be used in article classification based upon the graph structure of a revision network using the methods provided. Performing

prediction from two dyads alone provides produces poor results compared to that of triadic analysis. This supports the belief that the structure of triads (as a combination of dyads) encodes more information than dyads in isolation.

4.4.6 Key Findings

This analysis has given insights into the structure underlying revision networks from Wikipedia articles, and has shown that a relatively small number of features, in terms of substructures in revision networks, characterise controversial Wikipedia articles. Using the collaborative data structure and transitional network, the results have identified key clusters of editorial interactions to this effect, in support of the hypothesis of this thesis. These are distinctive and indicate that the revision networks for controversial and non-controversial Wikipedia articles have differentiated subgraph ratio profiles. This investigation gives understanding as to how prediction or classification of articles can be enhanced using the latent structures relating to editor behaviour, which supports Research Question 4 of this thesis. This also reaffirms the importance of the revision network as a simple but useful representation for assessment of Wikipedia articles based on the transitional network. Furthermore, these results establish the reliability and efficacy of using transitional networks for modelling periodic behaviour, as shown in the investigation on Wikipedia which addresses Research Questions 1 and 2 as a result.

4.5 User-oriented Transitional Networks on Reddit

Compared to content-oriented transitional networks (see Section 4.4), user-oriented transitional networks are designed to focus on interactions that are centred around a single user according to the feed data structure. The activity they produce over time is analysed by observing how they transition between different pieces of content.

Social media platforms such as Reddit allow users to post and share content with

communities of other like-minded individuals. Reddit is segregated into smaller communities known as subreddits, which are dedicated to a particular topic or theme. A user can interact with these communities in the form of posting a submission (either sharing a link or publishing a text post) and commenting (commenting on another user's post or leaving a reply in response to another user's comment).

As a result, this mechanism opens up the possibility of bad actors seeking to cause disruption at scale by distributing content to a collection of different subreddit in an attempt to falsely influence the opinions of others and cause disruption. In recent years, Reddit has highlighted the issue of users with malicious intent as part of their 2017 Transparency report[5].

Much like Wikipedia, Reddit shares a similar structure by providing subreddits as a mechanism to allow users to congregate around a specific topic with other users, to discuss shared items. As seen within Wikipedia, (see Section 4.4) an article takes on the form of a piece of content where a collection of users seek to provide edits, forming a community around said content.

As a result, further modifications can be made to form transitional networks around users and the transitions that are made between Reddit's subreddits, the "content". In this case, the concept of a transitional network is applied in a user-oriented fashion.

We address the hypothesis of this thesis by investigating, firstly, alternative ways in which behavioural networks can be explored on social media using Wikipedia and Reddit (see Research Question 1). Secondly, by constructing alternative network-based representations using two different orientations (see Research Question 2). Thirdly, by capturing different types of affordances (see Research Question 3) associated with vandalism and disruptive accounts. And fourthly, by utilising methods derived from motif analysis to detect disruptive behaviour from controversial topics using the transitional network representation developed in this investigation (see Research Question 4). For

[5]2017 Transparency Report: `https://www.redditinc.com/policies/transparency-report`

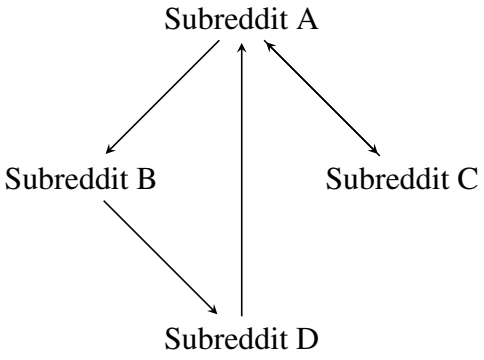

Figure 4.23: Much like Figure 4.6, a network generated using the subreddit order A, B, D, A, C, A, from the newest edit to oldest. An edge is formed between the current subreddit and its adjacent neighbour in the sequence, forming the directed edges $(A, B), (B, D), (D, A), (A, C), (C, A)$.

this reason, as addressed in Section 4.1, alternative formulation of transitional networks is therefore relevant.

In this investigation, subreddits are represented by nodes and a directed edge between $SubredditA$ and $SubredditB$ indicates that "*a user posted in Subreddit A then posted in B after*". An example network is shown in Figure 4.23

In view of the key findings presented in Section 4.4, the notion of transitional networks provides utility for understanding temporal switches between two users around specific content of interest (content orientated).

As mentioned previously, this addresses Research Question 2 which, combined with content-oriented transitional networks, can be used as part of a wider framework of different network representation. Furthermore, the relationship between a user and content has an important role in characterising user behaviour and helps answer Research Question 1. This discussed further in Chapter 6.

This, in turn, motivates additional research to consider the reversal roles of users and content such that transitions are made between different items of content around an individual user. This is said to be user-oriented. This section is intended to determine whether the approach used as part of the section on controversial articles on Wikipedia

(see Section 4.4) can be transferred to focus on users as opposed to content.

4.5.1 Related Work

As a platform, Reddit has experienced disruption in a number of ways. Perhaps one of the biggest forms of disruption that has taken place on Reddit is trolling. The use of trolling on Reddit has been observed across a large range of topics, from religion to entertainment [138]. In recent years, this has evolved to include more contentious topics such as health and politics [407]. This became particularly dominant during the 2016 US Presidential Election and the rise of Donald Trump [141, 164].

In addition to trolling, a considerable amount of research has been invested in solutions to identify and tackle misinformation [234, 329]. Much like trolling, misinformation on Reddit is centred around both health and politics. Research on Reddit has shown how echo-chambers have contributed to the development of anti-vaccination conspiracies and the spread of political misinformation [376, 76].

The concept of building network representations to capture switching/migrations patterns has been explored in multiple ways. Within the literature, studies on switching-like behaviour have been observed from the context of migration theory. Migration theory describes user behaviour using the push, pull, and moorings (PPM) model for understanding what motivates users to migrate (push), what attracts them (pull) and prevents them (mooring) [31, 58].

The PPM model was designed to understand migration in an offline environment, but has since been extended to understand migration on online social media platforms. Research has shown how this model has been used to understand what causes people to migrate in a cross-platform manner between different social networking apps and instant messaging applications [147, 90, 304].

Similar research has been applied specifically to Reddit in an attempt to understand how users migrate between subreddits over time. This has to be achieved by observing

migration on different scales between subreddits (micro and macro) [110] and from Reddit to other platforms such as Digg and Slashdot during a time of community unrest [277].

As a result, user migration on Reddit is a non-trivial matter and is therefore an important factor to consider when understanding how users seek to cause disruption between subreddits. Within the literature, few studies consider the role of a network-based approach for capturing predictive signals to identify the presence of disruptive behaviour. For this reason, transitional networks serve as an appropriate representation to consider when modelling a user's movements over time. As a result, this motivates the work of this section.

The purpose of this investigation is to demonstrate the utility of user-oriented transitional networks for detecting disruptive behaviour. In doing so, this investigation considers Research Question 1 of this thesis by defining a behaviour network composed of normal and disruptive users though transitional networks for modelling migration behaviour. Furthermore, this section also demonstrates how the transitional network could be used in a different orientation to model patterns of migration as well as collaborative exchanges, which helps answer Research Question 2.

Within this section, transitional networks are used in an attempt to extract small sequential patterns to discover common migration patterns between two or more communities. These sequential patterns are used to find sequential patterns which are unique to disruptive users over normal users and has the potential for capturing statistically significant interactions and could reveal potential for classification (see Research Question 4) and the detection of events such as posting campaigns.

4.5.2 Dataset

As discussed previously, this section seeks to explore Research Question 1 and Research Question 2 of the thesis by understanding the role of user-orientated transitional net-

works for detecting disruptive behaviour. Labelled data to indicate disruptive behaviour on Reddit is important as a basis for investigation. Accordingly, Reddit has published a publicly available list of accounts which they deem to be suspicious[6] as outlined in their 2017 Transparency Report[7]. This dataset is used as the ground truth for the "suspicious" users dataset. This returned a total of $N = 938$ users.

To form a baseline comparison, data from a random set of users is collected by iterative sampling from a random collection of subreddits. This group of randomly picked users are considered to be a fair representation of the entire Reddit population and to ensure that no two users are alike. Additionally, this will produce more diversity with respect to activity. This produced a total of $N = 896$ users. In doing so, the baseline comparison is used for drawing conclusions about the distinctions in behaviour between suspicious and random users.

The timestamp of a user's posting activity is a valuable feature for describing the direction of a switch between a pair of subreddits in the transitional network. As a result, Figures 4.24 and 4.25 reveal the events plots for both random and suspicious users according to the timestamp of a submission. Each event plot is between September 2011 to April 2018 where a cell has a window size of a week and indicates if a user was active within that period.

As observed in Figure 4.24, it is clear that user sessions are far more evenly distributed and longer-lasting. Random users are nearly constantly online, with very few breaks in between each period. The set of suspicious users in Figure 4.25 reveals that the timings are very much in alignment with other users, highlighting the presence of potential group co-ordination.

The results provided by the activity event plots (as shown in Figures 4.24 and 4.25 provide motivation for the overarching goal to differentiate between normal and disrupt-

[6]List of suspicious users on Reddit: `https://www.reddit.com/wiki/suspiciousaccounts`

[7]2017 Transparency Report: `https://www.redditinc.com/policies/transparency-report`

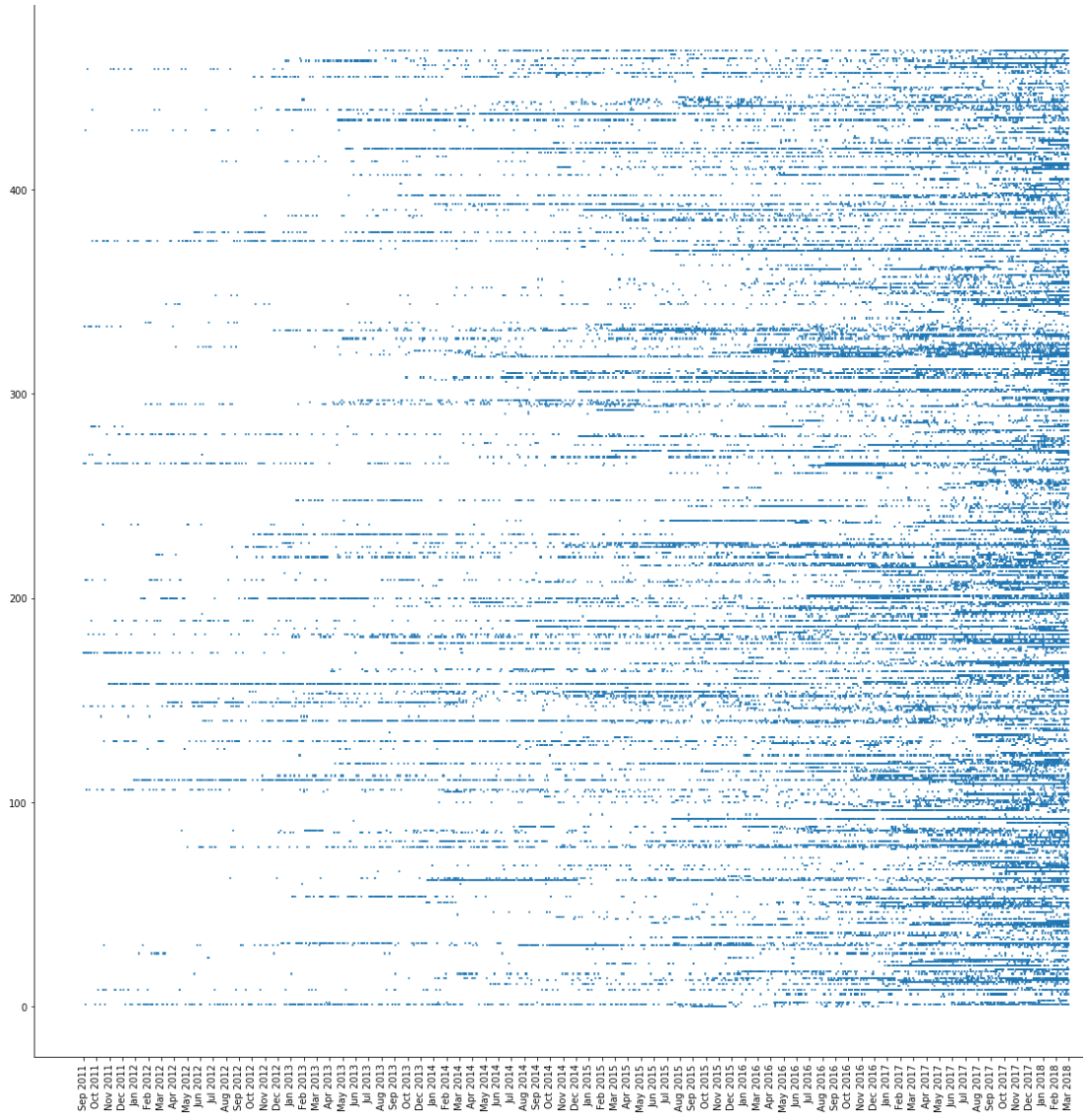

Figure 4.24: An event plot where a filled cell represents user activity within a window of a week. The figure reveals how random user activity shows near-constant usage with irregular periods between activity.

ive users based upon migration switching behaviour. It is evident that the timings of submissions contribute to the identification of suspicious users.

Overall, a total of $N = 1,135$ unique subreddits were used by suspicious users and $N = 12,565$ unique subreddits were used across all random users. Both suspicious and random users have a total of $N = 709$ subreddits in common.

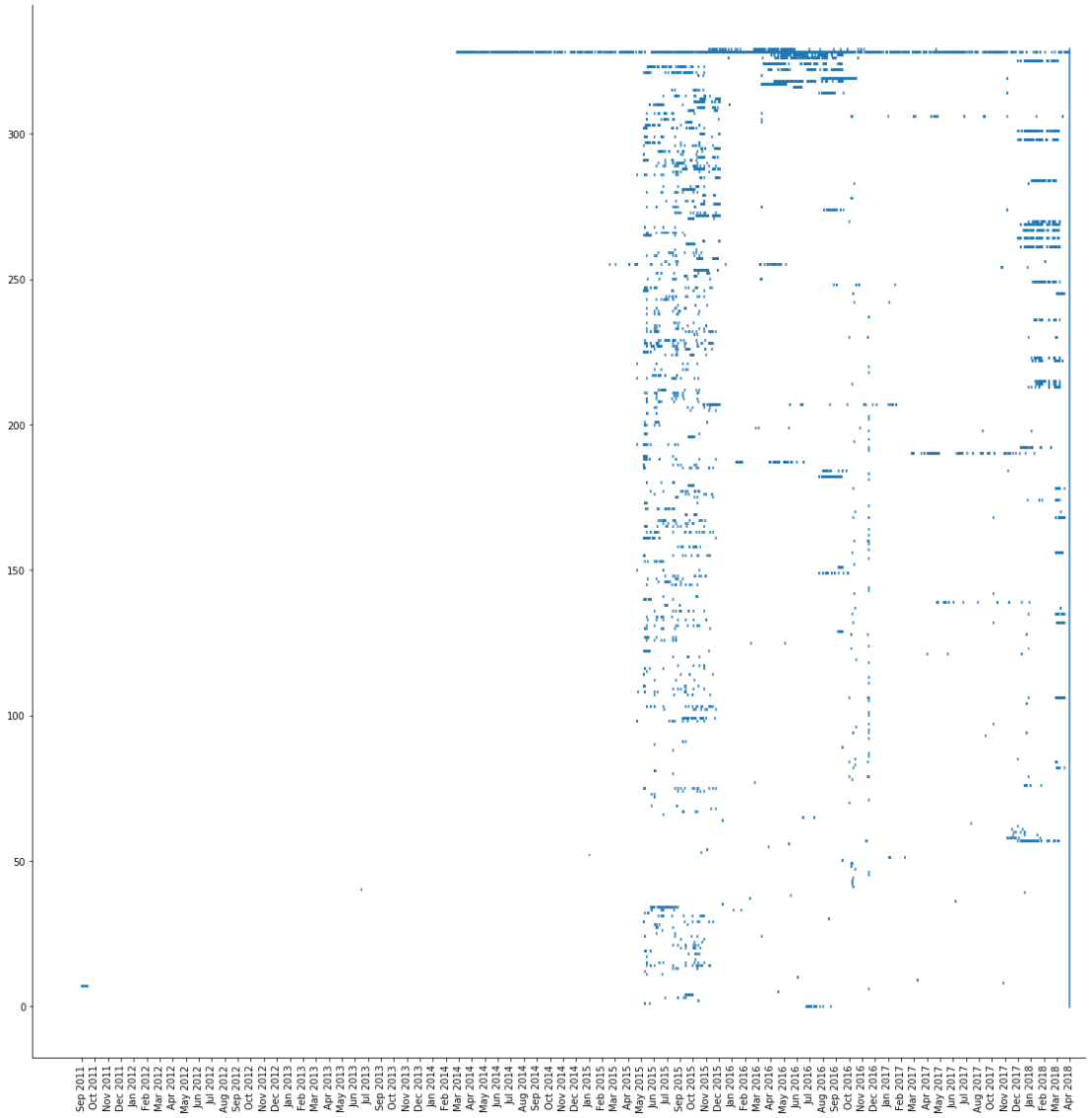

Figure 4.25: An event plot where a filled cell represents user activity within a window of a week. The figure reveals how suspicious users are active in distinct bursts which often overlap with other users in the set with clearly marked periods of inactivity.

4.5.3 Methods

User-Oriented Transitional Network Construction

Each user is represented by a user-oriented traditional network where nodes are used to represent unique subreddits and a directed edge shows the order in which an individual

user posted in a subreddit after another subreddit. The posting activity of an individual user's profile is traversed in the order in which submissions are created, spanning the entire posting history of said user's profile.

A network is formally defined by, $G = (V, E)$ where each subreddit is represented by a node $v \in V$. An edge $(v_i, v_j) \in E$ indicates that a user posts in subreddit v_i after posting in subreddit v_j. A simple example can of the resulting network can be found in Figure 4.23.

Consider the example provided in Figure 4.26 where the entire activity of subreddit switching is aggregated cross all users in the dataset of suspicious users. Although very little detail is provided considering the scale of the network, initial observations reveal frequently visited subreddits as indicated by node size.

Switch Motifs

Much like the previous investigation (see Section 4.4), to differentiate between suspicious and random users, motif analysis is used to discover statistically significant switch "patterns" which are unique to each user group. The distinction between the two orientations of transitional network is that rather than focusing on unique triad formations (Section 4.4), this analysis considers the role of simple switches in a chain-like linear fashion, which in turn explores Research Question 1 and Research Question 2 of the thesis. To achieve this, this section focuses on combinations of subreddits (in groups of two) which appear in a sequence similar to an n-gram model or Markov chain, in which transitional networks are used to model / count the unique switches.

This is motivated by the results in Section 4.4.4 of the previous investigation, where the SRP profile indicated that the simple linear, chain-like connections and dyads are more statistically significant than the remaining triad formations in the complete set. As a result, these linear formations provided the strongest signals and motivates the methodology of this investigation through the concept of "switch motifs".

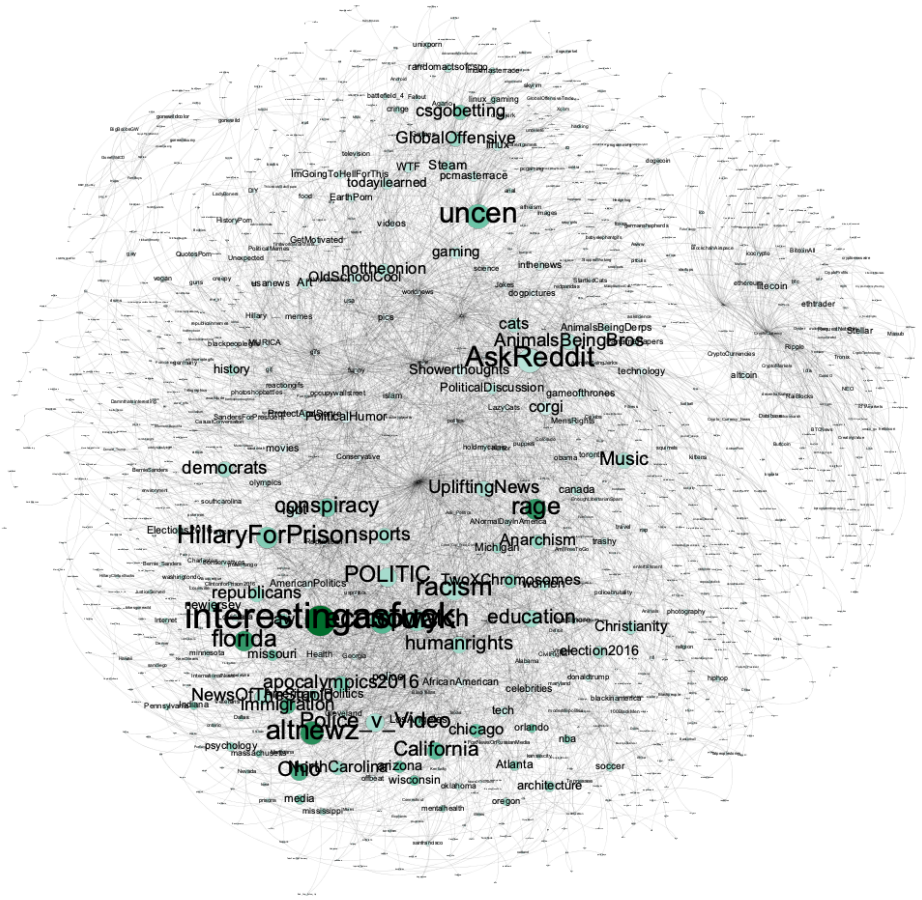

Figure 4.26: A switch between two subreddits is made when they appear together as neighbours in time-series, forming a directed network. The size of a node reflects the in-degree value and the shade of the node represents the strength of the eigenvector centrality.

This approach, as has been used in other studies whereby, the sequential patterns of purchases in credit card data can be used to reveal different types of lifestyles in urban populations [117]. Furthermore, similar techniques are widely used within bio-informatics as a method for discovering gene sequences [28].

One of the leading examples in gene sequencing algorithms is MEME (Multiple Em for Motif Elicitation) which uses a probabilistic mixing model as a null model. The primary purpose of MEME is for processing DNA, RNA and protein sequences with additional features such as the addition of a custom “alphabet” [28]. This approach is

capable of discovering statistically significant sequences in the same way network motif analysis can find statistically significant subgraphs. As a result, this approach motivates the methodical structures of this investigation, whereby a similar approach is used to find significant subreddit switches in the same way as significant gene sequences.

Much like traditional motif analysis, random sequences of subreddits can be generated to form a baseline comparison (otherwise known as a null model) combined with the relevant statistical measures used throughout the motif literature. Within gene sequencing algorithms, the notion of an "alphabet" is introduced and maps directly to the total set of subreddits used in the investigation. The algorithm for the solution is approximated as follows:

1. Define an alphabet of subreddits $S = \langle S_1, S_2, \ldots, S_n \rangle$
2. Generate all possible sequences where $n = 3$ for sequence length
3. Count occurrences for all possible sequences in the target network
4. Generate random sequences using the same alphabet (usually 100 switch sequences) from a uniform distribution
5. Determine Z_i or Δ_i scores for each distinct combination (see Equations 3.3 and 3.4).

4.5.4 Results

Figures 4.27 and 4.28 provide a overview of the top 40 popular subreddits (according to in-degree) across both user sets. Initial observations from random users (Figure 4.27) show a significant increase of activity within the subreddit r/AskReddit compared to others.

It's possible to suggest that this may be a sign of a small-world network, as there is a possibility that any pair of users in the random set may have engaged with each other.

Further analysis is needed to better understand activity on a local level to reaffirm this observation.

Figure 4.27 also shows random users are far more likely to engage in broad subreddits contrary to suspicious users. As observed in Figure 4.28. These subreddits can be broken down into three main themes: *Politically divisive* communities (e.g. r/The_donald, r/HillaryForPrison); *activism* (e.g. r/Bad_Cop_No_Donut, r/Police_V_Video and r/uncen) and *cryptocurrencies* (r/Cryptocurrencies, r/Blockchain and r/Bitcoin).

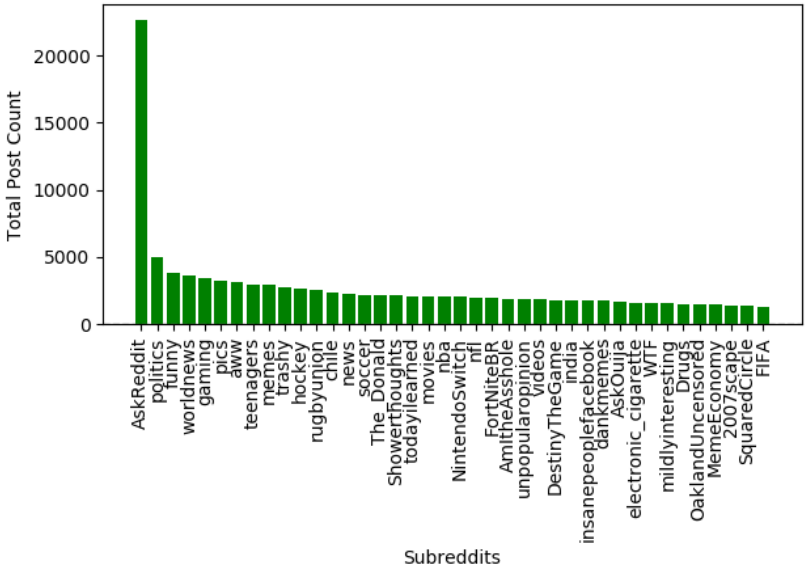

Figure 4.27: The set of random non-suspicious users post the vast majority of content in the r/AskReddit subreddit. Other top subreddits include r/politics, r/funny and r/worldnews which are fairly generic with respect to discussion.

Without considering duplicate subreddits in an individual switch, a high-level overview of basic switching patterns reveal somewhat distinct behaviour between the two groups. Random users exhibit somewhat consistent behaviour by producing switches between two subreddits that are of a similar theme. The top four most popular switches according to r/moderatepolitics $\rightarrow$ r/politics, r/ukpolitics $\rightarrow$ r/unitedkingdom, r/cryptocurrency $\rightarrow$ r/nanocurrency and r/army $\rightarrow$ r/airforce. These examples show that user sessions typically fall within the same topic category.

Contrary to random users, suspicious users produce switches that are more unusual

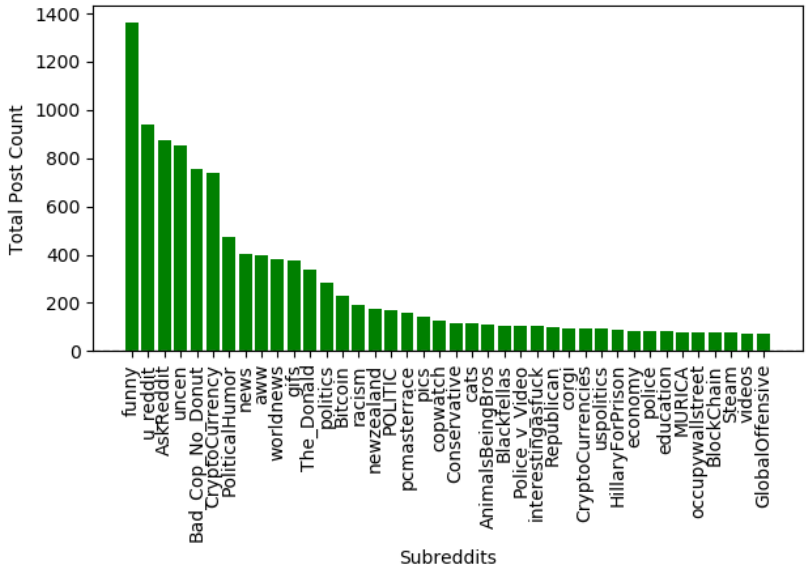

Figure 4.28: The suspicious user set is somewhat similar to that of normal users however, a few of the top ranking subreddits include specialised communities such as r/uncen and r/Bad_Cop_No_Donut.

and inconsistent compared to that of "normal" behaviour. For example, suspicious users frequently engage with political and cryptocurrency related subreddits such as r/cryptocurrency, r/blockchain, r/bitcoin, r/politics and r/conservative.

Due to the large quantity of subreddits (total of $N = 12,991$) it was only possible to count distinct switches with a maximum length of $N = 2$ that occurred in the dataset. In the interest of speed and efficiency, this analysis is concentrated on the union of subreddits between the two sets, focusing on the communities that both random and suspicious users have in common. After this reduction, it can be observed that both user groups have a total of $N = 709$ subreddits in common.

The total number of switch combinations of length $N = 2$ produces $709^2 = 502,681$ possible combinations to enumerate over. For this reason, it is simply not possible, with conventional resources, to perform motif analysis as expected due to the computational overhead of generating a minimum of $N = 100$ additional random permutations as a baseline. This issue elaborated further in Section 4.5.5.

4.5.5 Discussion

The methods developed in this investigation attempt to determine the feasibility of differentiating between random and suspicious users according to migration patterns through the concept of switching motifs as a method for discovering statistically significant switches. Ultimately, this did not support the hypothesis that a user-oriented transitional network can be used to model migration on Reddit. The analysis proved that this was not the case, as motif analysis can not be used to differentiate the activity of normal and disruptive users.

While motif analysis on switching networks using user-oriented networks failed to yield insights, this was not the case for content-oriented transitional networks (see Section 4.4) where it can be observed that motif analysis provides utility for applications such as prediction.

The results from this analysis allude to a number of key issues and possible solutions. These are outlined as follows:

High Volume of Subreddits

As mentioned previously, a total of $N = 12,991$ unique subreddits overall, meaning that the possibility of any pair of users sharing the same switch sequence is incredibly low, making it extremely difficult to compare different users. Consequently, comparing the results may distort the analysis as this will likely feature switch pairs which may not be present within the other user group (e.g. random and suspicious).

Furthermore, data collection was performed on a content-centric basis (similar to Section 4.4 by sampling via subreddits) however, the analysis was performed on a user-centric basis meaning that the two processes (data collection and analysis) were misaligned.

One solution to combat this issue would be to categorise subreddits to emphasise switching between similar topics (e.g. Hobby, Politics, Sport). Another option would be

to maximise the overlap of possible switches by taking the intersection of the subreddits between the two user groups.

Possibility of Duplicates

Duplicated switches are frequently occurring switch sequences which contains the same subreddit in the entire sequence (e.g. r/AskReddit $\rightarrow$ r/AskReddit). Duplicates mostly explain high bursts of activity within the same subreddit, which is unhelpful when finding distinct switching behaviours. Observing switches within the same subreddit across all users enforces the idea that users are persistent in their engagement within the same community. While it is possible for a user to favour posting in a certain subreddit over others, less frequent observations may be overlooked.

Scalability

Given the number of subreddits involved in the set, this solution doesn't scale well with respect to the number of comparison that are needed for matching across all subreddits. Finding occurrences of pairs will scale in quadratic time $O(n^2)$ and triples in cubic time $O(n^3)$. For example, using the $N = 709$ shared between the two user groups, the algorithm will compare $N = 502,681$ pairs and $N = 356,400,829$ triples. Even with a reduction, the solution does not scale well and is not suitable for application which perform in real time.

Considering the size of the sample space, with such a large set of subreddits, this will mean that speed is significantly reduced as services like MEME operate using a far smaller alphabet. While this solution provides a quick alternative for discovering over-represented switches, it is subject to criticism as it is not adapted for true sequential motif discovery.

4.5.6 Key Findings

As mentioned in Section 4.5.5, the results produced from this investigation reveal how the use motif analysis is unable to identify statistically significant motifs to differentiate between random and suspicious users using a user-oriented transitional network of subreddit switches. Consequently, this means we are unable to answer Research Question 4 as a result of not being able to perform any form of classification using the techniques developed in this investigation on user-oriented transitional networks.

Each user account featured in the analysis has a broad range of subreddits which are dissimilar and diverse by comparison to other user's subreddits. Even though a user may post in a diverse range of subreddits, ultimately not enough data is provided for each account to provide a detailed overview of user activity, making it harder to extract switching signals that are unique to a particular user or group.

In addition to this, the results from the switch motif analysis revealed several issues relating to time complexity and sample space of enumerating over every possible switching pairs within the set. While switch motif analysis may provide promising results, future research that is based on the optimisation of motif analysis and switching networks is needed and goes beyond the scope of this thesis. Instead, this investigation addresses these issues and offers suggestions for future research.

4.6 Conclusions

To conclude, this chapter provides two investigations which have been invoked to answer the hypothesis and research questions of this thesis by emphasising the utility of transitional networks on two unique yet valuable platforms for mining collective user behaviour. Both Wikipedia and Reddit, although different, can model user behaviour in similar ways using transitional networks using the collaborative and feed data structures respectively (as introduced in Chapter 3). This chapter demonstrated that it is only

possible to observe user activity from one perspective using a content-oriented transitional network. As a result, this solution provides versatile methods for understanding user activity by observing latent substructures through subgraphs of content-oriented transitional networks, while further research of user-oriented networks is needed.

In the case of Wikipedia, transitional networks provide a mechanism for interpreting editorial exchanges of content among a collection of editors. Nodes are used to represent editors, with directed edges being used to link to the previous editor's contributions. This in turn can be inverted to observe how a single editor can transition over multiple articles.

On the other hand, user-oriented networks on Reddit reveals that the methodology does not easily translate to other platforms using a different orientation. Data for these networks is generally content centric, and user centricity is much harder to acquire and analyse, noting that the large pool of required users is a significant impediment to analysis. As a result, a lot of data is needed to build an accurate representation of the activity due to limited quality data. This is further compounded by the issue that the algorithm for detecting unique switches does not scale well with respect to size. For this reason, it is possible to conclude that Reddit, and indeed other community-oriented platforms, the method can not be transferred to user-centric analysis in the same way as demonstrated on Wikipedia for content-centric application.

In conclusion, only content-oriented transitional networks support the hypothesis through Research Questions 1, 2 and 4, yet this is not the case for user-oriented transitional networks due to being unable to construct and extract behavioural signals which can be used to detect the presence of disruptive activity.

Chapter 5

User-To-User Networks

5.1 Introduction

In the previous chapter (see Chapter 4), the concept of transitional networks were introduced in an attempt to model switches (also known as "transitions") between two or more different states. These were examined from two perspectives, content-oriented and user-orientated. The results revealed how content-oriented networks on Wikipedia produced promising results by focusing on transitional networks composed exclusively of pairs of users collaborating on a Wikipedia article. This led to the success of being able to detect controversial from non-controversial articles using just the network representation as a feature vector. As a result, these findings demonstrate that behavioural networks can be defined in the form of content-oriented transitional networks (see Research Question 1) and also help contribute to the development of a concise framework (see Research Question 2) and classification (see Research Question 4).

In view of the results from Chapter 4, the purpose of this chapter is to investigate the hypothesis of this thesis by addressing Research Questions 1 and 2 by maximising the involvement of user-led interactions. This is achieved by exploring user-to-user networks which represent simple user-directed interactions which take place between other users. Furthermore, this chapter seeks to investigate whether user-to-user networks can be used as part of a wider framework for examining diverse affordances (see

Research Question 3) and the extent to which they can be used to predict the presence of disruptive behaviour (see Research Question 4)

These networks are created with the intention of discovering potential signals for disruptive activity such as trolling (see Chapter 2, Section 2.2.2). This network representation is introduced in this chapter as a *user-to-user network*, which primarily focuses on the role of message-based interactions which takes place between a set of users. This chapter refers to "message-based" as interactions which contain a body of text directed at a specific user. The research conducted in this chapter closely examines user-to-user networks, using Twitter (as shown in Section 5.4) and Reddit (as shown in Section 5.5) to examine this network representation.

User-to-user networks are perhaps the most widely studied form of network represented within the literature, and are employed for a variety of reasons. For example, similar configurations have been applied to explore friendships / following, favourites and sharing [213, 275, 237]. As a result, the notion of constructing networks excessively from user interactions can be used to cover a broad range of scenarios for modelling complex social networks. However, in the context of detecting disruptive behaviour, the rationale behind user-to-user networks in this thesis is to study interactions within a message-based environment to consider latent signalling concerning disruptive behaviour.

While user-to-user networks could be used to model interactions which don't explicitly involve content (e.g a user liking another user's content) disruptive activity typically takes place over text as acts such as spreading misinformation and trolling are manifested through written content [355, 169, 133, 98]. While text-based analysis, has some utility in detecting disruptive behaviour [169, 133], consequently this produces a few defects. These include factoring in language, inconsistencies and spelling. Contrary to this, user-to-user networks are language-agnostic making it possible to expand their usage across multiple languages and regions. Furthermore, this network representation makes it possible to observe patterns of exchanges between users. This can be exploited to

provide a better understanding of how users behave on an individual (egocentric) and collective basis.

As outlined in Chapter 3, both Twitter and Reddit align with the message data structure. These are presented in Table 5.1 relative to the remaining work of this thesis through the use of the message data structure.

		Data Structure			
		Community	**Message**	Collaborative	Feed
Platform	Wikipedia	N/A	N/A	*See 3.4.1*	N/A
	Reddit	*See 3.4.2*	***See 3.4.2***	N/A	*See 3.4.2*
	Twitter	N/A	***See 3.4.3***	N/A	*See 3.4.3*

Table 5.1: Relationship between platforms of interest and all data structures with the appropriate cells concerning the work of Chapter 5 highlighted in bold.

The use of the message data structure aligns with the use of the user-to-user network representation (see Chapter 3, Section 3.6) which, in turn, informs the analysis of this chapter. As a result, this helps address Research Question 1 by defining behavioural networks from social media and Research Question 3 by representing diverse affordances.

As described in Chapter 3, Section 3.3.3, Twitter (as used in Section 5.4) is a widely used microblogging service used by many. Although, its primary purpose is to allow users to post 280 character-long pieces of text (otherwise known as tweets), what makes Twitter a particularly interesting platform to study is how users use the platform to communicate with others using three distinct types of message-based interactions.

As shown in Section 3.4.3 of Chapter 3, Twitter's message-based interactions take place between a *User* (the source) and a *Recipient* (the target) include interactions in the form of **mentions** (a tweet which contain "*mentions*" another person's username), **replies** (a user responding to another user's tweet) and **quote retweets** (tweeting another person's tweet with a comment) and these are distinguished using the *Interaction* attribute. As a result, these pairwise user interactions support different modes of communication and user engagement and an example can be found in Figure 5.1.

Figure 5.1: An example of a tweet demonstrating how quote retweets, mentions and replies can be represented by a user-to-user network using Twitter. This example reveals how a user-to-user network can be used to represent multiple directed interactions between a pair of users.

Similarly, as mentioned in Chapter 3, Section 3.3.2, Reddit (as used in Section 5.5), is described as a social news aggregation and discussion site where users can submit posts to individual communities known as subreddits. As shown in Section 3.4.2, when a user submits a post, this creates a new discussion thread where users (as represented using the *User* field) can interact directly with others (the *Parent* field) in the form of replies (the *Interaction* field) using the comments section below the post. These discussion threads are oriented around a specific topic (the post) within a certain theme (a subreddit). An example of a discussion thread on Reddit can be found in Figure 5.2.

By comparison, both Twitter and Reddit share the reply-based interaction where a user can respond directly to another user's piece of content using the message data structure. The distinction between the two platforms is that Twitter facilitates open conversations where anyone can partake in the discussion whereas Reddit is more community-oriented such that discussions are centred around a specific theme or topic. Furthermore, Twitter provides an additional two interactions in the form of quote retweets and mentions which, in turn, describes different semantics connections and can be used to improve

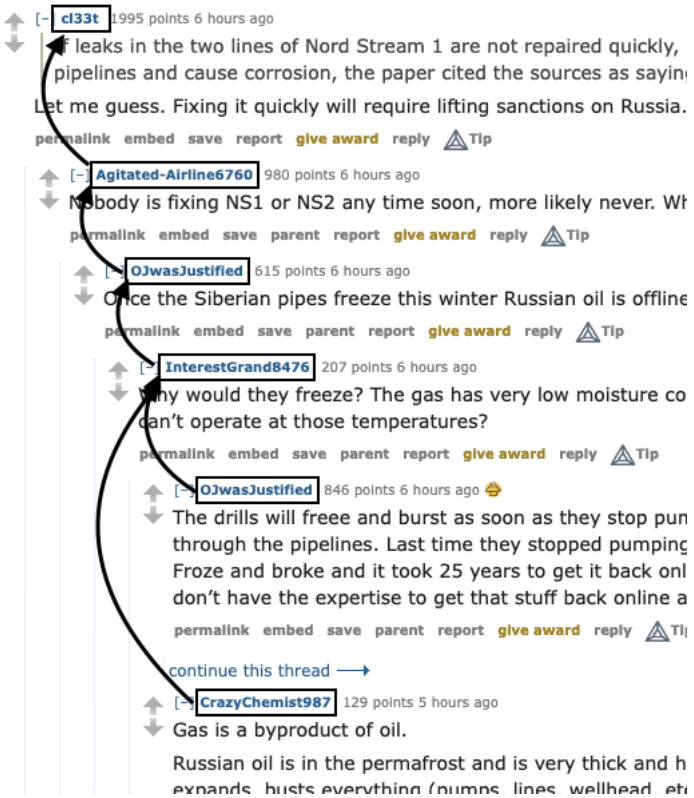

Figure 5.2: Example of a discussion thread on Reddit demonstrating how reply-based interactions can be reproduced using a user-to-user network based upon a nested conversation.

our understanding of how users behave and interact with each other.

5.1.1 Contributions

This chapter seeks to address Research Question 2 by demonstrating how user-to-user networks can be used as part of a wider framework of network representations to identify the presence of disruptive behaviour within an informal conversational "message-based" setting. This is achieved by considering platforms which can be observed through the message data structure (see Chapter 3, Section 3.2.3).

This chapter considers different types of user-to-user interactions to assess their utility for differentiating between topics that are likely and unlikely to cause disruption. In doing so, this offers new insights regarding by considering multiple interactions in an

attempt to survey as much activity as possible (see Research Question 1) and can to capture diverse affordances, as a result (see Research Question 3). The implications of this chapter demonstrate the versatility of user-to-user networks by using Twitter and Reddit as examples by detecting disruptive behaviour (see Research Question 4) and signals how this could be modelled on similar social media platforms (such as other microblogs and discussion forums).

5.1.2 User-To-User Network Construction

As defined in Chapter 3, Section 3.5.2, user-to-user networks are composed exclusively of users which are represented by nodes within the network and are formed according to the message data structure. In this case, using Twitter as an example, whether a user is featured in the network is determined by one of the following rules at a specific point in time:

- A user created content (e.g. publish a tweet, create a comment), the *author*
- A user engaged with another user's content (e.g. leave a reply), the *engagement*.

For example, with a behavioural network on Twitter (see Research Question 1), a user (User A) can compose a tweet which mentions another user (User B) to express their approval or disapproval in the form of a simple message. In doing so, this forms a directed connection between users A and B such that A *mentions* B. Likewise, on Reddit a user (User A) can reply to another user (User B) in a discussion thread which forms a directed connection between users A and B such that A *replies to* B. Both of these interactions are represented by the connection $A \rightarrow B$.

Networks are defined by $G = V, E$, where a node $V = v_1, v_2, \ldots, v_N$ represents users, such that a directed edge $(v_i, v_j) \in E$ produces the interaction "v_i *engages with* v_j". Furthermore, this can be expanded to capture multiple instances of an interaction

between pairs using metadata such as time t_i and frequency f_i such that $(v_i, v_j, t_i, f_i) \in E$ for greater accuracy and reparation.

5.2 Motivation

The analysis used as part of this chapter to investigate the hypothesis makes use of real world instances where disruptive behaviour takes place. This chapter attempts to explore the hypothesis by investigating user behaviour from a collective (see Section 5.4) and individual (see Section 5.5) perspective. In doing so, as discussed previously, this addresses Research Question 2 by examining different types of interactions / perspectives as part of a larger framework of network-based representations, Research Question 3 by representing a diverse range of affordances and interactions and Research Question 4 by using classification techniques to detect the presence of disruptive behaviour using networks features derived from the user-to-user network representation.

Both Twitter and Reddit have been used for causing disrupting typically in the form of spreading misinformation and trolling [183, 17, 85, 261, 138] where non-network-based methods (such as NLP) are used. For this reason, these two platforms motivate the research questions pursued by this thesis which, in turn, further demonstrates the versatility of user-to-user networks as one respiration for modelling many different types of interaction.

The use of social network analysis on Twitter, more broadly, has very important implications to this investigation for considering the utility of Twitter's three types of message-based interactions - retweets, mentions and replies. To begin, the role of retweet networks have been used to identify communities of like-minded individuals [91] and understand the spread of information [236] which, consequently, also includes misinformation [404]. Retweet networks provide predictive signals for predicting retweeting behaviour [402] and can be used as a reliable proxy for gauging popularity, social capital and friendship formation [317, 9].

In addition to this, the use of mention interactions are primarily used to direct a tweet to a user (or group of users) by mentioning their username in their tweet. This type of interaction has been observed in multiple settings demonstrating that mentions provide utility for predicting links between users [55, 185, 324] and shares similar structural properties to networks produced using the "favourite" button [213]. Finally, replies can also be treated as a network representation for modelling conversational dynamics. Reply networks have been used to study patterns in multiple Q&A discussions [327], monitor an audience's approval or disapproval [270] and to gain more followers [325] in a language-agnostic capacity [71]. Research has demonstrated how replies can be used to validate Dunbar number revealing that much like offline interactions, online interactions can reach a limit to the number of possible interactions that can be preserved [154]. Furthermore, additional research has highlighted how reply networks provide predictive qualities using local based features stating the need to go beyond simple structures based upon triangles (triads) [326].

With respect to Reddit, this platform provides a mechanism for allowing users to express both positive or negative sentiment towards the comments and replies of other users though voting. As discussed previously, users within a community can up-vote or down-vote a post depending on how well it is received within the community, in turn producing what is known as *karma*. Furthermore, users can produce comments and leave replies forming a nested discussion tree. As a result, Reddit serves as a rich platform to support direct user-to-user interaction through discussion threads. However, this functionality means that users can be exposed to sophisticated echo chambers [164, 57] and filter bubbles [70] as well as more serious problematic behaviour such as *trolling*.

Consequently, both of the two platforms are the recipient of disruptive activity and the supporting literature signals that user-to-user networks serve as a potential candidate in an attempt to detect disruptive activity. In doing so, various network statistics such as node centrality measures (e.g. popularity and influence [410]), induced triad counts and

reciprocity can be used to infer high-level social behaviours [293] thus strengthening the importance of this work and further enhancing our understanding of user behaviour on the context of disruption.

5.3 Approach

		Data Structure			
		Community	**Message**	Collaborative	Feed
Network	Transitional	-	-	See 4.4 (Wikipedia)	See 4.5 (Reddit)
	User-to-user	-	**See 5.4, 5.5 (Twitter, Reddit)**	-	-
	User Association	See 6.4 (Reddit)	-	-	-

Table 5.2: A replica of Table 3.20 featured in Chapter 3 outlining the investigations of this thesis (with respect to data structures and network representations) with the appropriate cells highlighted in bold which refers to the problem space which Chapter 5 seeks to investigate.

As shown in 5.2, relative to this thesis, this chapter attempts to understand the utility of user-to-user network representations for detecting disruptive activity using platforms which align with the message data structure. In doing so, this chapter is composed of two parts and is considered as follows. Firstly, by performing a comprehensive analysis of multiple message-based interactions. Secondly, by considering smaller egocentric networks (centred around a single user) discussions combined with temporal features.

As described in Chapter 3 Section 3.5.2, user-to-user networks and the message data structure are used to capture interactions (such as replies) in a content-driven manner (e.g. a tweet). Furthermore, Section 3.4 describe how Twitter and Reddit were selected due being aligned with the message data structure. As a result, this is used to demonstrate the versatility of the user-to-user network representation in a cross-platform manner which, in turn, can be used to define behavioural networks (see Research Question 1) and can be used as part of a coherent framework (see Research Question 2), as

previously mentioned. Both of these approaches involve using network structures as a feature vectors for predicting between disruptive and non-disruptive activity and address Research Question 4 as a result. This approach is tested by investigating the following:

- **Section 5.4: Twitter Message-based Interaction Networks**: Uses a combination of quote retweet, mention and reply networks to differentiate controversial "anti-vax" content from non-controversial content on Twitter using both global and local-based network metrics.
- **Section 5.5: Egocentric Reply Networks and Temporal Features**: Focuses on the reply behaviour of a user on an individual basis based upon their "egocentric" immediate neighbourhood supplemented with temporal information to flag disruptive users against normal users.

In this chapter, the investigation of the hypothesis is explored using Twitter and Reddit which both share the same overarching network representation (see Table 5.2) however, each of the two studies utilises these in two different ways. The first investigation uses Twitter to consider multiple static networks (in the form of quote retweets, mentions and replies) using the entire network in an attempt to find the most suitable representation and discover underlying structural properties suitable for classifying controversial from non-controversial networks. An example of this can be observed in Figure 5.3 and Table 5.3 for edge types.

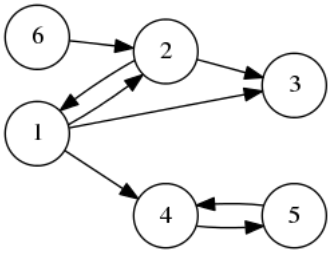

Figure 5.3: Example of a basic user-to-user interactions in the form of a reply network where a directed edge represents the direction of the conversation such that $A \rightarrow B$ indicates A *replies to* B.

Application	Quote Retweet Network	Mention Network	Reply Network
Source Node	User	User	User
Edge (directed)	*quote retweets from*	*mentions*	*replies to*
Target Node	User	User	User

Table 5.3: Edge list definitions for each of the network representation used as part of the exploration of the hypothesis using Twitter (see Section 5.4).

On the other hand, as shown in Table 5.2, the second investigation explores the hypothesis using Reddit to consider just one form of message-based interaction, replies. Similar to Twitter, Reddit also aligns with the message data structure and meaning that replies can be modelled through a user-to-user network.

By comparison, the size of the networks used in this analysis are reduced to focus exclusively on individual egocentric connections - interactions that take place around a single node. The results of the investigation on Twitter (see Section 5.4) provide strong evidence that the subgraphs which closely resembling egocentric networks provide the most predictive utility when considering "local" subgraph features for reply networks. For this reason, the methodology of the Reddit investigation (see Section 5.5) only considers egocentric reply networks in an attempt to simplify the process and to maximise classification performance.

In addition to this, temporal features (such as mean duration, account age and comment rate) are included to aid classification accuracy. In doing so, this makes it possible to better understand conversational dynamics which, in contrast to the first investigation, it's harder to understand within a static context. An example of this can be found in Figure 5.4 and Table 5.4 for edge types.

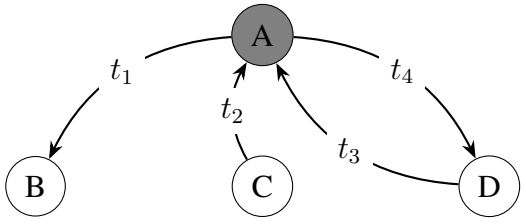

Figure 5.4: Example of a temporal egocentric network focused around a target user (gray) with edges occurring at timestamps t_1, t_2, t_3 and t_4.

Application	**Ego Comment Networks**
Source Node	Central User
Edge (directed)	*reply to/from* ($\leftrightarrow$)
Target Node	Other User

Table 5.4: Edge list definitions for each of the network representation used to investigate the hypothesis using Reddit (see Section 5.5).

5.4 Twitter Message-Based Interaction Networks

As discussed in Chapter 2, social networks have a fundamental role in the way in which users communicate with one another online. As a result of this, issues such as misinformation begin to emerge due to the size and heavily connected nature of social media platforms [404, 384]. More specifically, since the COVID-19 pandemic, discussions surrounding vaccine usage has attracted highly emotive positive and negatives view-points which, consequently, increases the potential for misinformation to emerge [190, 272, 209, 125].

This can have far-reaching consequences in an offline setting, as ill-informed decisions as of a result of misinformation can be a threat to public health [374, 378, 384]. There have been known issues where microblogging platforms such as Twitter, have been used to spread misinformation surrounding vaccines thought the use of message-led interactions (e.g. replies, retweets etc) [337, 386, 197].

The three interaction mentioned above (mention, reply and quote retweet) are the focus of this investigation and can be modelled using network-based methods to understand user-to-user interactions. Furthermore, little research has been performed in an attempt to study the utility of these three types of interaction - both combined and in isolation - by cross comparison. This investigation contributes new knowledge to this domain by using COVID-19 vaccines (see Section 2.1.4) as the basis for representing disruptive activity and to extract social networks from tweets using a collection of relevant hashtags and keywords.

This investigation of the hypothesis focuses on a combination of known, crowdsourced

and custom COVID-19 anti-vax hashtags and keywords (referred to as "terms" throughout the investigation) used on Twitter which are ranked by a small group of participants. A term is labelled either "controversial" or "non-controversial" depending on how it scores on a Likert scale. Each term is represented by three individual networks constructed for each of the three interaction types of interest. These are then used as part of a classification task. This methodology is used to support the investigation into the hypothesis (see Section 1.2) by detecting anomalous activity related to disruption (see Research Question 4) and is later used as part of an investigation into the role of capturing diverse affordances on social media (see Research Question 2).

The following hypothesis is set: *network metrics and substructures can be used to differentiate between controversial and non-controversial terms relating to COVID-19 vaccines using either or a combination of networks constructed from quote retweets, mentions, and replies*. In doing so, this investigation attempts to identify social networks which are constructed from a given term and interaction type which, as a result, are language-agnostic and independent of content.

The aim of this investigation is two-fold. Firstly, to understand how network-based features can be used to observe nuances between each network term and interaction type (see Research Question 2). For example, are you more likely to discover reciprocated ties in replies than quote retweet interactions? And secondly, to find signals between each set which are likely to lead to disruption (see Section 1.2 and Research Questions 1 and 4). Are there a subset of features which can be used to differentiate between controversial and non-controversial terms.

5.4.1 Background and Related Work

The ability to detect misinformation and fake news on social media is by no means a novel idea and there has been a wealth of research invested in computational solutions that can identify and detect user accounts responsible for spreading such information in

a semi autonomous fashion [336, 356, 115]. As discussed in Chapter 2, Section 2.1.4, misinformation surrounding vaccine usage has been a historical issue which has only been more exacerbated in recent years due to the COVID-19 pandemic [144].

Within the literature, a significant component for predicting anti-vax content relies on the use of NLP as a solution for analysing textual information. Research has demonstrated that the use of NLP and other methods have the potential to identify tweets containing misinformation [183, 17, 85, 258, 335, 106], conspiracies [274, 240, 109] and hate speech [417, 303]. Within the NLP-based literature, a subset of research focuses specifically on the role of sentiment analysis for identifying both positive and negative view points. Sentiment analysis has been used to understand the various themes and trends surrounding support and opposition towards vaccines [190, 272, 209, 125]. The research shows that users promoting anti-vax related content frequently engaged in replies and, overall, were more negative showing emotions such as rage and sorrow as one of the key themes [265, 94]. Alternatively, a network science-based approach has been used to demonstrate how anti-vax users form echo chambers of polarised communities with other like-minded users by retweeting each other's content [262].

As mentioned previously, the existing research on Twitter suggests that a gap in the literature for a network-based methodological approach for detecting anti-vax is present. Furthermore, very few studies consider the importance of considering multiple user-to-user network representations based upon different types of interaction.

Overall, this investigation of the hypothesis expands upon existing literature by using Twitter to assess the ways in which network-based approaches can be used to model interactions surrounding disruption (see Research Question 2). The literature demonstrates how these three interactions have potential to provide predictive utility for detecting controversial and non-controversial networks among various anti-vax topics.

5.4.2 Dataset

All tweets that are used in this investigation are a part of the Coronavirus (COVID-19) Tweets Dataset [232]. This dataset serves as a baseline for our analysis as it uses a set of broad, predetermined, generic keywords which are of relevance to the COVID-19 pandemic. In particular, subsequent analysis is focused on a specific window between 9th November 2020 and 8th December 2020. This date range refers to the period when the initial vaccines were first approved for use in the United Kingdom thus starting the conversation and reactions around vaccines on Twitter [104].

In addition to this, this investigation in introduces a set of potentially controversial hashtags and keywords (referred to as "terms" within this investigation). There are two particular papers within the supporting literature which provide useful examples of terms which align with disruptive activity. These papers were selected due to their transparency in reporting the set of terms they used to conduct their analysis. As a result, this simplified the process of discovering relevant terms for our analysis.

The first paper of interest ("#Scamdemic, #Plandemic, or #Scaredemic: What Parler Social Media Platform Tells Us about COVID-19 Vaccine") uses data originating from Parler an "alt-tech microblogging" site for Trump supporters and conspiracy theorists alike [29]. While the data in this investigation didn't originate from Twitter itself, Parler as a platform is well known for its issues regarding echo chambers, filter bubbles and lack of fact-checking [68, 11] which, in turn, can lead to disruptive behaviour. For example, one Hashtag in particular "#echo" is used to purposely encourage users to reinforce their existing beliefs on COVID-19 vaccine efficacy and vaccine acceptance.

The second paper of interest ("COVID-19 Vaccine Hesitancy on Social Media: Building a Public Twitter Data Set of Antivaccine Content, Vaccine Misinformation, and Conspiracies") focuses on attempts to build a set of anti-vax related content by manually collecting keywords used exclusively within the context of vaccine hesitancy from previously known / observed anti-vax disruptive behaviour such as #vaccineskill or

#vaccinedamage [274]. In doing so, these were treated as seed keywords to find other, co-occurring terms that are of relevance and to avoided less relevant tweets through manual intervention. Overall, the tweets collected from their analysis discovered that the vast majority of keywords mentioned in the paper are related to known (yet, debunked) conspiracy theories based upon shared websites with questionable credibility as verified through the Iffy+ database of low credibility sites[1].

In addition to these two papers, we extend our set of terms to include the original keywords / hashtags used as part of the Coronavirus (COVID-19) Tweets Dataset along with a few additional terms that were manually selected using the Twitter search tool. The search tool was used to broaden our existing set of terms and to ensure that we did not exclude potentially relevant terms which were not featured in previous analysis.

Overall, the complete set of terms used as part of this investigation can be found in Table 7.6 and will be used as part of a crowdsourced ranking task where participants rank each keyword according to how controversial they are.

The combination of an applied date range and relevant terms ensures that the tweets are of relevance to vaccine related content Furthermore, the scope of the dataset will ensure that ambiguous terms such as #microchipping, #arrestbillgates and #populationcontrol will focus on a very specific subset of tweets which are both of relevance to COVID-19 and the anti-vax community and will unlikely be confused with anything else.

5.4.3 Methodology

The methodology used as part of this investigation can be broken down into several stages. To begin, each of the keywords featured in the dataset are ranked as part of a crowdsourced task to determine controversial terms. Secondly, tweets are "hydrated" to retrieve the original content of the tweet from its ID. Thirdly, using the hydrated tweets, user-to-user networks are generated for each of the given terms broken down

[1] `https://iffy.news/`

by interaction type. Fourthly, both global and local network features are then extracted for subsequent analysis. Finally, the metrics extracted as part of the global and local analysis are used as part of classification task in an attempt to detect controversial from non-controversial terms.

Ranking of Terms

To begin, a total of ($N = 5$) participants were recruited and were asked to rate each unique term listed in Table 7.6 on a Likert scale where each term is scored according to a weight $w \in [0..4]$ based upon the following scale: "Neutral" ($w = 0$), "Somewhat Controversial" ($w = 1$), "Controversial" ($w = 2$), "Very Controversial" ($w = 3$), "Highly Controversial" ($w = 4$). A 5-point scale was used to ensure that enough detail was provided to assign an accurate label to a term to determine the extent to which a term is controversial across a spectrum. Furthermore, The design of the Likert scale makes it possible to score many terms at scale.

Similar to Wikipedia (see Chapter 4, Section 4.4), this approach exploits the "wisdom of the crowd" whereby the collective opinion of a particular term is considered for finding controversial terms [359].

The results from each participant are then aggregated to include the total (sum of scores), mean and standard deviation of each score. Each of the terms were then ranked according to mean score. To reduce the score to a binary classification, a suitable threshold is determined according to the distribution of mean scores. A score which exceeds the threshold is considered "controversial" and those which are below the threshold are considered "non-controversial".

Hydrating Tweets

In order to retrieve the specific content and metadata (such as the timestamp, reply to, retweet and mention fields) from the Coronavirus (COVID-19) Tweets Dataset, the

tweets need to be "hydrated" from the original ID where the Twitter API is required to lookup and retrieve the original tweet. In doing so, the process of "hydrating" takes the list of IDs provide by the IEEE dataset and transforms them into a set of tweets complete with the information need for subsequent analysis.

Network Generation

As mentioned previously, this investigation focuses on three different types of interaction of interest: quote retweets, mentions and replies. A single network $G_i = (V, E)$ is generated for each of one of these interactions. A node $v_i \in V$ represents a user and the presence of a directed edge $(v_i, v_j) \in E$ indicates an interaction towards another user. For example, $v_i \rightarrow v_j$, can be interpreted as "v_i *mentions/replies to/quote retweets from* v_j".

For each term used in this investigation (see Table 7.6 for complete set) the three interaction network are generated conditioned on the presence of the term appearing in the body of a tweet. For example, a subset t of all tweets $t \subset T$ are determined by only focusing on tweets which contain the term "#covid19". For all tweets in this subset, three distinct networks are extracted according to the presence of one of the interactions of interest. Overall, a total of $N = 199$ terms are considered focusing on $M = 3$ interactions of interest producing a total of $N \times M = 597$ unique networks.

To evaluate the utility of these network representations, a classification task is used to evaluate how well they perform in predicting controversial terms from non-controversial terms. This is achieved using two sets of network features - global and local network features. These are explained as follows.

Global Network Features

This investigation considers the network-based metrics at a global-level by observing how users in each interaction network behave collectively. These include density (the

capacity of how many interaction edges occupy the network), reciprocity (the ratio of bidirectional ties in the network), transitivity (the extent to which nodes for transitive edges in a triad), in degree (mean,max,min), out degree (mean,max,min).

These metrics provide genetic structural properties and also serve as a baseline to determine the predictive utility in comparison to the local network features. Properties such as density, reciprocity and transitivity are fundamental for capturing social traits such as trust, friendships and communities within social networks [242, 50, 357, 203]. These, can be used to provide predictive signals for differentiating between controversial and non-controversial which, in turn, addresses Research Question 4 of the hypothesis.

Local Network Features

In addition to global network features, local network features are derived by counting the frequency of all induced subgraphs of a fixed size. These address Research Question 4 by considering the role of both local and global network features for detecting disruptive behaviour. As mentioned in Chapter 3, Section 3.7 due to the subgraph isomorphism problem [102], subgraph counting does not scale well with time and this therefore resource intensive. This means that counting subgraphs grows exponentially with time and that larger subgraphs take much longer to compute. For this reason, only subgraphs containing 3 and 4 nodes are considered producing $N = 13$ and $N = 199$ possible combinations respectively with a total of $N = 212$ subgraphs overall. Each subgraph is assigned a label from `S1` to `S212`. Each interaction network for a given term produces a vector V_{G_i} where:

$$V_{G_i} = (v_1, v_2, \ldots, v_{212}) \tag{5.1}$$

and where v_i represent the frequency of the ith subgraph in the set. In addition to this, each of these vectors V_{G_i} are normalised making it possible to compare to networks of

different sizes using the following:

$$V_{G_i} = \frac{1}{\sum_{j=1}^{212} v_j}(v_1, v_2, \ldots, v_{212}) \tag{5.2}$$

As a result, V_{G_i} is used to represent the ratio of subgraph frequencies and provides the basis for discovering under and over-representations of induced subgraphs with respect to other networks. Furthermore, this approach can be used to determine the extent to which interactions and terms share similar structural features.

Prediction

In line Research Question 4, this task is performed to assess the feasibility regarding whether it is possible to detect controversial terms from non-controversial terms and to understand the predictive utility of global and local features. The purpose of these two types of features is to address the predictive capabilities of user-to-user networks using each set of features (global and local) as baseline comparison to help answer Research Question 4 by extracting behaviour and signals.

As described in Research Question 4 we use binary logistic regression (BLR), support vector machine (SVM) and a random forest classifier (RFC) are the classifiers of choice. To assess the classification performance 10-fold cross-validation is applied where the accuracy, sensitivity, specificity, positive predictive value (+PV) and negative predictive value (-PV) of each of results is reported.

5.4.4 Results

Discovery of Anti-vax Terms

A crowdsourced ranking task was used to discover controversial Twitter terms from non-Controversial terms. The combined results from all participants reveal that most

terms were labelled as "Neutral" which would appear, on average, 37.9% of the time. The second most frequent label was "Somewhat Controversial" appearing 20.8% of the time. The most uncommon label is "Highly Controversial" at 8.74%.

The full results are compared in Figure 5.5 and reported in Table 5.5.

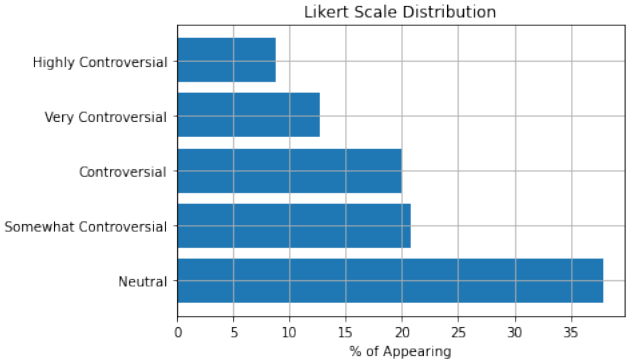

Figure 5.5: Distribution of all labels used within the Likert scale based upon appearances. Overall, terms considered "Neutral" appeared the most.

Label	**% of Appearing**
Neutral	37.89%
Somewhat Controversial	20.8%
Controversial	19.9%
Very Controversial	12.6%
Highly Controversial	8.74%

Table 5.5: Full list of labels used within the Likert scale and the probability of appearing.

As mentioned in Section 5.4.3 the results of the ($N = 5$) participants are aggregated to include the total score, mean and standard deviation. The complete results for each term can be found in Table 7.7. The distribution of mean term score is shown in Figure 5.6.

The distribution is used to determine the position of a threshold t as a cut-off point for separating non-controversial and controversial terms. Based upon the distribution of mean scores (as shown in Figure 5.6) terms are partitioned into the two groups according to a set threshold of $t = 0.95$. Terms were assigned the label "non-controversial" if the average score was $< t$, otherwise considered "controversial" to some extent.

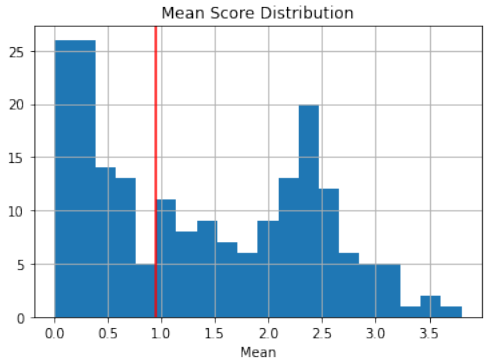

Figure 5.6: Distribution of mean score for all terms in the set reveals two distinct peaks in values which are centred around 0 and 2.4. A threshold of $t = 0.95$ is marked in red and is used to indicate how the data is partitioned in two.

Using this classification technique, a total of ($N = 115$) controversial and ($N = 84$) non-controversial terms were discovered producing a 58/42 split. The complete classification details are shown in Table 7.8.

Using the labels provided as part of the classification task, the data was split into the two sets according to the appropriate label. As a result of computing the global and local metrics a few additional observations of interest emerged which are outlined as follows. Furthermore, principal component analysis is performed to determine the spatial relevance for both global and local network features.

Global Network Features

Among all the global features, density, transitivity and reciprocity reveal the most distinct characteristics between each of the three interaction types in isolation and provide initial insights on how social interactions are structured between each of the two classification labels. The distribution of values for density, transitivity and reciprocity can be found in Figures 5.7, 5.8 and 5.9 respectively.

The results in Figure 5.7 reveal how all three interactions share a similar distribution of density among all controversial networks with a slight rise in value towards the end of the distribution.

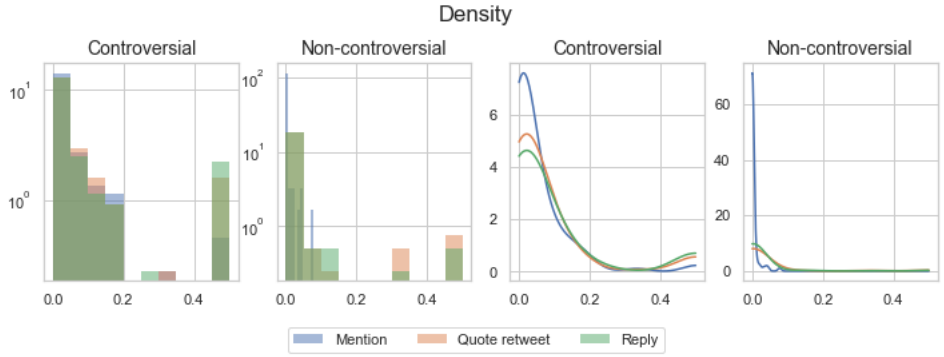

Figure 5.7: Histogram comparing the density distribution of all networks, grouped by controversial and non-controversial terms for each interaction (mention, quote retweet and reply).

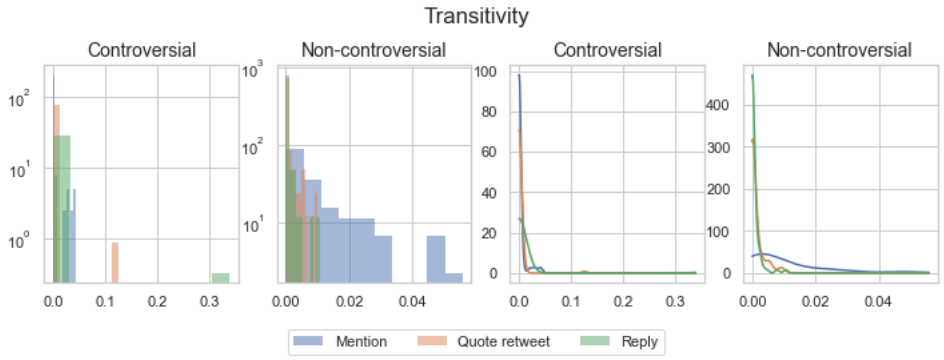

Figure 5.8: Histogram comparing the transitivity distribution of all networks, grouped by controversial and non-controversial terms for each interaction (mention, quote retweet and reply).

The distribution of transitivity in Figure 5.8 reveals how mentions are more likely to features transitive connections within non-controversial networks than controversial networks.

The results featured in Figure 5.9 indicate how reciprocity is more significant among the reply integration as opposed to mentions and quote retweets and has a strong presence in both networks equally.

Each of the interaction types can be combined such that a single network can represent by all three interaction types as opposed to in isolation. As a result, principal component analysis (PCA) is used to determine the spatial relevance for each local network feature

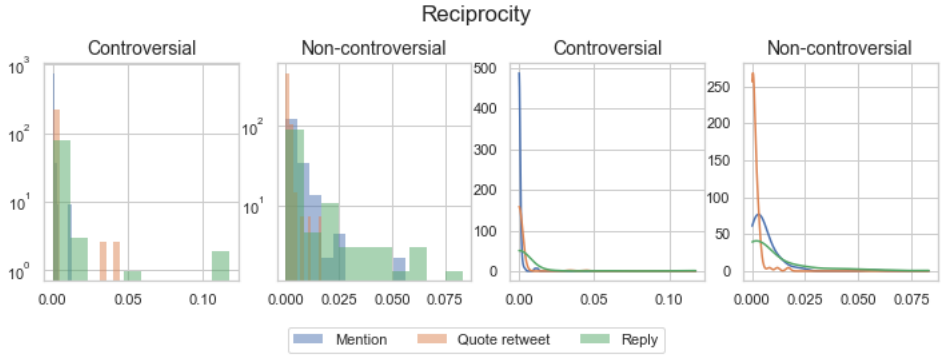

Figure 5.9: Histogram comparing the reciprocity distribution of all networks, grouped by controversial and non-controversial terms for each interaction (mention, quote retweet and reply).

using a two-dimensional projection. The results can be observed in Figure 5.10 with supplemented with the corresponding eigenvector values in Figure 5.11.

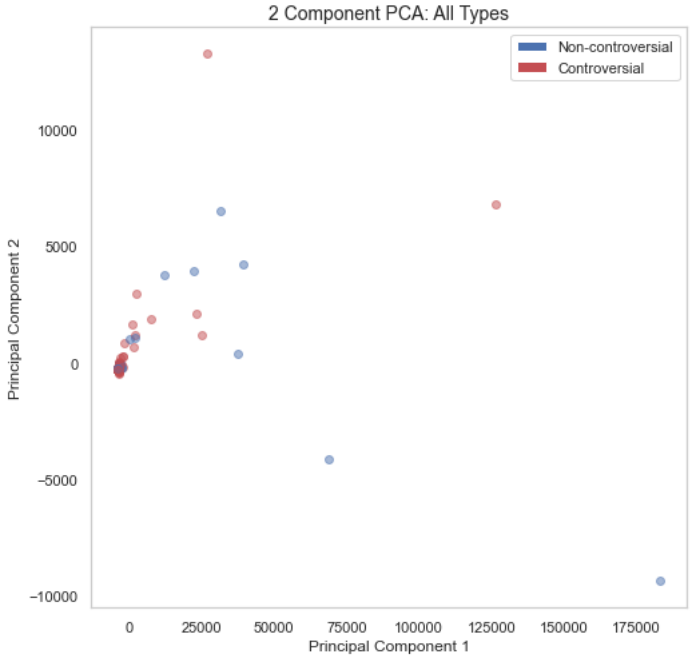

Figure 5.10: Two-dimensional principal component analysis is performed on all global network features where each interaction (reply, mention and quote retweet) is combined into a single feature vector.

The PCA scatter plot in Figure 5.10 reveals how there are no obvious spacial clustering or patterns which emerge between the two types and provides little spatial utility. The

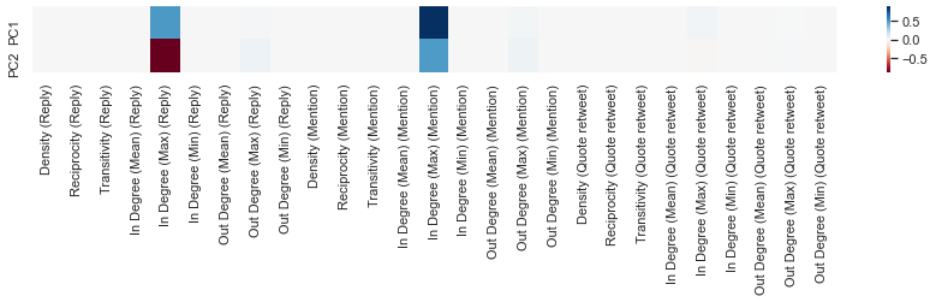

Figure 5.11: The corresponding eigenvector values for each principal component used in Figure 5.10 is shown to indicate important metrics which contribute to the spatial positioning of networks in the PCA plot.

PCA eigenvectors demonstrate how "In Degree (Max) (Reply)" and "In Degree (Max) (Mention)" appear as the strongest features used for each principal component. This means that networks which are characterised by nodes with a high in degree (according to reply and mention) are mostly likely to contribute to the detection of controversial and non-controversial networks.

Local Network Features

Due to the size of each of the feature vectors ($N = 212$ for each type of interaction), PCA is performed to reduce the size of the feature space making it possible visualise the data in two dimensions for each type of interaction. Furthermore, the PCA eigenvectors are used to determine inflectional subgraphs which contribute to the spacial positioning of each feature vector. These results are presented in Figures 5.12 and 5.13 for the PCA scatter plots and eigenvector values respectively.

The results in Figures 5.12 and 5.13 suggest that there are no obvious cluster formations which are unique between each of the three interaction types. This means that there is little indication of clustering potential and that the use of local features may be unsuitable for detecting nuances between each of the networks.

Additionally, the coefficients in Figure 5.13 reveal that same set of subgraphs are dominant throughout each of the interaction types. By setting a threshold of $t = 0.1$

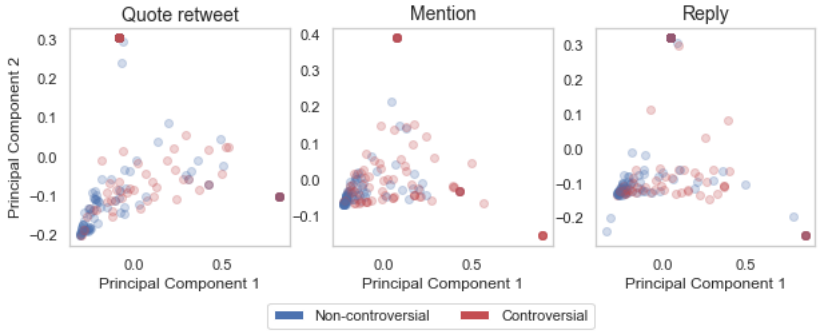

Figure 5.12: Two-dimensional principal component analysis is performed on all local network features for each interaction type reply, mention and quote retweet) in isolation.

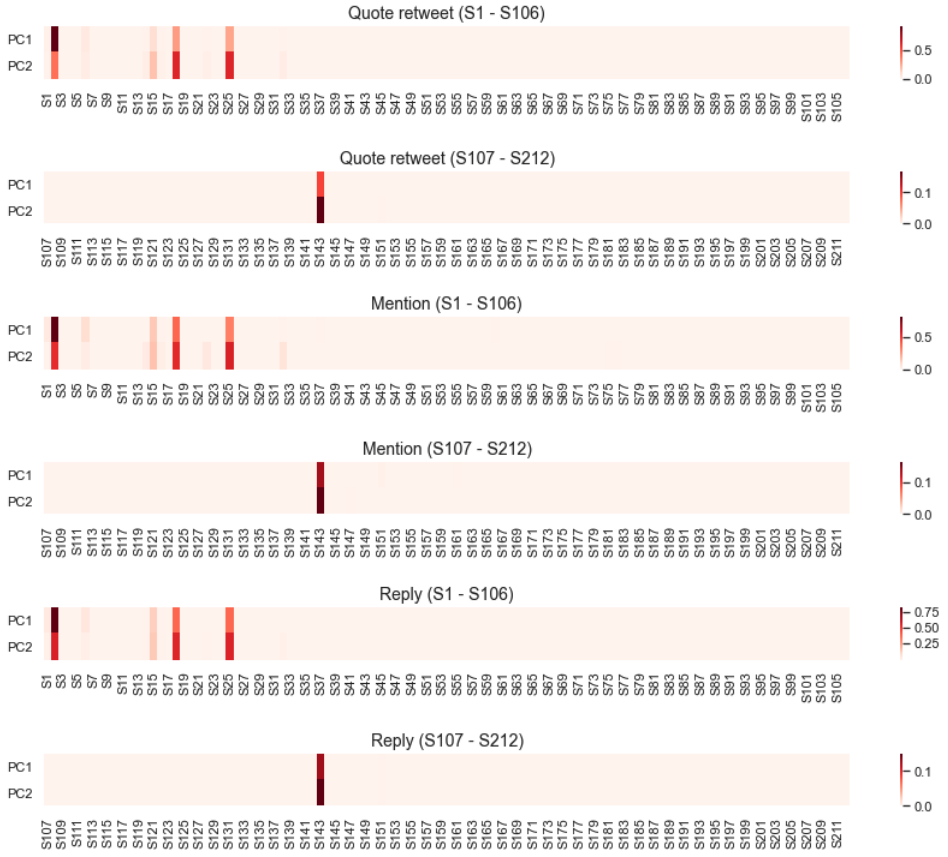

Figure 5.13: The corresponding eigenvector values for each principal component and interaction type used in Figure 5.12 is shown to indicate important subgraphs which contribute to the spatial positioning of networks in the PCA plot.

(determined by observing the distribution of values), across all the eigenvector values, a total of five subgraphs emerged which exceeded this threshold. These include subgraphs

S2, S15, S18, S25 and S143 which appear consistently across all three interactions and are shown in Figure 5.14 for reference.

Figure 5.14: A subset of subgraphs which are dominant within the PCA eigenvector values of local features for all three interactions. Each of the subgraphs features interactions centred around a single user in an egocentric fashion.

5.4.5 Classification of Controversial Terms

Using the data produced in earlier tasks, the feature vectors produced from global and local features (see Sections 5.4.3 and 5.4.3) are used as part of a classification task with the intention to predict controversial terms from non-controversial terms using network-based features exclusively (see Research Question 4). This is done with the intention of understanding the predictive utility of both global and local network features by comparison.

Global Network Features

As discussed earlier (see Section 5.4.3), global features as used to address the predictive utility of user-to-user networks along with local features as part of a baseline comparison. Each of the three classification models are trained using the raw metrics defined in Section 5.4.3. The classification results are reported for each type of interaction and combined. The results can be found in Table 5.6 and 5.7 for each interaction in isolation and combined respectively.

The results in Table 5.6 show the RFC outperforms both SVM and BLR consistently across each type of interaction using global features. As highlighted in Table 5.6, mention interactions combined with an RFC produces the best performing classifier

	Mention			**Quote retweet**			**Reply**		
Classifier	BLR	SVM	RFC	BLR	SVM	RFC	BLR	SVM	RFC
Accuracy	0.744	0.619	**0.81**	0.68	0.609	**0.746**	0.698	0.598	**0.716**
Sensitivity	0.544	0.203	**0.734**	0.432	0.235	**0.605**	0.475	0.15	**0.65**
Specificity	0.921	**0.989**	0.876	0.909	**0.955**	0.875	0.899	**1.0**	0.775
+PV	0.695	0.583	**0.788**	0.635	0.575	**0.706**	0.656	0.567	**0.711**
-PV	0.86	**0.941**	0.841	0.814	**0.826**	0.817	0.809	**1.0**	0.722

Table 5.6: Complete classification results for all three interaction types using global features reporting the performance for each classifier. The best performing classifier is highlighted in bold.

with an accuracy of $p = 0.81$. This is then followed by quote retweets with an accuracy of $p = 0.746$ and finally, reply with an accuracy of $p = 0.719$

Across all interaction types and models it can be observed that overall classification performs better at detecting non-controversial networks (negative class) compared with detecting controversial networks (positive class) according to negative predictive value (**-PV**) and positive predictive value (**+PV**) respectively. Furthermore, this is reflected in sensitivity which, in most cases, is much lower than specificity except for RFC in which sensitivity is much higher.

Additional classification tasks are performed whereby all global network features for each interaction are combined into a single feature vector. All interaction types are combined to assess whether this has an impact on classification performance. These results can be found in Table 5.7.

Classifier	**BLR**	**SVM**	**RFC**
Accuracy	0.847	0.852	**0.886**
Sensitivity	0.192	0.154	**0.269**
Specificity	0.96	0.973	**0.993**
+ Predict Value	0.873	0.869	**0.887**
- Predict Value	0.455	0.5	**0.875**

Table 5.7: Classification results combining all three interaction types using global features. The best performing classifier is highlighted in bold.

The results in Table 5.7 indicate that by combining all interaction types, the accuracy of the RFC increases to around $p = 0.886$. This produces a performance gain of

approximately 9.3% compared with the mention RFC accuracy - the best performing classifier of all interaction types.

While accuracy has improved, the sensitivity of all three classifiers has dropped significantly and only yielding $p = 0.269$ (RFC) at best. This suggests that many controversial networks are being classified as false negatives. This may be due to a combination of class imbalance and over-fitting as a result of high-dimensional data.

Local Network Features

As mentioned previously (see Section 5.4.3), local network features are used to assess whether it is possible to extract predictive signals from the user-to-user a network representation in the same manner as global network features. Following the same format as Section 5.4.5, the same three classifiers (RFC, BLR and SVM) are used with local network features exclusively. The classification results for each interaction type can be found in Table 5.8.

	Mention			**Quote retweet**			**Reply**		
Classifier	BLR	SVM	RFC	BLR	SVM	RFC	BLR	SVM	RFC
Accuracy	0.72	0.732	**0.786**	0.663	0.663	**0.704**	0.663	**0.692**	0.615
Sensitivity	**0.785**	**0.785**	0.759	**0.679**	0.593	0.556	**0.662**	**0.662**	0.45
Specificity	0.663	0.685	**0.809**	0.648	0.727	**0.841**	0.663	0.719	**0.764**
+PV	0.776	0.782	**0.791**	**0.687**	0.66	0.673	0.686	**0.703**	0.607
-PV	0.674	0.689	**0.779**	0.64	0.667	**0.763**	0.639	**0.679**	0.632

Table 5.8: Complete classification results for all three interaction types using local features reporting the performance for each classifier. The best performing classifier is highlighted in bold.

Much like Section 5.4.5, RFC is the highest performing classifier for each interaction type except reply interactions where SVM performs the best. Similarly, mention interactions outperform quote retweets and replies with respect to predictive utility and performance. Each of the feature vectors used as part of the classification task in Table 5.8 are combined and used in a separate classification task with the results shown in Table 5.4.5.

Overall, both the positive predictive value (+PV) and negative predictive value (-PV) are fairly consistent across all interaction types and classifiers. Similarly, the sensitivity is also fairly consistent across each of the three classifiers and interaction types and is not as low compared to the global results in Table 5.6. Contrary to the results in Table 5.6, RFC is best at detecting true negatives compared with other classifiers.

Classifier	**BLR**	**SVM**	**RFC**
Accuracy	0.852	0.852	**0.858**
Sensitivity	0.0	0.0	0.115
Specificity	1.0	1.0	0.987
+ Predict Value	0.852	0.852	0.865
- Predict Value	-	-	0.6

Table 5.9: Classification results combining all three interaction types using local features. The best performing classifier is highlighted in bold.

The results in Table show that RFC is the best performing classifier when all interactions are combined using local features. The accuracy of the RFC model increased to $p = 0.858$ which, in turn, produces a performance gain of 9.16% compared with the best performing result in Table 5.8.

In view of these result, it is also important to acknowledge the low sensitivity values. Much like the results in Table 5.7, when all interaction types are combined for local features, the sensitivity is incredibly low and, in some cases (BLR and SVM), non-existent. Furthermore, the negative predictive value for both BLR and SVM are absent due to the classifiers incorrectly reporting zero true and false negatives.

As discussed in Section 5.4.5 it is reasonable to speculate that this is due to the imbalance between each class which is further compounded by the possibility of over-fitting due to high-dimensional data.

Comparing Feature Performance

Overall, the classification results between both local and global network metrics area fairly consistent however, it is evident that global features marginally outperform local

network features when predicting between controversial and non-controversial terms. When interaction networks are combined, global features ($p = 0.886$, best performing classifier) have a higher overall accuracy than local features ($p = 0.858$, best performing classifier). This produces a performance gain of 3.26%.

In addition to this, the relative classification gain Δ_{ij} between interaction type is determined by taking the best performing classifier and obtaining the difference between classification values p_i, p_j scaled by the original value of the classifier of interest p_i such that $\Delta_{ij} = \frac{(p_j - p_i)}{p_i} * 100$. These results are shown in Figure 5.15 for both local (left) and global (right) features.

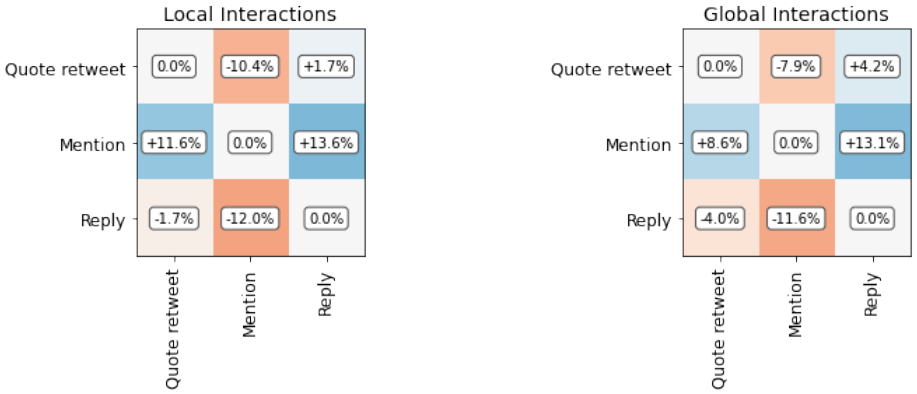

Figure 5.15: Pairwise comparison of the classification gain (or loss) for the best performing classifier for each interaction type for both local (left) and global (right) network features. Interaction types on the y-axis are compared with interactions on the x-axis.

It is evident from Figure 5.15 that mention interactions consistently outperform both quote retweets and reply interactions for both local and global network features. Reply interactions provide the least predictive utility - especially when compared with quote retweets and mentions.

The calculations used as part of Figure 5.15 are repeated to determine the relative classification gain between local and global network features for each classifier based upon the overall performance of each interaction type by pairwise comparison (Figure 5.16) and independently (see Figure 5.17).

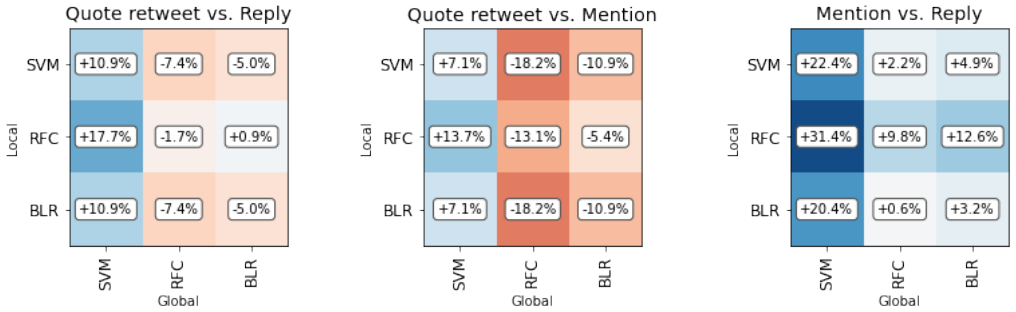

Figure 5.16: Pairwise comparison of the classification gain (or loss) for each classifier comparing local features (y-axis) against global network features (x-axis). Each heatmap represents a pairwise comparison between two types of interaction network.

Figure 5.16 demonstrates that local features using all three classifiers outperform SVM using global features for each of the cross-compared interaction types. Furthermore, it is also evident that local features lack performance when quote retweets and mentions are cross-compared. However, this trend is reversed when local network features are compared with global features when mentions are replies are cross-compared.

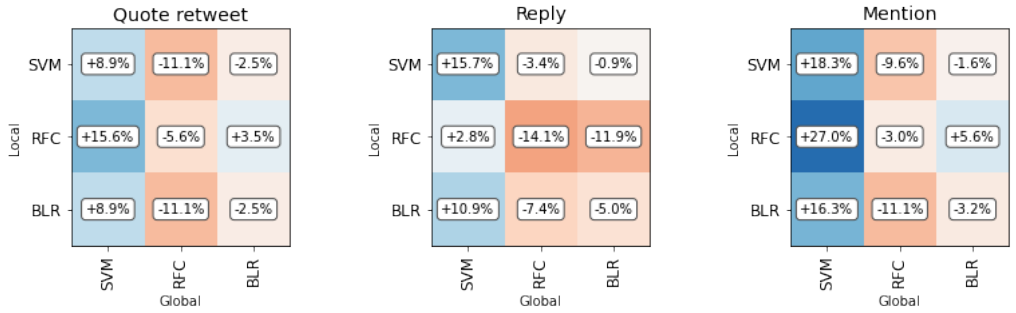

Figure 5.17: Pairwise comparison of the classification gain (or loss) for each classifier comparing local features (y-axis) against global network features (x-axis). Each heatmap represents a single interaction network.

Much like Figure 5.16, Figure 5.17 reaffirms the observation that local features using all three classifiers outperform SVM when global features are used. It is clear that generally, global features provide the most predictive utility when compared with local features and that mention interactions highlight the magnitude of classification gain between local and global features.

5.4.6 Discussion

Both the data overview and classification results (see Section 5.4.4) provide meaningful insights on the utility of using network representations for predicting controversial and non-controversial terms. As a result, a number of key observations are explored and discussed further in this section which relate to the utility of both global and local network representations and their prediction performance. These are discussed as follows.

Global Metrics Overview

As addressed in Section 5.4.3 and Research Question 4, global metrics are considered for studying properties which characterise the entire "global" structure of a network using simple metrics such as transitivity, density and reciprocity.

The results demonstrate that the distribution of density in Figure 5.7 is consistent across all three interaction types. These results suggest that there are a few controversial networks which are denser than others. Non-controversial networks share similar properties. The general trend is that mentions produce the least dense networks of the three network interactions. Furthermore, the distribution of transitive ties in Figure 5.8 indicate that non-controversial networks are more likely to feature tightly connected communities and triads which suggests that users are engaging with one another using directed messaging through mutual connections

Finally, the distribution in Figure 5.9 reveals how reciprocity is used to capture conversational signals where a user is likely to respond back using reply-based interactions. In this context, reply-based interactions are most likely to appear in a non-controversial network as opposed to controversial networks suggesting that users are more likely to participate in a conversation by responding to a reply made by someone else in a discussion thread. These results are reaffirmed by the PCA eigenvectors in Figure 5.11

whereby both mentions and replies appear as the most dominant features according to the maximum in degree.

Overall, the role of global metrics in this investigation demonstrate how the user-to-user network can be used to capture a diverse range of affordances across different types of interaction (see Research Question 3).

Local Metrics Overview

Much like global metrics, the use of local metrics serve the purpose of understanding structural properties by examining a network's substructures, as addressed in Research Question 4.

As described in Section 5.4.4, the PCA eigenvectors shown in Figure 5.13 reveals five subgraphs (see Figure 5.14 for subgraphs `S2`, `S15`, `S18`, `S25` and `S143`) which exceed an arbitrary threshold of $t = 0.1$. These subgraphs reflect those that resemble a tree-like structure where all interactions are centred around one node (similar to the previous investigation).

This potentially correlates with features such as the maximum in degree of a network where many nodes are directed towards a single "central" node. This is evident based upon the features which emerged in the PCA eigenvectors for global features in Figure 5.11 where the maximum in degree is dominant.

As a result, these findings reaffirm the utility of user-to-user networks for capturing a diverse range of affordances based upon the three types of interaction (see Research Question 3) which are used in this investigation.

Global Prediction

As described in Research Question 4, prediction is an important process for investigating the hypothesis which allows of the detection of disruptive activity. Within this invest-

igation, the use of global features for predicting controversial and non-controversial networks reveals rather promising results. The result indicate that an RFC (Random Forest Classifier) consistently outperforms alternative classifiers across each of the three interaction types. The results in Table 5.6 show that mention interactions can best differentiate between the two types with a classification accuracy of $p = 0.81$.

Based open the eigenvector values in Figure 5.11, it is possible to speculate that the maximum in degree for mention interactions (one of the most dominant features) provide the best spatial distribution of networks in order to separate the two groups. As a result of combining all three interactions, it is also reasonable to imply that improved accuracy of $p = 0.886$ is due to the maximum in degree for replies also being a dominant feature in the PCA eigenvectors in Figure 5.11.

Local Prediction

By using local features for differentiating between controversial and non-controversial networks a similar trend can be observed compared with global features. The results are fairly consistent across each of the interactions which is unsurprising considering that the PCA eigenvectors in Figure 5.13 for each interaction are almost all identical. Similarly, mention interactions provide the best result with respect to accuracy ($p = 0.786$) which is improved to $p = 0.858$ when all interactions are combined.

Prediction Comparison

Based upon the results in Figure 5.15, it is clear that between each of the three type of interaction mentions outperform both quote retweets and mentions but quite a significant margin. These results are reflected in both sets of features where global features overall provide the most predictive utility when it comes to differentiate between controversial and non-controversial networks. In addition to this, the distinction between global and local features are further exacerbated in Figure 5.17 as most results indicate that local

feature significantly lack perforce when each of the classifiers are cross-compared with each other between the three interaction types. With respect to the hypothesis, these findings address Research Question 4 by demonstrating that global features outperform local features for providing signals in which the prediction of disruptive behaviour can be made.

5.4.7 Key Findings

The results gained provide meaningful insights towards understanding the utility of using social network representations for differentiating between controversial and non-controversial networks using platforms which align with the message data structure (see Chapter 3, Section 3.2.3). This investigation uses COVID-19 vaccinations as the central point of discussion - a known target of disruptive behaviour (see Section 2.1.4 in Chapter 2). This helps address the hypothesis of this thesis in numerous ways. In particular, this investigation supports Research Question 4 by using a prediction task to show how three message-based user interactions (quote retweets, mentions and replies) provide network-based representations (see Research Question 2) to understand how users behave collectively based upon their underlying network substructures and metrics.

The results help address Research Question 4 by providing evidence that simple graph-based metrics such as in/out degree, density, reciprocity and transitivity are sufficient for differentiating between controversial and non-controversial networks. Each type of global metric yields different results depending on the type of interaction. For example, non-controversial networks more likely to feature reciprocated ties and transitive using reply and mention-based interactions respectively. By combining all three interactions into one, it is clear that global features adequately capture the nuances between each of the networks using relatively few features - contrary to subgraph “local” approach. As well as providing a performance gain, the set of global features used in this investigation are relatively easy to calculate and can almost be done in near real time.

The implications of this investigation impact how we are to consider using quote retweets, mentions, and replies (or a combination of the three) in future work as part of a framework of diverse affordances (see Research Questions 2 and 3). The research featured in this investigation clearly demonstrate that it is possible to determine controversial networks for non-controversial networks using interactions derived from human behaviour exclusively (see Research Question 1).

The clear advantage of this is that very little textual analysis (e.g. NLP) is needed making it possible to replicate results using a non-English speaking corpus. The methods used as part of the investigation can be applied to social media platforms beyond Twitter due to widespread adoption of features such as replying and sharing - fundamental user-to-user interactions which apply to almost all social media platforms. Additionally, these results can be applied to content moderation on social media whereby moderators can use similar techniques to identify the presence of controversial terms based upon the signature of the global network structure.

5.5 Egocentric Reply Networks and Temporal Features

Much like previous work, the focus of this investigation is centred around Reddit as the platform has been the subject of recent foreign interference from social bots manipulating users through political propaganda[2]. As mentioned previously, this is of significant concern given that it has the potential to disrupt the functioning of democracy and has been observed through events such as Brexit (see Section 2.1.2) and the 2016 US Presidential Election (see Section 2.1.3).

As mentioned in Chapter 3, Section 3.3.2, Reddit allows users to conveniently comment and share content within communities of interest. Participants can establish a reputation within a community of peers and engage with others having similar interests. This is

[2]Reddit Transparency Report 2017: `https://www.redditinc.com/policies/transparency-report`

achieved through so-called *subreddits* where users publicly post content of relevance to a particular topic. Furthermore, Reddit encourages users to express different opinions (both positive and negative) through the use of voting whereby uses can up-down and down-vote as post or comment which further in answers his popularity. Consequently, this means that users are subject to complex echo chambers which could be further compounded by other issues such as trolling.

Perhaps one of the most important features to Reddit is the ability to form and participate in conversations through the use of the discussion or comments section. As described in Chapter 3, Section 3.4.2, this activity aligns with the message data structure (see Section 3.2.3) which, in turn, motivates the use of the user-to-user network representation.

The purpose of this investigation is to explore the hypothesis by contributing to the previous investigation on Twitter by demonstrating the versatility of the user-to-user network representation which can be used as part of a framework (see Research Question 2) for modelling activity through behavioural networks (see Research Question 1) and to discover signals which could lead to the detection of disruption (see Research Question 4).

5.5.1 Related Work

The methods presented in this investigation are similar to techniques used for processing bulletins board by modelling a directed network of users replying to others within nested discussion threads [412, 256]. These networks exhibit valuable metrics which provide the basis for analysing user behaviour such as leadership within discussions [75], assessing topic discovery with hierarchical quality [392] and predicting user interactions [151]. These network structures facilitate the discovery of interaction patterns and can be used to model the behaviour present within basic discussions (see Research Questions 1 and 4).

Furthermore, temporal features also provide valuable insights towards better understand-

ing user behaviour on social media. This is achieved by assessing the timings of various activities using spike train analysis to observe an ordered sequence of events. Evidence suggests that distinct temporal features contribute to the detection of social media bots and spammers [103, 137, 267]. These features can be used to aid the detection of disruptive behaviour in a language-agnostic manner (see Research Questions 4).

In addition to this, research has been performed to understand the value of significant patterns in behavioural networks with examples such as situational understanding [61], information fusion [195], conspiracy theories [330], user influence [273, 290] and quality [399, 108]. However, little research has been performed to study the impact of user behaviour from the perspective of a single user within the discussion thread. Contributions show that the use of egocentric networks can be used to classify different types of behaviour (Research Questions 1 and 3) in an online social setting [140] and, in so doing, explores the hypothesis of this thesis.

As a result, this investigation contributes to the literature by using Reddit as a platform to combine network-based and temporal features in detecting disruptive behaviour from bot-like activity (see Section 2.2.1). This address Research Question 4 and has the potential to improve the prediction accuracy for classification of social bots and supports situational awareness for social media users and the platform itself.

5.5.2 Dataset

To investigate the role of user-to-user networks on Reddit, the Reddit API was used to extract a sample of disruptive and non-disruptive normal users ($N = 794$ and $N = 850$ respectively). To begin, 100 random subreddits were selected where 10 random posts were sampled within each subreddit with the intention to get a uniform sample of users across the platform. From this, we used the overall 'comment karma' score k assigned by Reddit for each user in our dataset. This derived from a calculation based upon the ratio of positive 'upvotes' and negative 'downvotes' given by other users. Random

sampling was repeated until a near 50:50 split between disruptive and non-disruptive users was achieved. Overall, a total of $N = 469,606$ comments with approximately 454 comments per non-disruptive user and 75.6 comments per disruptive user. The karma score k was used as it represents the receptivity of the user's contribution to a discussion in a community, where something that is negatively received can be seen as being disruptive to the norms of that community.

5.5.3 Methods

To explore the hypothesis, this analysis investigates whether network and temporal-based features of a user's behaviour can be used to predict if they are considered disruptive or not (see Research Question 4). An egocentric reply network is defined (see Research Question 1) and is generated for each user X by forming a directed edge between user X and the other users that X has either replied to or has received a reply from. For each such network, the frequency of all induced 4-node subgraphs in a star formation are counted, where X is central to the induced star. 4-node subgraphs are chosen as they are much easier to process (in terms of computational overhead) and still preserves a suitable level of detail compared to that of 3-node subgraphs. There are 10 possible alternative configurations (see Figure 5.21). These capture the alternative edge configurations surrounding a target user, and the interactions taking place in terms of direction of communication. For each user X, the frequency counts are normalised of each type of induced 4-node subgraph. This expresses the proportions of 4-node subgraphs in which X is the centre of the star, such that each subgraph profile represents a single user (e.g., Figure 5.21). These results are combined with further frequency statistics relating to user activity on Reddit, including account age, number of historical comments, mean comments per week, and mean duration between comments. Together these features form a basis to characterise user activity and to predict disruptive behaviour based on karma.

	# Comments	Age	μ Comment rate	μ Duration
Normal	0.459	0.441	0.559	-0.682
Disruptive	-0.173	-0.099	-0.08	0.114

Table 5.10: Spearman correlation coefficients of temporal features with overall comment karma.

5.5.4 Results

The results focus on three important of this investigation: the role of temporal features, egocentric reply networks and prediction. Firstly, although temporal features are not network-based, they are language-agnostic and could be used to further enhance prediction results as part of Research Question 4. Secondly, the egocentric networks are considered to understand which subgraph formations are unique to disruptive and normal users. Finally, prediction is performed to determine the feasibility of detecting the two types of users using these features and address Research Question 4 as a result. The results are discussed further as follows:

Temporal Features

To begin, attributes associated with user activity and comment timings are considered. Figure 5.18 shows histograms for comment count (bottom right), account age (top right), mean comment rate (top left) and mean comment duration (bottom left), split between the two user groups.

The histograms in Figure 5.18 reveal distinct behaviour between normal and disruptive users. It is apparent that normal users produce a uniformly distributed number of comments, whereas disruptive users are less likely to have a large comment history. This “long-tail" log-normal type distribution is consistent, to varying degrees, across each histogram, where disruptive users are far more likely to less active between commenting and much less mature in age (see Figure 5.18).

To understand the relationship each variable has with respect to comment karma, we

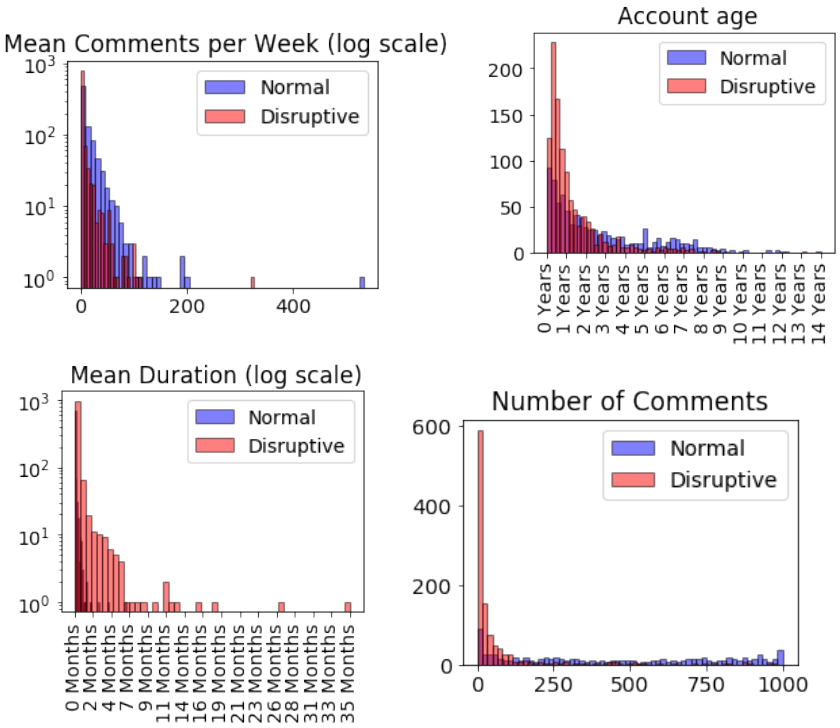

Figure 5.18: Histogram comparison of normal and disruptive users with respect to user activity. The first (top left) represents the distribution of mean comments per week per user. The second (top right) represents the distribution of account age by years per user. The third (bottom left) represents the average duration between activity per user. The fourth, and final, (bottom right) represents the average number of comments per user.

assess the correlation between each variable using Spearman tests (see Table 5.10). Contrary to disruptive users, variables such as age and comment rate reveal a partial correlation suggesting that more-mature accounts are less likely to be used for disruptive behaviour on Reddit.

Egocentric Reply Networks

Figure 5.19 provides two examples of user interactions, where the centre node is our target user and edges going out indicate a reply and edges being received represent receiving a reply. These are featured around individual users rather than groups. The examples provided here are used to illustrate that normal users are more active than

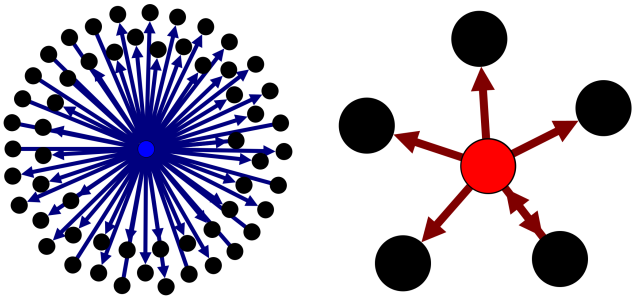

Figure 5.19: Examples of typical normal (left) and disruptive (right) egocentric user reply networks. Normal users are typically characterised with many interactions (high degree), whereas disruptive users have very few interactions by comparison (low degree).

disruptive users and are more likely to participate in a two-way conversation.

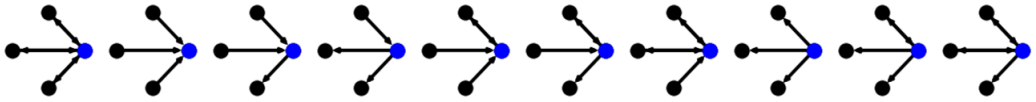

Figure 5.20: The complete set of all 4-node 'star' subgraphs featuring the target node highlighted in blue for every possible edge combination.

We observe the user interactions by counting and enumerating over all 4-node subgraphs resembling a tree-like structure where the root node serves as our target user of interest (see Figure 5.21). This allows us to examine the occurrence of edge configurations surrounding an individual user's interactions, when we consider the induced 4-node star configurations. The proportion of all induced 4-node star configurations for each user is presented in Figure 5.21.

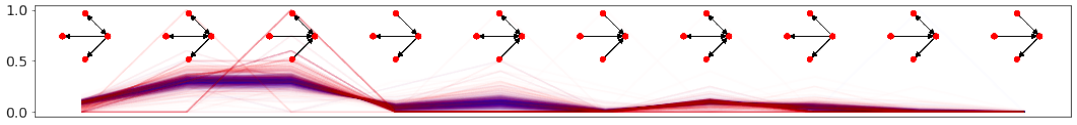

Figure 5.21: Frequency plot of all featured subgraphs represented as a ratio of disruptive (red) and normal (blue) users as an overlapping line plot. The profiles suggest that normal users are more consistent in comparison to disruptive users.

Each of the subgraph profiles presented in Figure 5.21 provide the basis for inferring the structure of social interactions. The frequency plot reveals how the third (one-in,

two-out), fifth (one-in-out, one-in, one-out) and seventh (all-out) subgraphs from Figure 5.20 stand out compared with normal user subgraphs. Normal users produce more consistent activity on the second (two-in, one-out), third (one-in, two-out) and fifth (one-in-out, one-in, one-out) subgraphs in Figure 5.21.

Prediction

Combining subgraph frequencies and temporal features together three binary classifiers are used to classify user behaviour as either non-disruptive "normal" (ND) or disruptive (D). In line with Research Question 4, binary logistic regression (BLR), a support vector machine and a random forest classifier (RFC) with $N = 100$ trees at a max depth of $D = 2$ are used. Different features are combined with the intention to understand if the accuracy of these models can be improved. These result can be found in Table 5.11 where we compute the precision (P), recall (R), F1-score (F1) and accuracy (A) using a train-test split ratio of 75:25.

		Temporal				**Subgraphs**				**Both**			
		P	**R**	**F1**	**A**	**P**	**R**	**F1**	**A**	**P**	**R**	**F1**	**A**
BLR	**ND**	0.51	0.87	0.64	0.55	0.66	0.98	0.79	0.7	0.64	0.95	0.76	0.7
	D	0.72	0.28	0.41		0.94	0.35	0.51		0.88	0.41	0.56	
SVM	**ND**	0.71	0.38	0.5	0.67	0.74	0.96	0.83	0.78	0.72	0.92	0.81	0.77
	D	0.63	0.87	0.73		0.91	0.56	0.7		0.88	0.61	0.72	
RFC	**ND**	0.82	0.75	0.78	0.81	0.79	0.87	0.83	0.8	0.85	0.89	0.87	0.86
	D	0.8	0.86	0.83		0.81	0.71	0.76		0.88	0.83	0.85	

Table 5.11: Prediction results of three leading classifiers for temporal features, subgraph features and the two combined improving the overall accuracy.

From the prediction results, it is clear that subgraph features perform better in compar-

ison to temporal features. The temporal features appear to lack support for prediction in the case of BLR and SVM however, RFC consistently outperforms BLR and SVM on every data set. Overall, the accuracy of the model is significantly improved when both subgraph counts and temporal features are combined.

5.5.5 Discussion

The results from this investigation address Research Questions 1 and 4, by demonstrating the utility of behavioural networks (in the form of egocentric subgraphs) and temporal features for classifying user behaviour. The temporal features collected provide early insights into the subtle differences between normal and disruptive activity. Simple attributes such as comment count and age provide initial indications whether behaviour is suspicious.

Counting the 4-node star subgraphs that are induced by users' communication provide a means to discover the relationship a user has with other users irrespective of the temporal domain. These results show distinct differences between the group of disruptive users and the "normal" users, noting that some subgraphs are much more likely to be present for disruptive users and vice versa.

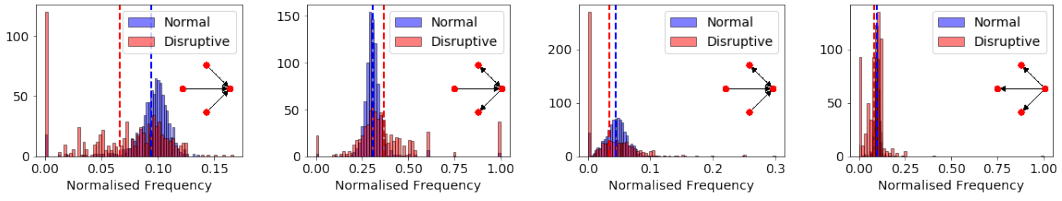

Figure 5.22: Collection of motifs discovered comparing disruptive against normal user activity displayed as frequency histograms shown with mean marked by dashed lines. These findings reaffirm the obseravtions of 5.21 that normal users are more consistent in activity compared with disruptive users.

In the first example, the subgraphs featured in Figure 5.22 indicate that normal users are more likely to receive a reply and form a two-way conversation. In addition to this, disruptive users vary much in the subgraphs and provide evidence that disruptive users

are likely to initiate a reply to a user and are less likely to receive a reply or two-way response.

Furthermore, disruptive users are less likely to preserve at least one symmetric edge, hence the subgraph (second from the left) in Figure 5.22 is significantly lacking in appearance. In contrast, normal users are more active and consistent across the nearly all subgraphs featured in Figure 5.22 where normal users are likely to receive at least one reply during the discussion potentially leading to a two-way conversion at some point.

The prediction analysis for user classification provides strong evidence for basic temporal statistics complementing subgraph frequencies. While classification performs reasonably well using each feature set in isolation, the combination of the two produces a significant improvement overall and help answer Research Question 4 as a result. The three classifiers used in this investigation serve as a basis to demonstrate the potential for behaviour classification and proof-of-concept, however alternative classifiers can be used, and may further improve performance.

5.5.6 Key Findings

As discussed earlier in Chapter 2, Reddit has experienced disruption as a result of events including Brexit (see Section 2.1.2) and the 2016 US Presidential Election (see Section 2.1.3). The methods and results featured as part of this investigation of the hypothesis indicate that the behaviour of users on Reddit, is sufficient for relatively accurate categorisation of disruptive users, independent of the content and semantics which, in turn, reaffirms that user-to-user networks (as well as temporal features) can be used as part of a framework for predicting disruptive behaviour (see Research Questions 2 and 4). This is based on extracting subgraph features from user reply networks to describe the fundamental features surrounding user-to-user interaction. As a result, reply networks are a valuable tool for analysing discussion threads between users which

can be used in multiple scenarios.

However, this investigation doesn't consider repeated discussions and the depth of two-way debates between users and could therefore be improved in response to Research Question 1. Future work which explores behavioural networks should involve considering weighted edges (or timestamped edges) for modelling repeated replies. In addition to this, with respect to Research Question 4, more-advanced classifiers (such as artificial neural networks) could be used to further improve overall performance and classification accuracy. Overall, while significant improvements could be made, the results obtained from this investigation sufficiently explores the hypothesis of this thesis and provides key insights towards understanding communication patterns of disruptive users.

5.6 Conclusions

In conclusion, this chapter demonstrates the utility of user-to-user networks for detecting disruptive behaviour in the form of two investigations to explore the hypothesis. The results reveal that Twitter and Reddit, while two fundamentally different platforms both align with the message data structure (see Chapter 3, Section 3.2.3) and provide strong evidence that user-to-user networks can be used to adequately capture disruptive behaviour. These answers Research Questions 2 and 3 by demonstrating that user-to-user networks have utility for representing and capturing alternative affordances across both Reddit and Twitter. Furthermore, networks built from reply-based interactions are common to both the studies and have potential for extracting behavioural signals especially when supplemented with temporal features, as seen in the second investigation using Reddit.

The analysis on Twitter used a combination of basic network-based metrics and local metrics (e.g. subgraphs) with the intent on using these as feature vectors for differentiating between controversial and non-controversial terms. The findings concluded that

generic global features such as degree, density, transitivity and reciprocity are sufficient for classifications tasks. As a result, this partially addresses Research Question 4 by demonstrating that it is possible to provide signalling on which the prediction of destructive behaviour can be made although through this is achieved though global features, not local features. Within this investigation the use of local subgraph-based metrics had a secondary role for describing how smaller interactions took place around individual users. These results revealed that subgraphs with the strongest eigenvector coefficients were centred around an individual user and are considered egocentric due to the way interaction are positioned around a central user.

Consequently, the findings of the investigation on Twitter justify the methodological structure of the investigation on Reddit whereby only egocentric subgraphs are used as feature vectors to determine normal behaving users from disruptive users according to their karma score. Reply-based interactions using just egocentric subgraphs are sufficient for classification which answers Research Question 4, although this could be further improved, with respect to accuracy, when combined with temporal features. Because of this, the results of the Reddit investigation informs Research Questions 1 and 4 by reaffirm the finding of the analysis on Twitter that simple metrics around individual users (e.g. in/out degree and egocentric networks) are adequate proxies for defining behavioural networks (Research Question 1) and for differentiating between disruptive and non-disruptive behaviour (Research Question 4).

Overall, this chapter investigates the hypothesis of this thesis by demonstrating the versatility of user-to-user networks for modelling and capturing behavioural signals for detecting disruption (Research Question 4). The implications of this chapter demonstrate that the user-to-user network respiration can be used in a wide variety of situations for defining behavioural networks (Research Question 1) among users (both collectively and individually) and can easily be applied to other type of affordances and the interactions they facilitate (Research Question 3). For this reason, it is important to acknowledge that this approach has implications to platforms beyond Twitter and Reddit. Furthermore, it

is reasonable to conclude that this network representation applies to social media more generally due to the way these platforms are designed to facilitate interactions among other users, thus reaffirming the idea that this can be used as part of a wider framework for detecting disruptive (Research Question 2).

Chapter 6

User Association Networks

6.1 Introduction

In the previous chapter (see Chapter 5) both Twitter and Reddit were used to investigate the hypothesis by demonstrating the versatility of user-to-user networks to identify the presence of disruptive behaviour such as trolling (see Chapter 2, Sections 2.2.2) and misinformation.

Overall, Chapter 5 addressed each of the research questions used to investigate the hypothesis. Firstly, Research Question 1 was addressed by defining behavioural networks from social media activity in the form of a user-to-user network. Secondly, Research Question 2 was addressed by demonstrating that user-to-user networks can be used as part of a framework for capturing disruptive behaviour. Thirdly, Research Question 3 was addressed by modelling multiple interaction and affordances based upon activity derived from Twitter and Reddit. Finally, Research Question 4 was addressed by demonstrating the potential for classification algorithms to detect disruptive activity using a combination of local and global network-based features.

This chapter considers the role of bipartite networks to model these relationships for observing similar disruptive behaviour such as misinformation and the formation of echo chambers [76, 329, 376, 43, 78, 211, 365] - behaviour which takes place at scale across multiple communities (see Chapter 2, Sections 2.2.4).

As mentioned previously in Chapter 3, this chapter describes these bipartite networks as

User Association networks which are designed to focus on the association between a user and community by means of a direct interaction (e.g. a user posts a submission to a community). In doing so, this network representation is used to model the many-to-many relationship between multiple users and multiple communities - the association. This chapter examines this behaviour, and therefore, the utility of user association networks by using Reddit to investigate the hypothesis (as shown in Section 6.4).

This chapter seeks to address Research Questions 1 and 2 by utilising bipartite behavioural networks derived from social media activity to model associations between users and communities which can be used as part of a framework for detecting disruptive activity. Additionally, this chapter also explores the hypothesis by considering the role of both local and global features for examining the extent to which prediction can be performed (see Research Question 4).

As mentioned in Chapter 3, Reddit (the focus of this chapter) aligns with the community data structure. This is demonstrated in Table 6.1 which is relative to the work of this thesis by considering the role of the community data structure and user association network.

		Data Structure			
		Community	Message	Collaborative	Feed
Platform	Wikipedia	N/A	N/A	*See 3.4.1*	N/A
	Reddit	***See 3.4.2***	*See 3.4.2*	N/A	*See 3.4.2*
	Twitter	N/A	*See 3.4.3*	N/A	*See 3.4.3*

Table 6.1: Relationship between platforms of interest and all data structures with the appropriate cells concerning the work of Chapter 6 highlighted in bold.

As described in Chapter 3, Section 3.6, this chapter utilises the community data structure to model interactions as bipartite relations through the user association network. By considering the community data structure and the user association network, these can be used to define behavioural networks (see Research Question 1) and provide representation for diverse affordances (see Research Question 3).

User association networks, and indeed all bipartite networks, are designed to model

connections between two sets of nodes which are mutually exclusive and have different roles. Within the literature, these networks have also been known as *affiliation network* and have been used to model different types of relationships [387]. For example, an affiliation network could be used to represent football players and clubs, with an edge between the two if a player played for a club [387]. Within social network analysis more broadly, bipartite networks have been used for community detection [257], understanding and gaining trust [292], measuring influential users [416], predicting links [41] and making recommendations [411].

In the context of this thesis, the notion of an affiliation network is almost identical to that of a user association network, however a user association network focuses on user activity on social media for capturing disruptive behaviour. Within the supporting literature, the use of bipartite network structures have been used on Wikipedia to determine high-quality based upon local substructures embedded within a network modelling the relationship between a user and article [398]. With respect to disruption, there is a clear gap in the literature for using bipartite network structures to detect disruption. The research produced in this chapter (see Section 6.4) demonstrates that user association networks can be used for identifying communities where for misinformation is likely to develop.

The rationale of user association networks is to distil the relationship between a user and a corresponding topic, community or other item in a discreet form. In doing so, this representation considers two features which are fundamental for understanding user behaviour: $1)$ a user's variation of interests and $2)$ the similarity with other users. As a result, association networks consider the local neighbourhood of a user (the ego) and constructs edges recursively between it's linked associations and their users. Furthermore, the use of induced sub-structures are used to manage and navigate through the complexity of many-to-many relationships using a smaller sample size.

As recalled in Chapter 3, Section 3.3.2, Reddit, the focus of this chapter (see Section 6.4), is an interesting platform to consider, since it allows self-defined communities

to establish themselves, giving a unique basis for the analysis of online interactions. By design, Reddit is considered a "community-driven" platform [342] due to the way the platform encourages users to orientate themselves around communities known as subreddits. Users can submit posts in the form of links and text submission to a subreddit with the intention that like-minded people will engage. As a result, the relationship between users and subreddit can be modelled using a user association network, where an edge represents the submission of a post. An example can be found in Figure 6.1.

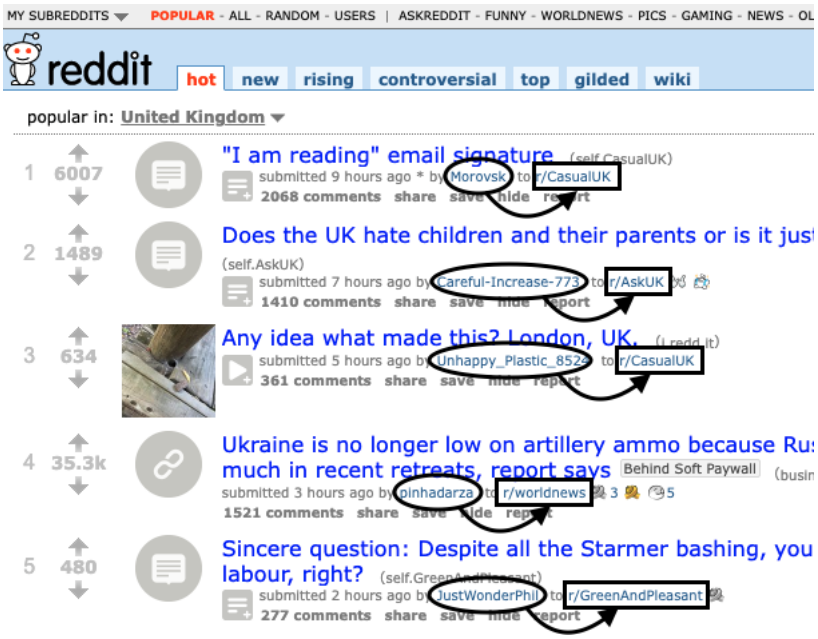

Figure 6.1: Example taken from the homepage of Reddit demonstrating how are users (ovals) are associated with subreddits (boxes) based upon posting behaviour. This demonstrates the basis for constructing a user association network based on Reddit activity.

6.1.1 Contributions

This chapter contributes to this thesis by investigating the hypothesis concerning the role of user association networks in detecting disruptive activity on social media.

To begin, this chapter demonstrates how disruptive behaviour such as the spread of misinformation (see Chapter 2 for more) can be detected through simple bipartite rela-

tionships using their underlying substructures by counting bipartite subgraphs known as "graphlets" which, in turn, addresses Research Question 4. The methodology introduced in this chapter helps address Research Questions 1 and 2 by providing a novel approach for building a detailed representation of community behaviour considering the activity of not just one user but of multiple users by considering their subreddits of interest too in a recursive manner as part of a framework of alternative network presentations. This makes it possible to model the spread of interactions at scale across multiple communities and to extract detailed signals such as the number of mutual connections (for example, users who post in the same set of subreddits) and distribution of interests which, in doing so, has the potential to capture diverse affordances and different types of interaction (see Research Question 3).

6.1.2 Network Construction

As demonstrated in Chapter 3 Section 3.5.3, the basis for user association networks are inherited from a bipartite network using mutually exclusive sets of users and communities. In response to Research Question 1, a user association network is formally defined by $G = (V_1 \cup V_2, E)$ where V_1 represents a set of users, and V_2 represents a set of communities where V_1 and V_2 represent a bipartite set of nodes (i.e., $V_1 \cap V_2 = \varnothing$). An edge $e_i = (v_i, v_j)$ where $e_i \in E$ exists if and only if user $v_i \in V_1$ has an association with $v_j \in V_2$. For example, a user v_i (User A) who posts in a community v_j (Community Z) forms an association with the community such that an edge $(v_i, v_j) \in E$ forms the connection $A \rightarrow Z$ or "*A posts in Z*".

6.2 Motivation

The investigation introduced in this chapter is motivated by the desire to detect communities which are likely to promote misinformation based upon user behaviour according to the structure of a user association network. As a platform, Reddit has been the

target of misinformation in recent years in the form of political misinformation [76], health misinformation [329, 224], and general conspiracy theories and echo chambers [376] (see Chapter 2, Section 2.2). Of these instances, misinformation emerges though a combination of targeted key narratives, little scientific consultation and algorithmically generated talking points. As identified in Chapter 2, Section 2.2.4, the community-centric design of Reddit can lead to the formation of echo chambers which, in turn, increase the possibility of misinformation to emerge.

The spreading of misinformation is a significant challenge and is frequently addressed though the use of NLP text analysis [329, 355, 255, 295]. With respect to bipartite networks (the basis of this chapter), research has shown how bipartite network configuration can aid the discovery of misinformation through a combination of community detection, unique connectivity patterns and motif analysis [43, 78, 211, 365]. This is achieved by modelling a user's connection with posts, pages, organisations, locations and news sources. Furthermore, this approach can be extended to focus on other networks which don't involve users by focusing on certain narratives, key terms and news entities [44, 107].

Bipartite network structures can be used to help better understand user behaviour, which can help aid the formation of language agonist solutions which can be deployed at scale (see Research Question 1). Very few studies consider using raw bipartite network features exclusively without considering additional features such as timestamps, text or other metadata which may not be available or consistent. As indicated through the supporting literature, bipartite network representations are a well-established component for community detection which is ideal for platforms such as Reddit, the central focus of this chapter. Being able to identify the characteristics of potential misinformation communities is highly valuable, particularly if this can be achieved with low overhead in terms of computational resources, in a language-agnostic manner and without the need for complex semantic analysis (see Research Question 4), and this aspect motivates the investigation of this thesis.

6.3 Approach

		Data Structure			
		Community	Message	Collaborative	Feed
Network	Transitional	-	-	See 4.4 (Wikipedia)	See 4.5 (Reddit)
	User-to-user	-	See 5.4, 5.5 (Twitter, Reddit)	-	-
	User Association	**See 6.4 (Reddit)**	-	-	-

Table 6.2: A replica of Table 3.20 featured in Chapter 3 outlining the investigations of this thesis (with respect to data structures and network representations) with the appropriate cells highlighted in bold which refers to the problem space which Chapter 6 seeks to investigate.

As presented in Table 6.2, this chapter, relative to this thesis, investigates the role of user association networks for detecting disruptive activity by analysing the spread of misinformation across communities. Reddit, a platform which aligns with the community data structure (see Chapter 3, Section 3.4.2), is used as the basis for our analysis and is also used as part of a wider framework of alternative network representations for defining behaviour on social media (see Research Questions 1 and 2). As described in Chapter 3, Section 3.5.3, user association networks and the community data structure are used to represent a relation (posting) between a user and a community (subreddit, in the case of this chapter).

To demonstrate the utility of user association networks, Section 6.4 of this chapter provides a real-world exemplar of how these networks can be used to understand disruptive behaviour in the form of an investigation of the hypothesis for detecting the potential for misinformation to emerge on Reddit regarding COVID-19 vaccinations (see Chapter 2, Section 2.1.4). Given the community-orientated nature of the platform, Reddit serves as an ideal candidate to study due to the presence of the community data structure (see Chapter 3, Section 3.2.2).

This investigation uses a combination of both local and global network features as feature vectors for classification. In doing so, this investigation helps answer Research Question

4 by using user association networks for modelling cross-community interactions between users and subreddits with the intention to discover signals which lead to the detection of misinformation. Furthermore, the use of user association networks can be used to classify additional types of subreddit as well as those that are likely to promote misinformation. An example of a user association network can be found in Figure 6.2 and Table 6.3 for edge types.

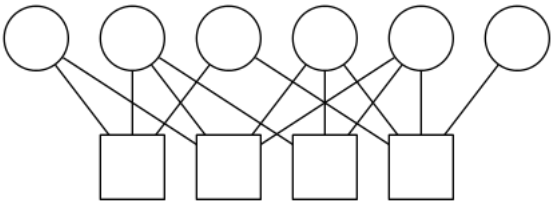

Figure 6.2: An example of a hypothetical user association network of users (circles) and communities (squares) based upon a bipartite network. In this example, an edge can be used to represent different interactions with the community (e.g. posting a link).

Application	**Subreddit Association**
Source Node	User
Edge (undirected)	*posted to*
Target Node	Community

Table 6.3: Edge list definition for a subreddit association network.

6.4 Detection of Misinformation on Reddit Using User Association Networks

As mentioned in Chapter 2, misinformation is a major cause for concern with potentially dangerous ramifications for social processes, including the stability of democracy [37, 322]. As discussed previously (see Section 2.1.4), the COVID-19 pandemic has been the subject of various "fake news" stories and conspiracies resulting in an "infodemic" as described by the WHO (World Health Organisation)[1] [298, 8, 294].

[1] `https://www.who.int/news-room/events/detail/2020/06/30/default-calendar/1st-who-infodemiology-conference`

This issue has become a serious threat to public health and has triggered multiple public responses including the destruction of 5G cellular towers in the UK[2] [8] and the proposition to reject potential vaccinations[3]. As part of this, the informality of online social media is well-suited to propagation of misinformation, which has been an unforeseen consequence of the technology's role in liberating global participation [95].

In this Chapter, Research Question 4 is addressed by considering both graphlet frequencies and global metrics are analysed to assess their utility in distinguishing between sets of subreddits potentially associated with misinformation and sets of sample subreddits that are not associated. In line with Research Question 4, various machine learning models are used to determine the predictive power of graphlet and global features. This gives a basis to assess the role of local features, including substructures, in capturing online behaviours aligned to potential misinformation.

Furthermore, the network-based methodology introduced uses a language-agnostic approach which provides important opportunities to support automation in the detection of misinformation in online communities. This investigation extends methods that were successful in classifying controversy in Wikipedia articles [22, 398] and further contributes to characterising potentially disruptive groups [398, 164]. As of this writing, little research has addressed communities of misinformation on COVID-19, particularly with reference to Reddit, making the current work both timely and relevant.

6.4.1 Associated Literature

The relative ease with which misinformation can be produced and become disruptive has motivated recent investigation into this phenomenon. From a psychological viewpoint, it appears that there are individual differences in how misinformation becomes potentially endorsed by individuals [331, 259]. This acceptance gives a basis for misinformation

[2] https://www.bbc.co.uk/news/technology-52281315

[3] https://www.telegraph.co.uk/global-health/science-and-disease/one-third-uk-may-not-get-coronavirus-vaccine-one-developed-new/

to become potentially promoted by like-minded others, leading to a compound effect where an informal group endorses particular content with a reinforcement across its participants.

For example, one subreddit in particular, r/Wuhan_flu, a subreddit which actively promoted the use of free speech gained a lot of attention with its anti-censorship agenda. Reddit took action and placed the subreddit in "quarantine" suggesting that it "*may contain misinformation or hoax content*"[4] therefore requiring users to "opt-in" to the community to view its content.

The effects of misinformation have taken place in various different contexts, with politics being particularly susceptible, as seen in the 2016 US presidential election [14, 16] (see Section 2.1.3). In previous events, disinformation concerning the funding of the UK's National Health Service (NHS) was circulated and brought to the attention of various political leaders during the 2019 UK General Election, which originated from Russian actors on Reddit[5].

However, once misinformation is embedded, echo chambers (see Section 2.2.4) are known to take hold and to support engagement of misinformation, using weak ties [371, 389] and lack of effective moderation [285] alongside "soft facts". These occur as a result of users sharing potentially misleading content without knowing the entire facts of an event [193, 119, 194].

The impact of misinformation surrounding COVID-19 has been observed on multiple platforms including the microblogging site Twitter [343, 221, 335, 358, 403, 89]. Evidence of social media analysis suggests that individuals fail to discern between truth and falsehood, prompting the argument that health information shared on social media should be regulated [306, 105, 372, 403].

As mentioned in Chapter 3, Section 3.3.2, users are encouraged to join "subreddits"

[4] `https://www.reddit.com/r/Wuhan_Flu`

[5] `https://www.reddit.com/r/redditsecurity/comments/e74nml/suspected_campaign_from_russia_on_reddit/`

which serve as individual communities dedicated to a topic or theme where users can share and comment on posts submitted to the community. This provides the freedom to connect and reinforce the views of others. In particular, the Reddit platform has played a role in the spreading of hoaxes originating from Wikipedia [225] as well as sharing misinformation across multiple platforms [406]. With respect to misinformation around COVID-19 on Reddit, research has addressed the difference in narrative and language within Reddit communities using NLP [409, 348] as well as the location [156] to assess the geographical influence.

To summarise, social media platforms provides innovative mechanisms allowing users to promote and share news and stores of current events which produce diverse affordances as a result (see Research Question 3). As a consequence, they produce conditions in which misinformation can develop at scale such that large audiences are potentially subject to misleading information . This is especially important considering the ease in which information can propagate through user interactions such as sharing and cross-posting across different communities.

6.4.2 Methods

As mentioned previously, both global and local features of network-based representations are examined to determine the predictive utility of user association networks (see Research Question 4). As described in Section 6.1.2, this is achieved through user association networks that link users with subreddits to which they contribute, which we term as *subreddit association networks*.

As defined earlier, Chapter 3, Section 3.5.3 demonstrates how bipartite networks can be used to capture user variations in interests and similarity with other users. In the context of this investigation, subreddit association networks capture two areas of interest: $1)$ a user's diversity in posting to different subreddits and $2)$ the overlap between users in posting to the same subreddit(s). An example of a hypothetical subreddit association

network is presented in Figure 6.3.

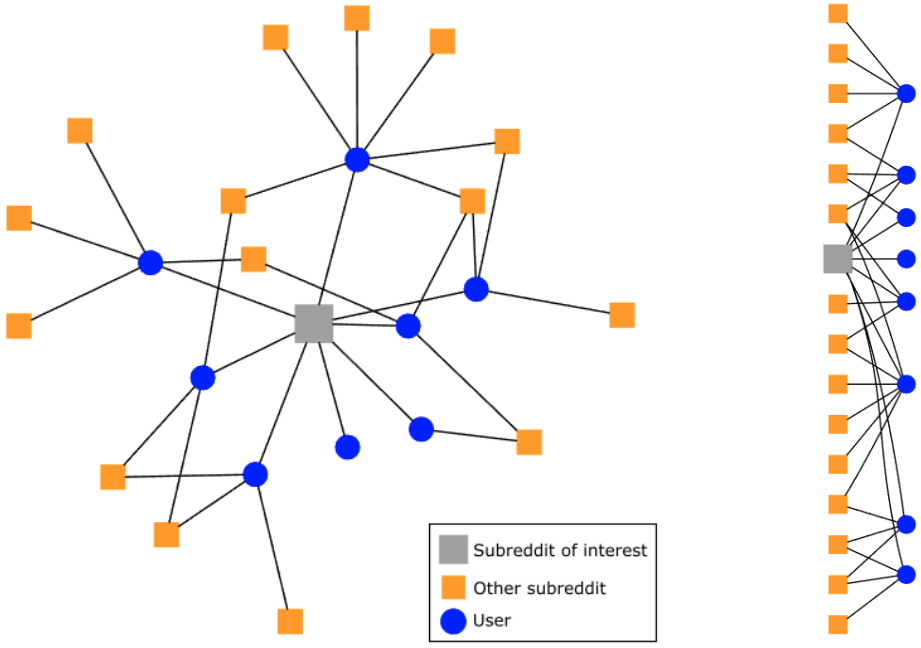

Figure 6.3: Example of a randomly-generated bipartite subreddit association network (full network, left, and bipartite arrangement, right) where the subreddit of interest is marked in grey, and it's surrounding users as blue circles. Other subreddits are represented as orange squares.

For a corpus of subreddits, a *subreddit association network* (also known as SAN) can be defined as a bipartite graph $G = (V_1 \cup V_2, E)$ where V_1 represents a set of Reddit users, and V_2 represents the set of subreddits to which the users in V_1 have collectively posted. V_1 and V_2 represent a bipartite set of nodes (i.e., $V_1 \cap V_2 = \varnothing$), and there exists an edge $(i, j) \in E$ if and only if a user $v_i \in V_1$ has posted in subreddit $v_j \in V_2$. The approach featured in this investigation is similar to work performed by Cheng et al. and Caldarelli et al. [89, 78] however, the network analysis is not limited exclusively to centrality and degree-based metrics as this investigation is extended to include graphlet analysis.

Local Network Features

As discussed earlier (see Section 6.1.1), classifying SANs is based on counting the frequency of graphlets that are induced within its structure. Defining graphlets to include

non-trivial induced substructures with up to six nodes provides a reasonable trade-off between the combinatorial complexity of counting (see [102] concerning the graphlet isomorphism problem and Section 3.7) and the presence of useful features for analysis. In the network science literature, the possible induced subgraphs of a fixed size are typically referred to as graphlets [65, 189, 198]. The same terminology is extended here to denote all connected bipartite graphs with 3 to 6 nodes, as presented in Figure 6.4, resulting in 43 possible alternatives.

The frequency of graphlets present in a given subreddit association network G is denoted by vector V_G where:

$$V_G = (v_1, v_2, \ldots, v_{43}) \tag{6.1}$$

and where v_i represents the frequency of the i^{th} possible graphlet from Figure 6.4. To enable comparison of networks of different size, we normalise V_G according to:

$$V_G = \frac{1}{\Sigma_{j=1}^{43} v_j}(v_1, v_2, \ldots, v_{43}). \tag{6.2}$$

Vector V_G gives a basis to consider the relative under or over-representation of induced graphlets, in comparison to other subreddit association networks. This is similar to network-motif analysis (see Section 3.7) approach for complex networks [263, 264], and gives a basis to compare networks based on their latent structural characteristics and helps address Research Question 4 by considering the role of local substructures for performing classification. The relatively high dimensional space associated with $V(G)$ means that dimensionality reduction is a useful tool to provide further insights into the relationships between different association networks. Therefore, V_G is analysed as derived from different subreddits, to establish the extent of similarity between different classes of association network.

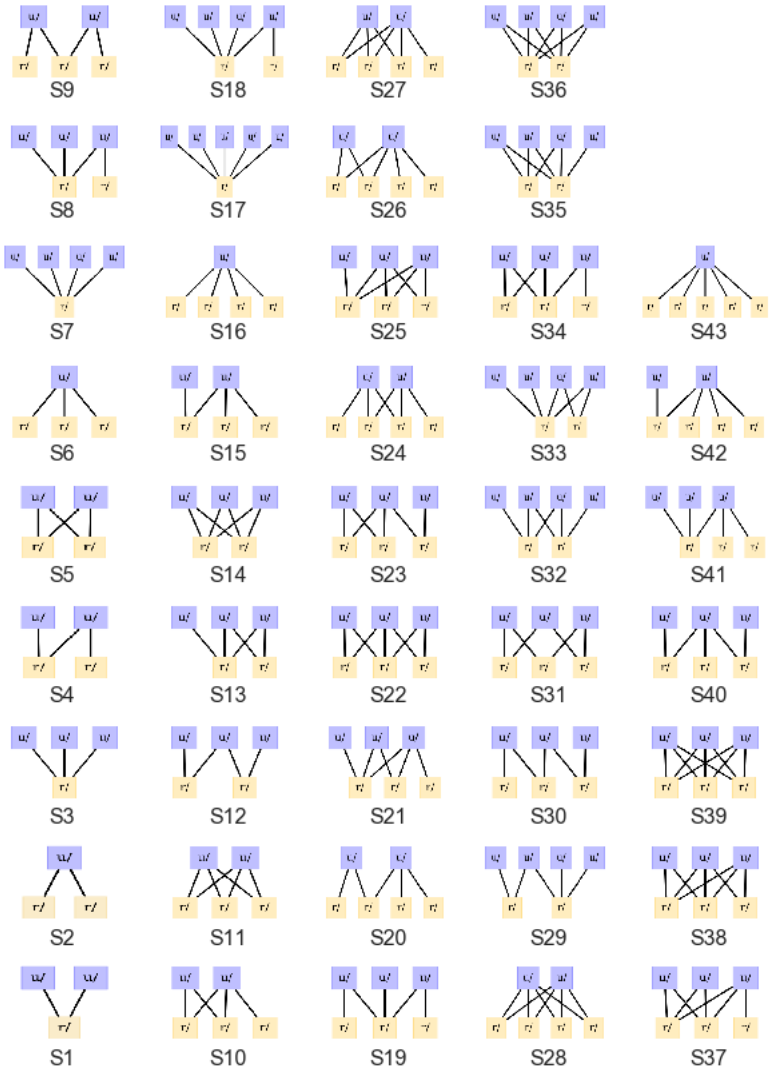

Figure 6.4: Collection of all 43 induced bipartite graphlets featuring graphlet sizes from 3 to 6 for every possible combination of nodes and edges. User nodes are labelled as blue and subreddits as yellow.

Global Network Features

Alongside the induced subgraphs represented through association networks, to address Research Question 4, network-based metrics are also considered which provide an understanding of how users and subreddits behave collectively in comparison to local graphlet features. These metrics include: subreddit and user degree, closeness centrality, clustering coefficient (i.e., local density between neighbouring connections), Latapy

clustering (to determine heavily clustered interactions between user and subreddits) and Robins-Alexander clustering [321] to determine the clustered interactions through an aggregation of cycles and paths. These metrics provide further ways in which user association networks can be classified and assessed.

Data Collection

To build a subreddit association network, the Reddit API was used to collect the data. Overall, a total of 257 subreddits were sampled. This resulting in a corpus of data consistent with the scale of other misinformation studies on COVID-19 e.g. [409]. For each subreddit, posts were sorted by date (most recent first) and a list of users who created the posts was extracted. Including all subreddits, the data spans between September 2017 and May 2020.

Subreddits were manually classified, aligned to their *potential for misinformation* (PFM) concerning COVID-19. There is no definitive way to achieve this meaning that the following criteria were applied to identify such subreddits. Either:

1. The subreddit generally had very few moderators (users who are responsible for maintaining a subreddit community) and applied little or no moderation given the size and age of the subreddit. This is relevant because it allows more freedom for misinformation to go unchecked.

2. The subreddit description used terms such as “anti-censorship” or “freedom of speech” (FOS) in the subreddit description with little moderator involvement. This leaves greater opportunity for misinformation to be established.

3. The subreddit had been placed in “quarantine” by the Reddit administrators for containing potentially misleading or harmful content for the community. This is relevant due to the potential detection of misinformation.

Additionally, the popular COVID-19 subreddits (e.g. r/Coronavirus, r/CoronavirusUK, etc) were included as they are considered subreddits with the potential to contain misinformation. These are highly topic-relevant subreddits that might be attractive to agents who are keen to express misinformation. In total, application of these criteria resulted in 27 subreddits being selected as having potential for misinformation. These are referred to the *PFM subreddits*. Appendix F.1 provides a list of the PFM subreddits used in this investigation. Note that alternative criteria for selection of the PFM subreddits could equally be applied.

To provide a basis for comparison of PFM subreddits, three other sets of subreddits were introduced such that benchmarking can be performed against alternative forms of user interaction with this social media platform. This is done with the objective of comparing against alternative subreddits to assess predictive utility. Furthermore, the aim of this investigation is to discover how our approach can represent subreddits of different taxonomies (e.g. Q&A compared with discussion). The benchmark subreddits are defined as follows.

- **PFM**: a total of 27 subreddits relating to COVID-19 which may contain misinformation.
- **Ask**: a sample of 30 Ask Q&A-based subreddits that involve interactions within in a highly moderated environment. This allows us to compare with PFM as posts made to Ask subreddits undergo strict moderation due to restricted posting rules meaning that they are unlikely to contain misinformation, contrary to PFM subreddits.
- **New**: a sample of 100 random subreddits created in the year 2020, covering the time period relevant to the creation of most PFM subreddits. This enables subreddit age to be controlled for in subsequent analysis.
- **Random**: a sample of 100 random subreddits without any constraints for subreddit age. This serves as a random baseline to include the diversity of Reddit content

in general.

For each of the users active in any of the subreddits in any of the four datasets, a list of all other subreddits that they also submitted to over the time period was produced. These were then aggregated to generate the bipartite user association network across the subreddits in the above samples, as described in Section 6.4.2. Across all the subreddits selected, a total of $7,876,064$ posts were processed across $96,634$ users.

Using non-network metrics, metadata (such as age and subscribers) were used to demonstrate that these features aren't necessarily the strongest indicators for clustering as observed in Figure 6.5. By comparison, network features provide better spatial relevance. Additionally, the availability of such data is limited and may not always be consistent meaning that the exclusive use of metadata does not serve as a reliable proxy for classification. Furthermore, this justifies the need for considering network-based metrics.

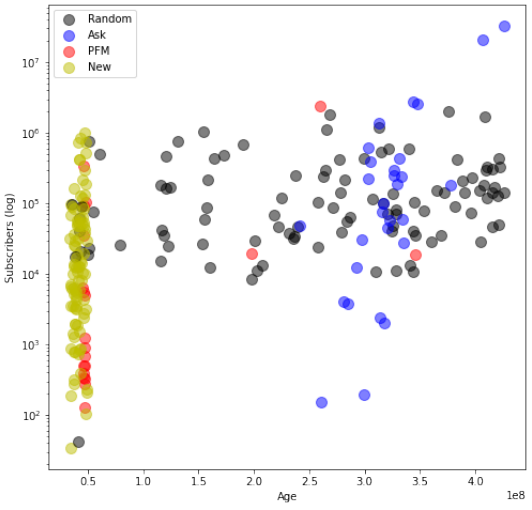

Figure 6.5: Using age (x-axis) and subscriber counts (y-axis) shows little clustering potential compared with network-based features. New subreddits are marked in yellow, Ask in blue, PFM in red and Random in black.

The results in Figure 6.6 reveal how PFM subreddits show lower maximum and average in subreddit degree meaning fewer users engage with these subreddits as compared with that of the Ask subreddit communities. Furthermore, it can be observed that PFM

subreddits also have a higher but varied average user degree and produce less clustering, contrary to Ask subreddits.

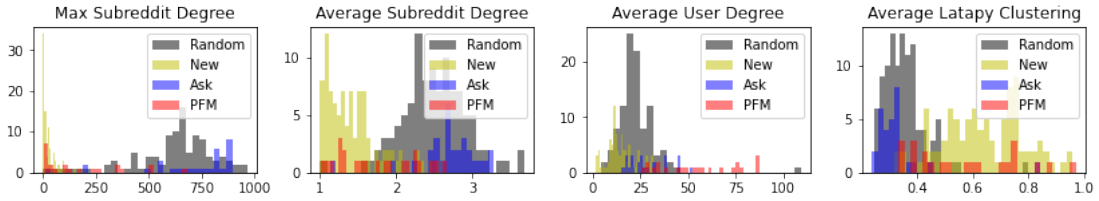

Figure 6.6: Subreddit degree distributions, user degree distributions and average Latapy clustering distributions for the Ask, PFM, new and random association networks. Each of these global metrics reveal distinct distributions for each of the four classes.

In Figure 6.7, the subreddit and user degrees as a normalised ratio of the maximum degree featured in each network are presented. For Ask subreddits, the maximum degree for user and subreddits nodes are fairly balanced with a partial swing towards having a slightly larger subreddit degree maximum. Furthermore, the PFM subreddits are heavily skewed towards a higher user degree and lower max subreddit degree.

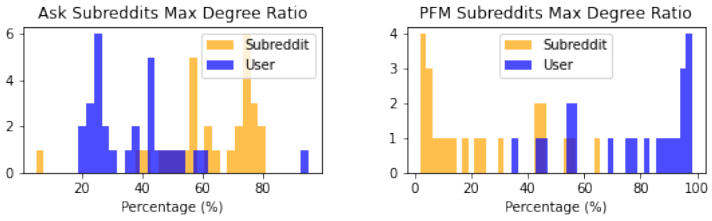

Figure 6.7: The maximum degree ratios for the Ask and PFM association networks reveal how PFM subreddits have a higher proportion of high degree user nodes compared to that of Ask subreddits.

6.4.3 Experimentation and Results

Using the data collected in Section 6.4.2, the profile of graphlets induced by the PFM, Ask, new and random sets of subreddits (Section 6.4.3) are examined. Principal component analysis (PCA) is then applied to examine the extent to which these different classes of subreddit can be distinguished under dimensionality reduction. This reveals

the contribution made by particular graphlets in support of the resulting new dimensions, indicating the dominance of particular graphlets, which, in turn, help address Research Question 4.

For comparison purposes, in Section 6.4.3 similar analysis is carried out but with global network metrics to compare prediction performance with local features (Research Question 4). Finally, Section 6.4.3, explores the extent to which it is possible to predict the classification of subreddits based on the dominant features of the PCA dimensions identified in Sections 6.4.3 and 6.4.3.

Association Network Profiling Through Graphlet Analysis

All induced bipartite graphlets (see Figure 6.4) are enumerated to produce normalised vectors of graphlet frequency (see Equation (6.2)) for all networks. The results in Figure 6.8 show some general similarities between different classes of subreddit. In particular, the `S17` and `S43` graphlets are dominant across all four sets of subreddit. A high variation across all graphlets for new subreddits can also be observed. Smaller graphlets such as `S1` and `S2` make more of an appearance in new subreddits. The graphlet profiles also reflect the sparse nature of the subreddit association networks, with high degree graphlets (both user and subreddit) being absent. However, despite similarities, at a more granular level significant differences are evident between the graphlet profiles for the different classes. This is apparent when principal component analysis is applied, which reduces the 43-dimensional feature vector to two principal components as shown in Figure 6.9.

The results in Figure 6.9 show clear differentiation between the Ask and PFM subreddits. Although both PFM and new subreddits were created around a similar time, it can be observed that their positioning in the scatter plot remains distinct by comparison which reaffirms that factors in addition to age distinguish these subreddits. The Ask subreddits exhibit clustering while the PFM subreddits generally exhibit higher values (greater than zero) against the second principal component.

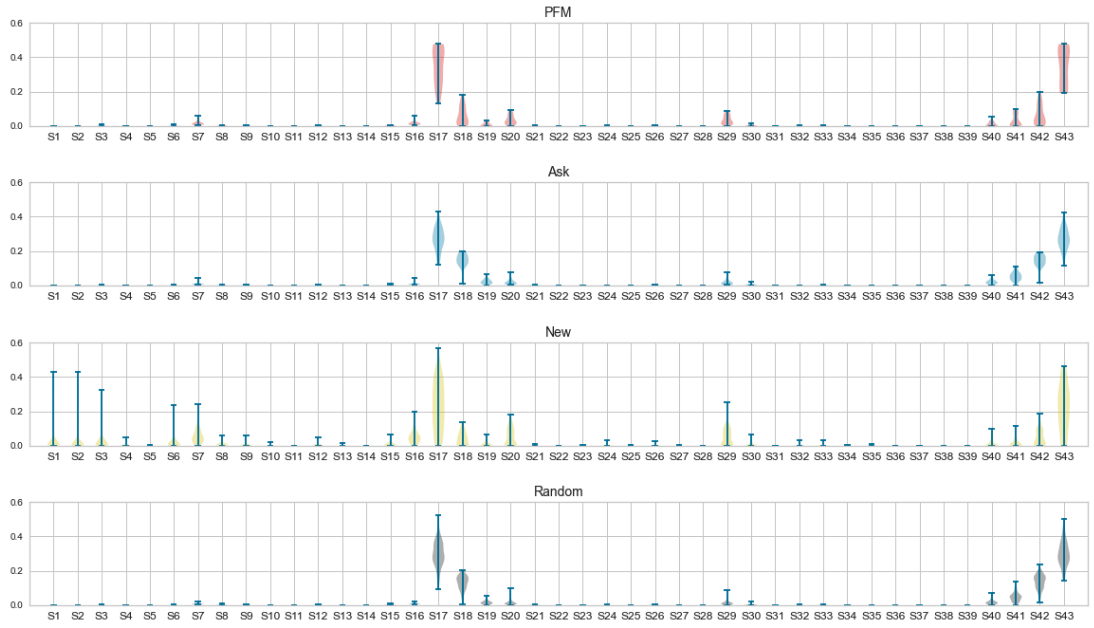

Figure 6.8: Normalised frequency as a violin plot of all 43 induced bipartite graphlets with PFM subreddits (first), Ask subreddits (second), New subreddits (third) and Random subreddits (fourth).

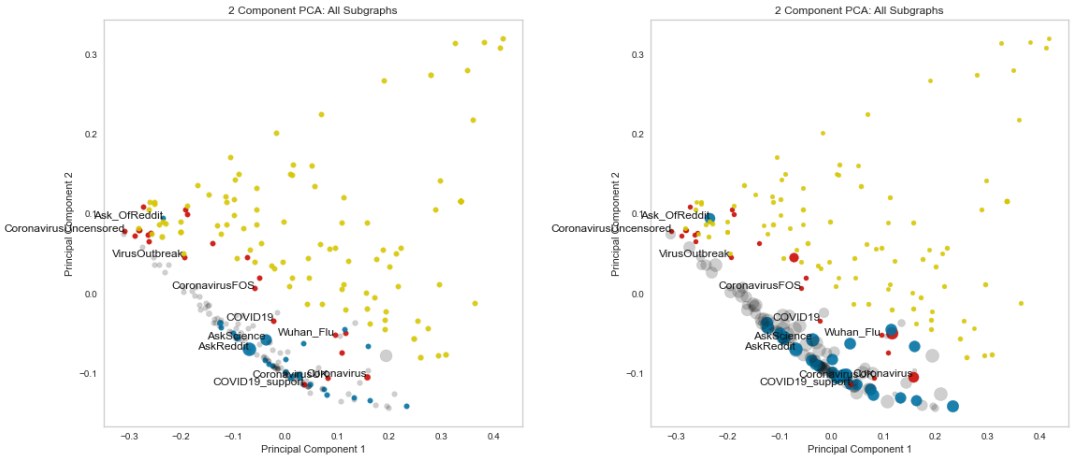

Figure 6.9: Scatter plot of two-dimensional PCA of graphlets counts producing distinct clusters with a few significant subreddits labelled. Ask subreddits are marked in blue, new subreddits in yellow, PFM subreddits in red and random subreddits in black. Nodes are sized according to (left) subscriber count (largest as most subscribed) and (right) age (largest as oldest).

More generally, these results indicate that graphlets can distinguish alternative classes of subreddits that appear similar in face value. Furthermore, it can be noted that principal component analysis positions the official coronavirus subreddits (r/CoronavirusUK and

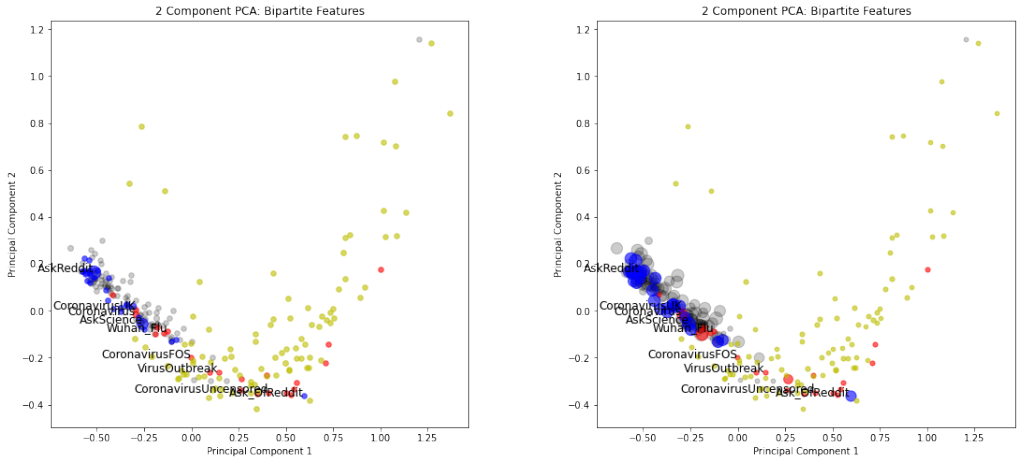

Figure 6.10: Scatter plot of two-dimensional PCA using only graph-based metrics as used in Section 6.4.2. Ask-subreddits are marked in blue, new subreddits in yellow, PFM subreddits in red and random subreddits in black. Nodes are sized according to (left) subscriber count (largest as most subscribed) and (right) age (largest as oldest).

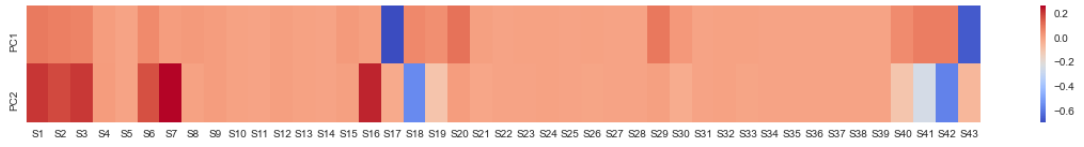

Figure 6.11: Extracting the PCA eigenvectors reveals specific graphlets which contribute to the spatial positing of subreddits within the two-dimensional space as shown in Figure 6.9.

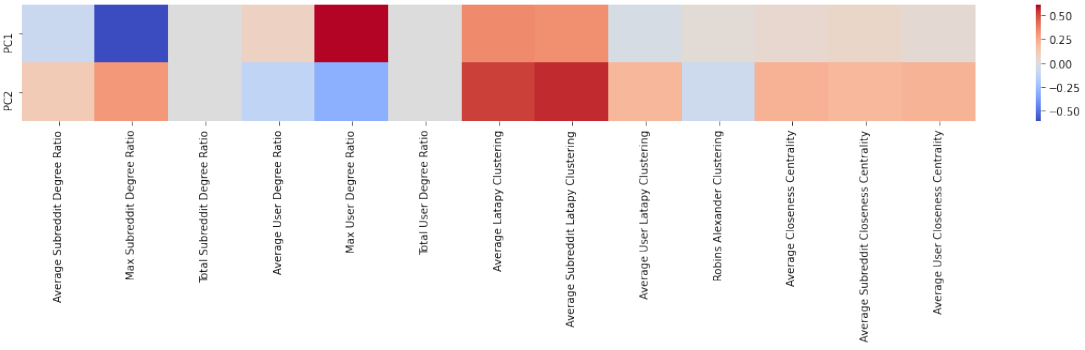

Figure 6.12: Extracting the PCA eigenvectors reveals specific global features which contribute to the spatial positing of subreddits within the two-dimensional space as shown in Figure 6.10.

r/Coronavirus) away from clusters of PFM subreddits. Subreddits with poor moderation such as r/CoronavirusUncensored, r/VirusOutbreak and r/CoronavirusFOS, are clearly

distinguished from Ask subreddits.

Taking into account data point sizes, Figure 6.9 indicates that subreddit subscriber count and age don't necessarily provide a strong indication of reliability or maturity compared with some subreddits which are much older and well established. The scatter plots reveal how younger subreddits share similar structures to that of well-established subreddits (such as the Ask communities). Furthermore, it is possible for older subreddits to align with the suspicious PFM subreddits cluster.

Figure 6.11 presents the eigenvalues for each graphlet with respect to their influence within each principal component. The first principal component is primarily characterised by the strong presence of graphlets `S17` and `S43` which describe a one-to-many (and vice versa) relationship. In other words, high degree from users to subreddits and high degree from subreddits to users are influential. Graphlets `S18`, `S20`, `S29` and `S42` are also highlighted, which suggests a partial overlap and mutual ties could contribute to distinguishing these alternative sets of subreddits.

Global Features of Association Networks

Using the bipartite metrics presented in Section 6.4.2 this section follows a similar approach to Section 6.4.3. In this section, user association networks are characterised using global metrics and apply PCA to create a reduced dimension space. The results provided in Figure 6.10 demonstrate similar clustering behaviour to that of the local features however differentiation between the spacing and clustering of particular subreddit groups is less pronounced. For example, the selected principal components provide little improvement in distinguishing between Ask and PFM subreddits. Additionally, greater spread is seen in the resulting dimensions, which is driven mainly by new subreddits.

Extracting the PCA coefficients (see Figure 6.12) demonstrates that the first principal component is heavily influenced by maximum degree ratios for both subreddit and user nodes. The second component is mainly positively influenced by Latapy clustering and

negatively influenced by maximum user degree ratio.

Predicting Class of Subreddit

To address Research Question 4, this section uses various prediction tasks to classify PFM, Ask and New subreddits respectively. Consistent with other approaches (e.g., [401]) and Research Question 4, this investigation uses binary logistic regression (BLR), support vector machine (SVM) and a random forest classifier (RFC) applied with 10-fold cross-validation. A complete list of prediction results can be found in Appendix F.1.

Local Features Prediction models are trained using normalised feature vectors of graphlet counts (as seen in Figure 6.8). Due to the imbalance between each of the subreddit groups, random under sampling is performed over $N = 100$ trials where the distribution, mean and standard deviation of various prediction metrics are reported [178]. The prediction metrics are reported in the form of violin plots (See Figures 6.13, 6.14 and 6.15) to capture the distribution of accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV).

Each subreddit group (PFM, Ask and New) are compared to a random baseline for benchmarking purposes. Pairwise comparison between each set of subreddits in isolation is also performed. This includes PFM vs Ask, PFM vs New and Ask vs New. As a result, this determines how robust prediction is between separate groups, in comparison to others without the need of a random baseline. This is used to understand how well classification of a particular set performs in isolation.

The classification results provided in Figures 6.13, 6.14 and 6.15 demonstrate that a RFC consistently outperforms BLR and SVM by comparison. This is reflected in the accuracy metrics as the distribution of values for RFC are much higher than those of BLR and SVM with an average accuracy of $P = 0.74$ for PFM, $P = 0.77$ for Ask and

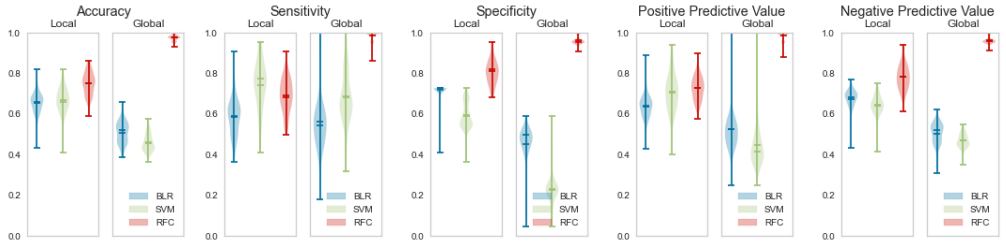

Figure 6.13: Classification performance for PFM subreddits comparing local and global features reveals a consistent performance for RFC.

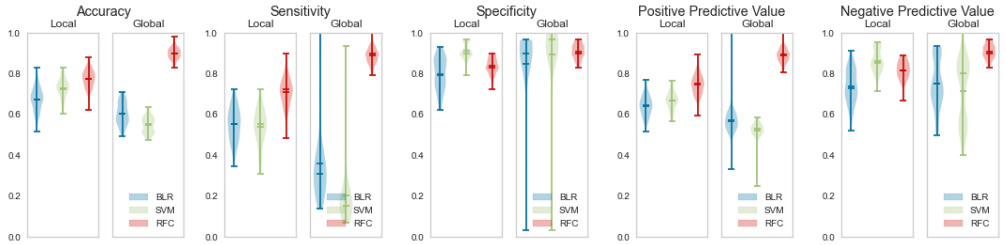

Figure 6.14: Classification performance for Ask subreddits comparing local and global features are a little more varied by comparison to PFM subreddits in Figure 6.13.

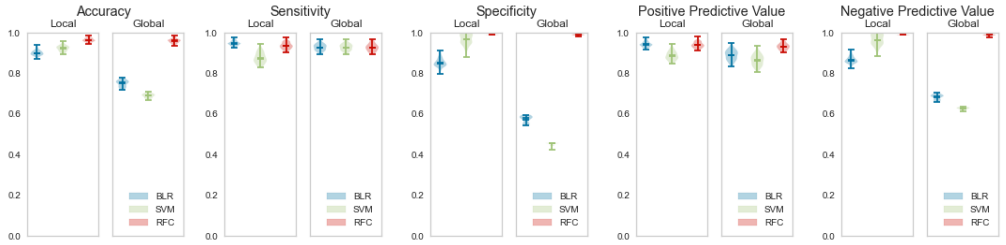

Figure 6.15: Classification performance for New subreddits comparing local and global features.

$P = 0.96$ for new using local features. Using global features, a RFC yields average accuracy values of $P = 0.97$ for PFM, 0.89 for Ask and 0.95 for new subreddits.

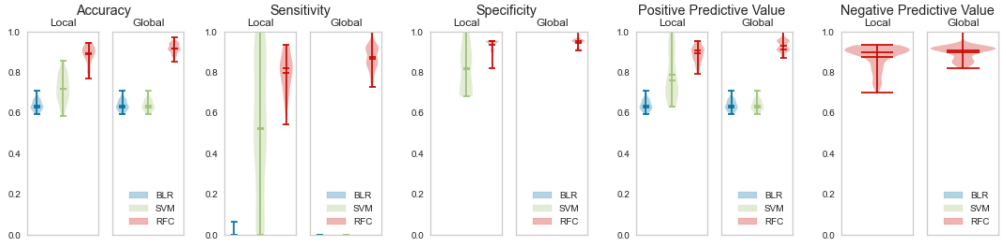

Figure 6.16: Classification performance comparing PFM with Ask subreddits comparing local and global features.

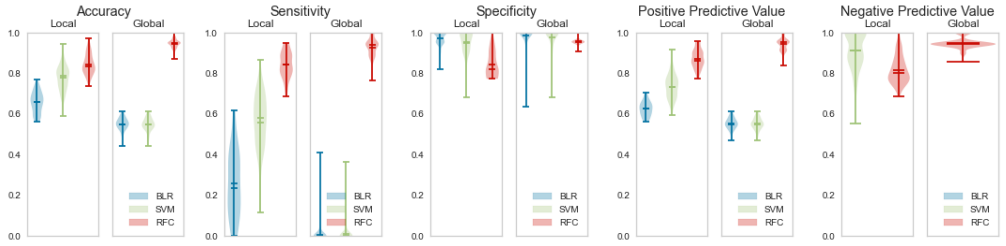

Figure 6.17: Classification performance comparing PFM with New subreddits comparing local and global features.

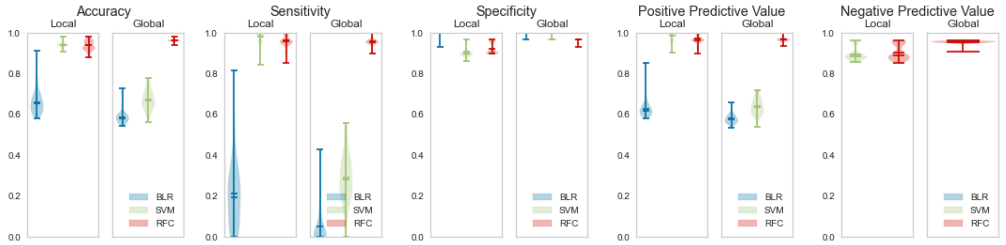

Figure 6.18: Classification performance comparing Ask with New subreddits comparing local and global features.

The results demonstrate a clear performance gain between classifying Ask vs New and PFM vs Ask subreddits as both the Positive Predictive Rate, PPR and Negative Predictive Rate, NPR are relatively stable with a mean PPV of $P = 0.96$ for Ask vs New and $P = 0.89$ for PFM vs Ask and a mean NPV of $P = 0.9$ for Ask vs New and $P = 0.87$ for PFM vs Ask using local features trained on a RFC. This demonstrates the effectiveness of local prediction with the ability to both separate and classify subreddits into groups with relative ease.

Global Features Using the same classifiers as in Section 6.4.3, the prediction models were trained using the global network features (see Section 6.4.2) with a view to understanding their predictive performance (see Research Question 4). As in Section 6.4.2, classification is performed in the scenario where each set is compared to a random baseline followed by classification when pairs of subreddit sets are involved.

By comparison to the pairwise prediction in Section 6.4.3, the violin plots indicate much greater dispersion of results in some cases: see Figures 6.16, 6.17 and 6.18. Here there

is much less consistency in results - for example, some results only yield one set of NPV's using a RFC. This reaffirms the idea that RFC performs consistently well across all sets. The results indicate that global features have much less stability as a basis for prediction in comparison to local features for categorisation involving two alternative sets of subreddits. This is especially true for predicting PFM with New subreddits as shown in Figure 6.17.

Comparing Local and Global Performance The classification results from Section 6.4.3 provide useful insights towards the effectiveness of using machine learning to predict characteristics of subreddits - such as potential behaviours correlating with potential misinformation activity. In this section, classification results are analysed by taking the average accuracy of each classifier and task cross-comparing prediction using local features over global features to help interpret comparing values. Accuracy was chosen specifically as it provides a reasonable metric for analysing prediction performance at a high-level.

The results presented provide evidence that local bipartite network features and induced graphlets play a significant role in understanding users' posting activity and similarity to others through subreddit association networks. It can also be observed that some classification tasks involving PFM subreddits perform better using local features whereas Ask subreddits perform better with global features. This suggests that PFM subreddits are dependent on more-detailed graphlet formations whereas Ask subreddits rely on simpler metrics, such as degree, using global features.

6.4.4 Discussion

The classification results from Section 6.4.3 address Research Question 4 and provide useful insights on the effectiveness of alternative approaches to accurately predicting the categorisation of different classes of subreddit. A number of key issues are evident and important to highlight. These relate to graphlet frequencies, predictability, the influence

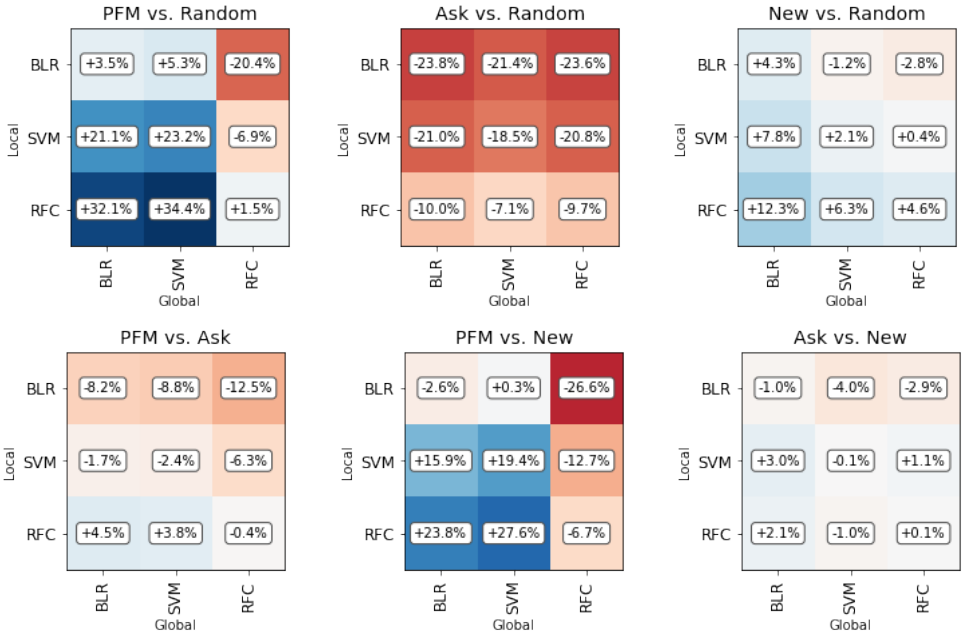

Figure 6.19: Pairwise comparison assessing the prediction gain (as a percentage) of local features over global features broken down into six prediction tasks. Percentage differences Δ_{ij} are derived by obtaining the difference between prediction values p_i, p_j scaled by the original value of the prediction task of interest p_i such that $\Delta_{ij} = \frac{(p_j - p_i)}{p_i} * 100$.

of local features, the potential for other classes of subreddits to be recognised and the potential for applicability beyond Reddit.

Firstly, it is important to note that the profile of graphlet frequencies differs between the PFM subreddits and other subreddits having a similar age. This supports the role of graphlets as a useful tool to detect distinguishing structural differences in the underlying subreddit association networks and answers Research Question 4 by providing evidence that local substructures are sufficient for providing signals for detecting disruptive behaviour. This also reaffirms the overarching utility in the representation that a network-based language-agnostic approach has potential for classification purposes as part of a framework of other network representation, as addressed in Research Question 2.

As shown in Figures 6.9 and 6.10, PFM and Ask subreddits remain distinct among

new subreddits despite being of a similar age. When considering age and subscriber counts (see Figure 6.5), the New and PFM classes of subreddit have substantial overlap. However, extracting features from the user association networks enables these classes to be distinguished. Note that this separation is arguably more pronounced when graphlet features are the basis for dimensionality reduction, as compared to standard graph-based metrics (i.e., Fig. 6.9 as compared to Fig. 6.10).

Secondly, the overall predictability of classification of PFM (and other) classes of subreddit are generally high while using alternative features and methods, based on both graphlets and global characteristics. This again supports the utility from the underlying representation of user association networks (see Research Question 2). It also further supports the creation of future monitoring agents for social media, with or without a human-in-the-loop. The results of this investigation reveal that a Random Forest Classifier overall provides the best and most consistent classification performance out of three alternative methods used. For example, classification of PFM and Ask subreddits can be achieved with relative ease as mentioned in Section 6.4.3 using this method.

Thirdly, both influential graphlet features and the influential global features for prediction of PFM subreddits appear to be related. The most influential global features involve node degree, while tree-like graphlets are the salient local features for prediction (see Figures 6.11 and 6.12). This indicates that local approaches (such as graphlet analysis) are well-suited to PFM classification and using alternative features (local and global) has helped to reaffirm this observation. The importance of local features opens up prospects for efficient real-time detection based on decentralisation and observation of graphlets in observable subnetworks. From extracting the eigenvector coefficients using PCA, it can be observed that aligned with global measures of degree, tree-like graphlets structures such as `S17` and `S43` are particularly important. Furthermore, the prediction results in Section 6.4.3 indicate that local features provide generally strong predictive utility (Research Question 4) for PFM and new subreddits.

Fourthly, from focussing on PFM subreddits and comparing these to other classes,

we can hypothesise that different classes of subreddit may leave particular underlying signatures in their corresponding user association networks, which reflect the different forms of user behaviour in which communities participate. For example, PMF and Ask fundamentally differ in the way in which users interact as one is used to distribute new articles whereas the other is primarily representative of stricter Q&A-like discussions. Consequently, it is possible to speculate that there may be a wider underlying taxonomy of significant graphlets for different classes of interaction in subreddits, but demonstrate nonetheless, that the user association network can be used to represent diverse affordances (Research Questions 2 and 3).

Finally, it can be noted that user association networks are a generalised approach to representing user behaviour in respect of social media content. This requires only an associative link between a user and another entity representing some form of content. This investigation has restricted the attention to Reddit (i.e., interactions with subreddits), but it is possible to believe that the general approach should yield insights into other forms of social media through modelling similarly (Research Questions 2 and 3). For example, these could include creating association networks based on user subscriptions, users tweeting certain hashtags on Twitter or users commenting across different articles. There are wide-ranging ways in which associations can be made and examined.

6.4.5 Key Findings

Overall, a general network representation of user association with social media content is used as a basis for prediction of important subclasses of content that align with the potential for misinformation (PFM). This has been applied to the Reddit social media platform, using the community data structure (see Section 3.2.2), utilising a number of alternative groups of subreddits for bench-marking purposes. The utility in this representation stems from being able to potentially categorise subreddits as having the potential for misinformation without undertaking any semantic analysis of content which helps answer Research Question 4.

This increases opportunities to detect classes of subreddits with agility, for example removing the need for translation in assessing foreign language social media. The analysis carried out in this analysis has identified that PFM subreddits are distinguished by the characteristics of their underlying user association networks - in other words the user-interaction with particular types of content has patterns associated with it that leave distinct signatures. These relate to the presence of high degree nodes which induce local tree-like structures that are seen in the form of particular graphlets that are strongly represented as compared to other induced substructures. The predictive capabilities of the graphlet census, alongside the global metrics, has been assessed, while employing PCA decomposition to identify the key features.

The methods included place an emphasis on the utility of network analysis where induced graphlets provide a fundamental topological representation. Furthermore, the use of the graphlet-based census serves as an ideal potential embedding technique for networks similar to that of tested techniques such as `gl2vec` [367]. By using PCA decomposition of graphlet-count feature vectors, this investigation highlights the latent differences across networks which open up insights that are not apparent when considering just the graphlet census in isolation. Through the use of dimensionality reduction, it can be observed how PFM communities and "anti-censorship" self-align and produce a distinct cluster in high-dimensional space aligned to the representation of induced subgraphs. In addition, moderation within the Ask subreddits appears to contribute to their distant positioning away from the PFM subreddits in the high dimensional space defined by induced graphlets as opposed to global features.

Numerous global features, such as degree-based metrics and clustering, are inherited as a consequence of simple local substructures, which has motivated the use of graphlets for analysis of complex networks across the wider literature. From this investigation there is evidence that graphlets have been effective because they are easily able to characterise the salient underlying features of the network related to node degree. More generally, graphlets also lend themselves to application in partially obfuscated network

scenarios, giving potential for their flexible deployment in wide-ranging scenarios.

This investigation establishes that classifying PFM subreddits can be achieved without the need for metadata, such as age and subscriber counts. It is evident that subreddits with a low subscriber count or young age (i.e., immaturity) aren't necessarily strong indicators of the potential for misinformation - subreddits with similar age or subscriber counts may have different network properties. The underlying behaviour of the users aligned to different classes of subreddit is the significant differentiating factor as this impacts the on the structure of interactions. This reaffirms the idea that the user association network can be used to model behaviour derived from social media activity according to Research Questions 1 and 3.

Finally, this helps assess Research Questions 2 as it is possible to believe that the exemplar presented in this investigation provides a useful proof of concept that could be extended to address other misinformation scenarios where the intent is to undermine public perceptions and rational behaviour on an alternative social media platform as part of a larger framework. The approaches considered in this investigation are relevant to future applications where subreddit classification can be performed at scale for situations such as automated moderation for growing communities, without recourse to semantic analysis.

6.5 Conclusions

To conclude, this chapter presents user association networks as a scalable and effective solution for modelling the dynamics between users and communities according to the community data structure as introduced in Chapter 3. This was demonstrated by investigating COVID-19 misinformation (see Chapter 2, Section 2.1.4) on Reddit where the concept of user association networks were used to form subreddit association networks (otherwise known as SAN's). These networks were used to model the posting interactions surrounding a subreddit of interest with the ability to accurately detect

subreddits with the potential for misinformation to emerge and others.

The investigation throughout this chapter has employed a combination of local (graphlet counting) and global (degree, centrality, density e.t.c) network-based metrics with the intention to form feature vectors that would later be used as part of a classification task. The results sufficiently answered Research Question 4 by concluding that graphlets provide the best results when trained on three different types of classifier. It was discovered the success of the classification results is partially due to simple global measures such as degree and local tree-like graphlets structures.

Overall, the implications of user association networks are significant for this thesis for a number of reasons. Firstly, this investigation addressed Research Question 1 by demonstrating that behavioural networks can be defined in the form of user association networks to provide a better representation of community behaviour and dynamics than simple metadata (such as age and subscriber counts, in the case of Reddit).

Secondly, it is important to acknowledge that the use of user association networks addresses Research Question 2 by demonstrating how this could be applied at scale and to include other platforms and/or websites which share the community data structure presented in Chapter 3.

Thirdly, the results drawn from the investigation address Research Question 3 by revealing that this network representation has the ability to classify a multitude of different types of community which represent diverse affordances and concludes by mentioning that there may exist a broader set of taxonomies each with their own distinct network profile. As a result, user association networks have broader implications for social media more generally and apply equally to other platforms as well as Reddit.

Finally, as addressed previously, this investigation offers a language-agnostic solution for detecting potential indicators of misinformation identified through characteristics of network-based representations (Research Question 4).

Chapter 7

Conclusions and Future Work

Disruptive behaviour, as seen throughout the context of this thesis, is a major issue for social media platforms and the individuals that use these platforms on a regular basis either for consuming news or socialising informally with friends (as addressed in Chapter 2). Social media has become such an influential part of the day-to-day proceedings of societal norms. As a result, it is clear that computational solutions are needed to combat disruptive behaviour in an autonomous or semi-autonomous (e.g. human-in-the-loop) fashion.

There have been many attempts to better understand and combat this issue yet very view studies consider using network-based features excursively as the basis for classification which is the focus of this thesis. This thesis provides a framework which offers a language-agnostic solution which can be performed at scale and across multiple platforms.

The hypothesis introduced at the start of this thesis in Chapter 1 is reconsidered in light of the findings presented. This is restated as follows along with the supporting research questions to help aid the investigation of the hypothesis:

Hypothesis*: Anomalous activity related to conflict or disruption in social media can be detected through the construction and analysis of networks representing different types of user behaviour and interaction, based on alternative affordances provided by social media.*

Consequently, this motivated the following research questions:

Research Question 1. *How can behavioural networks be defined from activity on social media platforms?*

Research Question 2. *Is it possible to produce a concise framework of alternative network-based representations capturing alternative affordances provided across social media platforms?*

Research Question 3. *How can diverse affordances and the interactions they facilitate be represented?*

Research Question 4. *To what extent can local features (i.e., subgraphs) provide signalling, on which prediction of disruptive behaviour be can be made?*

In response to the hypothesis and research questions, Chapter 2 provides context to the issue of disruptive behaviour on social media by identifying the different types of disruptive behaviour that can take place and how it has evolved over time. Following on from this, an overview of the different ways to model user activity and the role of complex networks is provided. This chapter concludes by acknowledging that a framework is needed to approach this problem (see Research Question 2).

This is addressed in Chapter 3 in which a novel framework for categorising social media platforms is induced though the concept of data structures and suggests three possible network representations suitable for capturing disruptive behaviour - transitional, user-to-user and user association. As a result, Chapter 3 addresses Research Questions 1, 2 and 3 by demonstrating how behaviour networks can be constructed from diverse affordances of social media activity using the three network representations as a framework.

Chapters 4, 5 and 6 are dedicated to the exploration of each representation in the context of detecting the presence of disruptive activity. These three network representation inform the core research of this thesis and explore the hypothesis by demonstrating firstly, how behavioural networks can be defined from social media activity (see Research Question 1). Secondly, how the three network representations can be used as part of a broader framework of detecting disruptive activity (see Research Question 2). Thirdly,

how each network representation can represent diverse interactions according to different affordances (see Research Question 3). And finally, how the detection of disruptive behaviour can be made through prediction using local network features including a combination of motif analysis and subgraph counting (see Research Question 4).

In Chapter 4, transitional networks were used to model user activity in the form of switches from the perspective of a user and content. Both Wikipedia (content-oriented) and Reddit (user-oriented) were used to investigate the hypothesis for modelling switching patterns to differentiate between disruptive and normal behaviour.

This was followed by Chapter 5 where user-to-user networks were used to model interactions taking place between other users. In doing so, these networks were primarily focused on conversational dynamics which took place in the form of replies, mentioning and quote retweets using Twitter and Reddit as platforms to investigate the hypothesis.

Finally, Chapter 6 uses bipartite networks to model the relationship between a user and a community on Reddit through the concept of user association networks. These were used to model the users' posting interactions surrounding a subreddit of interest to detect the potential for misinformation to emerge.

Consequently, Research Questions 2 and 4 were used to help conceptualise a framework of network-based representations for detecting disruptive behaviour and are addressed through the findings of Chapters 4, 5 and 6. These chapters demonstrate how each of the three network representations (as defined in Chapter 3) can be used to detect disruptive activity based upon a combination of local and global metrics.

To expand on findings against the research questions, the preceding sections are structured as follows:

- **Section 7.1: Summary of Results**: A detailed summary of the results gathered from the research with respect to the research questions to explore the hypothesis of this thesis.

- **Section 7.2: Discussion and Future Research**: A detailed discussion regarding the various changes of constructing and analysing social networks throughout this thesis and consideration of the future work of this research though the framework presented in Chapter 3.
- **Section 7.3: Research Impact**: A discussion focused on the implication of this thesis and how it contributes to new knowledge within the surrounding literature.
- **Section 7.4: Final Conclusions**: The overall conclusions which summarises the impact of this research and considers the implications beyond the scope of this thesis.

7.1 Summary of Results

As mentioned previously, the central findings of this thesis primarily relate to Research Questions 2 and 4 which are addressed in Chapters 4, 5 and 6. These chapters demonstrate how each of the three network representations can be used to capture **disruptive and non-distributive behaviour using a combination of simple network-based metrics** such as global features (including in-degree, transitivity and reciprocity) and local features (presence of certain subgraphs).

The results of this thesis are summarised in detail with respect to each of the research questions. These are discussed and evaluated as follows:

7.1.1 Research Question 1

How can behavioural networks be defined from activity on social media platforms?

As described earlier, Research Question 1 addresses the methodological structure regarding how behavioural networks are defined from activity collected on different social media platforms. These behavioural networks are initially defined in Chapter

3 and are instantiated and tested in Chapters 4, 5 and 6 as part of a framework of alternative network-based representations (see Research Question 2). These behavioural networks are defined according to the presence of certain data structures using three network representations - transitional, user-to-user and user association.

Firstly, a transitional network was used in Chapter 4 to model temporal switching behaviour using the collaborative and feed data structures. The results of Chapter 4 conclude that using a content-oriented approach (see Section 4.4), transitional networks on Wikipedia can be used to model exchanges between editors according to the collaborative data structure (as introduced in Chapter 3, Section 3.2.4).

In addition to this, it is important to recognise that this approach can not be transferred in the same way (see Section 4.5). This was discovered by using a user-oriented approach though the concept of switching motifs - a proposed technique for discovering statistically significant switches (see Section 4.5.5). This investigation used Reddit to observe switches between subreddits according to the feed data structure (as introduced in Chapter 3, Section 3.2.5). The results concluded that in order to build a detailed model of user activity over time, more data was required to include a broader and more diverse range of subreddits making it easier to compare with other accounts. Consequently, this produces a trade-off with computational complexity making the task of computing statistically significant switches resource intensive for such a large corpus of subreddits.

Secondly, a user-to-user network was used in Chapter 5 to model behaviour which takes place between users using the message data structure. Chapter 5 focused on the role of user-led interactions though user-to-user networks as a simple yet versatile network representation for modelling conversational dynamics - a widely used feature and common to many social media platforms. To begin, Twitter (see Section 5.4) was used to extract three user interactions (quote retweets, mentions and replies) according to the message data structure (see Chapter 3, Section 3.2.3) for understanding conversations with respect to controversial and non-controversial narratives regarding vaccine usage. Furthermore, the investigation on Reddit (see Section 5.4), which also aligns with the

message data structure, reaffirmed this observation that user-to-user networks are a versatile network representation for modelling communication between users.

Finally, a user association network was used in Chapter 6 to model posting behaviour initiated by a user and directed towards a community based upon the community data structure (see Chapter 3, Section 3.2.2). The results of Chapter 6 demonstrate how user association networks provide an innovative approach for analysing cross-community dynamics at large by using Reddit to investigate the hypothesis (see Section 6.4) to model the relationship between a user and a subreddit based upon the explicit action of a user submitting a post to a subreddit.

7.1.2 Research Question 2

Is it possible to produce a concise framework of alternative network-based representations capturing alternative affordances provided across social media platforms?

Using the behavioural networks defined as part of Research Question 1, Research Question 2 seeks to address how these network representations can be used collectively as part of a framework of diverse affordances across social media. Chapter 3 aids the structure of the framework by demonstrating how the three behavioural networks align with the data structures through different social media platforms.

To assess the suitability of the framework, each network representation is tested in Chapters 4, 5 and 6 by addressing different affordances using the data structures provided. This demonstrates that it is possible to produce a concise framework and is discussed in the preceding sections as follows:

Chapter 4: Transitional Networks

		Data Structure			
		Community	Message	**Collaborative**	**Feed**
Network	**Transitional**	-	-	**See 4.4 (Wikipedia)**	**See 4.5 (Reddit)**
	User-to-user	-	See 5.4, 5.5 (Twitter, Reddit)	-	-
	User Association	See 6.4 (Reddit)	-	-	-

Table 7.1: Reminder of how Chapter 4 aligns with the investigations used to explore the hypothesis of this thesis.

In Chapter 4, transitional networks were used to model activity derived from Wikipedia and Reddit. As shown in Table 7.1, relative to the rest of the thesis, the role of the transitional network is to examine how activity aligned with the collaborative and feed data structures can be considered as part of a wider framework.

This representation is designed to model interactions in the form of switching behaviour in a content-oriented and user-oriented manner. A content-oriented approach was used to observe switching between users in a revision network based around an article of interest - the content (see Section 4.4). Additionally, a user-oriented approach was used to observe how a user switches between different subreddits (see Section 4.5).

As shown in Chapter 4, only a content-oriented approach yields results suitable for differentiating between disruptive and non-disruptive behaviour. While this does not apply in a user-oriented manner, these findings address Research Question 2 by reaffirming the significance of content-oriented transitional networks as part of a wider framework for detecting disruption according to temporal switching behaviour.

Chapter 5: User-To-User Networks

		Data Structure			
		Community	**Message**	Collaborative	Feed
Network	Transitional	-	-	See 4.4 (Wikipedia)	See 4.5 (Reddit)
	User-to-user	-	**See 5.4, 5.5 (Twitter, Reddit)**	-	-
	User Association	See 6.4 (Reddit)	-	-	-

Table 7.2: Reminder of how Chapter 5 aligns with the investigations used to explore the hypothesis of this thesis.

As discussed earlier, user-to-user networks were used to represent interactions which take place among a set of users based upon activity derived from Twitter and Reddit. As highlighted in Table 7.2, user-to-user networks contribute to the framework by modelling activity which align with the message data structure.

In Chapter 5, user-to-user networks used the message data structure to observe how a pair of users communicate with each other. As mentioned previously, this was observed using Twitter to examine a range of different message-based interactions in the form of quote retweets, mentions and replies (see Section 5.4) and on Reddit using egocentric networks based upon replies surrounding a single user (see Section 5.5).

Consequently, the results from Chapter 5 reaffirm the significance of the user-to-user network as a result of assessing their utility with interchangeable interaction types across two different platforms. These results address Research Question 2 which establish user-to-user networks as an important representation for message-based interactions which contributes to a framework for detecting disruptive behaviour.

Chapter 6: User Association Networks

		Data Structure			
		Community	Message	Collaborative	Feed
Network	Transitional	-	-	See 4.4 (Wikipedia)	See 4.5 (Reddit)
	User-to-user	-	See 5.4, 5.5 (Twitter, Reddit)	-	-
	User Association	**See 6.4 (Reddit)**	-	-	-

Table 7.3: Reminder of how Chapter 6 aligns with the investigations used to explore the hypothesis of this thesis.

As demonstrated through Chapter 6, a user association network can be used to model the relationship between a user and community based upon activity recorded on Reddit. As demonstrated in Table 7.3, this also contributes to the framework by representing activity which aligns with the community data structure.

Chapter 6 utilises the user association network to model behaviour in the form of posting interactions on Reddit between a user and a subreddit (see Section 6.4). In doing so, this approach was used to observe how users post to a subreddit of interest and others based upon mutual connections.

The results of Chapter 6 demonstrate how the user association network is an important network representation for detecting communities which have the potential to spread misinformation and can be used as part of a framework (among other network representations) for detecting disruptive behaviour based upon cross-community dynamics.

7.1.3 Research Question 3

How can diverse affordances and the interactions they facilitate be represented?

As mentioned in Chapter 1, different social media platforms can be used in different ways meaning that not all interaction are alike. To address Research Question 3, Chapter 3 introduced the concept of data structures to capture different type of affordances and their resulting interactions which, in turn, helps inform the network representation. This also provides the basis for helping characterise different social media platforms according to the affordances they facilitate.

Table 7.4 reveals how each of the platforms used in this thesis align with each of the four data structures.

		Data Structure			
		Community	Message	Collaborative	Feed
Platform	Wikipedia	N/A	N/A	*See 3.4.1*	N/A
	Reddit	*See 3.4.2*	*See 3.4.2*	N/A	*See 3.4.2*
	Twitter	N/A	*See 3.4.3*	N/A	*See 3.4.3*

Table 7.4: Overview of how the platforms of interest align with different data structures as used within this thesis for capturing diverse affordances.

As shown in Table 7.4, all four data structures (community, message, collaborative and feed) can be used collectively to represent diverse affordances across Wikipedia, Reddit and Twitter. This was demonstrated in Chapters 4, 5 and 6.

As shown in Chapter 3, Section 3.2.2, social media activity which include communities (such as Reddit) are important for understanding how like-minded individuals develop connections. The community data structure was used in Chapter 6 to focus on activity with respect to users and how they interact with subreddits (the communities) on Reddit in the form of posting. In doing so, this made it possible to understand the community-centric interactions which take place on Reddit using the user association network (see Section 6.4).

Secondly, the message data structure was used in Chapter 5 to analyse user activity in the form of conversations or "messages" which are exchanged between a pair of users. As discussed earlier (see Chapter 3, Section 3.2.3), communication between users is an important component for any social media platform and is examined on Twitter (see

Section 5.4) and Reddit (see Section 5.5) in detail by focusing on messages which take place through the use of commenting and replying using a user-to-user network.

Thirdly, collaboration on social media is important for understanding how users seek to engage with others in an attempt to improve the state of a piece of content (see Chapter 3, Section 3.4.1). As a result, the collaborative data structure was used within Chapter 4 to study exchanges between Wikipedia editors. These exchanges were oriented around a single article in a content-oriented fashion (see Section 4.4) with a transitional network.

Finally, as mentioned in Chapter 3, Section 3.2.5, most social media platforms utilise some form of news feed algorithm to present relevant information to users. The feed data structure was proposed in Chapter 4 to understand how information presented over time can be interpreted using a transitional network to observe how users switch between different communities (subreddits) on Reddit (see Section 4.5). Similar comments can be made on Twitter regarding how tweets presented in the form of a "timeline" which also applies to the feed data structure.

7.1.4 Research Question 4

To what extent can local features (i.e., subgraphs) provide signalling, on which prediction of disruptive behaviour be can be made?

In response to Research Question 4, various classification algorithms were used to assess the ability to predict disruptive from non-disruptive activity with the aid of the appropriate data structures and network representations (see Research Questions 3 and 1). As initially outlined in Chapter 1, this was achieved using a combination of local (motif and subgraph analysis) and global analysis and trained using classification algorithms including a support vector machine (SVM), binary logistic regression (BLR) and a random forest classifier (RFC). The results are featured in detail within Chapters 4, 5 and 6 and are summarised as follows.

In Chapter 4, through the use of network motif analysis, the SRP (subgraph radio

profile) identified that simple linear connections (such as chain-like triads and dyads) are statistically significant and facilitates the classification of controversial and non-controversial articles on Wikipedia (see Section 4.4). As a result, content-oriented transitional networks have the potential to capture behavioural signals of users collectively when centred around a piece of content. However, it is important to note that this approach does not apply to user-oriented transitional networks as a result of being unable to perform local analysis due to limitations such as volume and scalability (see Section 4.5.5 for more).

In Chapter 5, the results of the investigation on Twitter (see Section 5.4) concluded that user-to-user networks derived from Twitter's mention interaction provide the most predictive utility (along other interaction types including quote retweets and replies) when using global metrics (including in/out-degree, density, reciprocity and transitive) as a feature vector. The results reveal how global features out perform local features with respect to classification performance. In addition to this, the second investigation used Reddit (see Section 5.5) to study reply networks in an egocentric manner by centring activity around a single user. By counting and enumerating over all 4-node subgraphs resembling a tree-like structure, the results reveal that it is possible to differentiate between normal and disruptive users when using normalised subgraph counts as a feature vector. The results produce a classification accuracy as high as $p = 0.86$ when supplemented with temporal features. As a result, both Twitter and Reddit demonstrate that user-to-user networks can be used to detect disruptive behaviour according to both global and local metrics which, as mentioned earlier, can be improved further when combined with temporal features.

Finally, in Chapter 6 the results of the investigation on Reddit (see Section 6.4) concluded that local network metrics (normalised bipartite graphlet counts) can be used to identify subreddits with the potential for misinformation to emerge based upon the presence of certain bipartite graphlets. Furthermore, the findings demonstrated how this approach has success in differentiating between other types of subreddit including Q&A

("Ask") and new subreddits, thus exposing the possibility of classifying other types of subreddit and communities in future work. These results reveal that it is possible to extract behavioural signals, making the task predicting disruption, when centred around a community, possible with reasonable accuracy.

Overall, the results of Chapters 4, 5 and 6, demonstrate that network-based representations (including other language-agnostic metrics such as temporal features) can be used to detect disruptive activity using a combination of both local and global metrics.

7.2 Discussion and Future Research

This section discusses various aspects of the research conducted in this thesis and exposes a number of different challenges and limitations associated with studying complex social networks in an attempt to address the hypothesis. Additionally, this section considers the framework introduced in Chapter 3 and discusses alternative solutions using this thesis as the foundation for future research. These are aligned to three points of discussion, which are as follows:

7.2.1 Challenges of Generating Networks

Through the concept of data structures, the framework introduced in Chapter 3 provided a convenient approach for understanding which platforms aligned to specific data structures which, in turn, informed the appropriate network representation(s). While this approach simplifies the process of selecting appropriate network representations, this fails to consider the challenges associated with generating the networks from raw data in an attempt to define behavioural networks (see Research Question 1) and representing diverse affordances and their resulting interactions (see Research Question 3).

Scale

By nature, many complex social networks are large and as they are used to represent many different types of interaction derived from real-world human activity [12, 280]. This presents a challenge with respect to the time required to generate the networks and the space needed to store the network in memory. Many networks can take several hours to construct and produce networks which are large with respect to computational storage. Consequently, this has a negative impact on subsequent analysis.

This was partially evident in Chapter 6 whereby subreddit association networks we are generated in a recursive manner and increasing in size per iteration. The design of the approach meant that the size of the network scaled exponentially making it increasingly difficult to analyse in real-time. Similar comments can be made for the investigations featured in Chapter 5 as simple interactions like replies can populate in quick succession depending on the topic.

Sample Size

In addition to the scale, another issue which impacts generating networks is sample size. In order to build an accurate representation, a reasonable sample of data is required in order to form a detailed model of a user's behaviour or signature. As a result, this produces a trade-off between level of detail and size and to find a compromise between the two is a challenge.

These issues were observed throughout the thesis as platforms such as Reddit produce very little in the way of user activity. This is particularly evident in Chapter 4 where the use of user-orientated transitional networks were used in an attempt to model common switching patterns. The investigation ultimately concluded that in order to build an accurate model more diverse detail was needed in order to find distinctive behavioural signals. This was due to including user accounts that were created recently meaning that very little activity was recorded. Similar observations were made on Wikipedia as

articles were included that were either created recently or not as well known meaning that not enough information was provided to determine whether it was controversial or not thus contributing to Type I and Type II errors.

Data Loss

One of the challenges of generating network representations is to ensure that the network preserves as much detail as possible without loss of information. For example, many social networks (including the networks used in this thesis) are static and do not feature some form of temporal component. As a result, this fails to consider how the network evolves over time and could disregard key features which further improve the detection of disruptive activity.

In Chapter 5, user-to-user networks were generated to model common message-based interactions on Twitter and Reddit which did not include timestamps as part of the representation. This meant that there was no temporal context to the order in which interactions, such as replies, took place. Consequently, this means that features such as conversation depth (the number of exchanges between a pair of users over time) were not included which may have had a significantly improved prediction results if the information was preserved.

Temporal Networks

Within this thesis, the behavioural networks are generated and analysed as static representations without considering any form of temporal component (e.g. temporal networks). The reason for considering static networks is that they simplify the process of defining behavioural networks and that many of the algorithms used for performing local and global analysis can only be performed using static networks.

As discussed previously, the main limitation of static networks is that they fail to represent the dynamic nature of human behaviour and their social media interactions

which evolve over time. Studying temporal networks are feasible, and one possible solution would involve applying timestamps to nodes and edges within a network and to consider interaction which take place within specific windows of time. Alternatively, temporal networks can be utilised using techniques derived from the Barabási–Albert model for studying the growth of scale-free complex networks over time [12] or through temporal network motifs, a novel approach for studying ordered sequences among triadic subgraphs [301].

7.2.2 Challenges of Analysing Networks

In response to Research Questions 2 and 4, the methodological structure of this thesis involves analysing complex social networks using standard methods derived from network science. In an attempt to strengthen the quality of the research and to maintain scientific rigour, the methods featured in this thesis are applied consistently to all network representations. In doing so, this thesis exposed a number of challenges, many of which are synonymous with generic complex networks in an attempt to build a concise framework of network representations (see Research Question 2) for detecting disruptive behaviour (see Research Question 4).

Local Metrics

Within this thesis, the role of local metrics were used to discover statistically significant underlying substructures present within each of the networks with the intention of finding unique structural properties to detect the presence of disruptive behaviour and answer Research Question 4. In line with Research Question 4, this was explored using either network motif analysis or induced subgraph counting. Both solutions expose the issue of the subgraph isomorphism problem [102] whereby subgraph counting is performed in polynomial time [128]. To combat this issue, the work featured in this thesis considers a smaller subgraph size typically containing three or four nodes in

the interest of speed and preserving resources. Furthermore, the use of motif analysis requires generating many random networks in the form of a null model which increases computation and resource usage. For this reason, subgraph counting was sufficient as networks were compared with each other meaning that there was no need for a random null model.

Global Metrics

Local metrics were used to describe the entire network structure using a single metric. By convention, complex social networks primarily focus on generic properties such as in/out-degree (max, min, average), transitivity, density and reciprocity as proxies for gauging the presence of popular/influential users, community structures and conversational dynamics [366, 242, 167]. This is also addressed in Research Question 4 to determine the predictive utility of local metrics over global where appropriate. Unlike local metrics, global metrics can be calculated at a much faster rate by comparison, making it easier to perform this analysis in real-time. This comes at a trade-off as many of the structural properties embedded within the network can be found using local metrics. Furthermore, it is possible to conclude that global metrics provide less detail as metrics which are calculated on a node-by-node basis (such as in/out degree) need to be aggregated by calculating the minimum, maximum, or average value. This makes it difficult to compare between other networks as the true distribution of the values is not considered. In Chapter 6, the results of the investigation on Reddit reveal how bipartite graphlets (local) outperform global metrics for differentiating between different classes of subreddit. The results from the same investigation also suggest that the graphlets which provided the most predicted futility correlate directly with simple global features such as in/out degree.

Real-Time Processing

As discussed earlier, techniques such as motif and subgraph analysis are resource intensive and are therefore not suitable for obtaining results in real-time. This has a drawback for applications and which will time processing is essential (e.g. content moderation for large-scale platforms) and to identify disruptive behaviour as and when it appears. A solution for this would be to develop algorithms in which motif or so graph analysis can be performed on incremental steps such that local features can be extracted or a smaller portion of the network.

For example, platforms such as Twitter allow users to stream tweets using the API using a set of keywords or accounts of interest. As tweets appear they can be added to the network and analysed concurrently in real-time using the techniques featured in this thesis.

These techniques and could potentially be used to save time and resources, and it is possible for the network representations introduced in this thesis to be transferred to use a similar process. Computational methods could be adapted to achieve this however this goes beyond the scope of this thesis but has implications for future research.

7.2.3 Data Structures and Network Representations

As previously mentioned, Research Questions 1, 2 and 3 are addressed in Chapter 3 by demonstrating how behavioural networks can be defined to capture diverse user affordances which, in turn, can be used for detecting disruption. In doing so, this led to the formation of three network-based representations which are informed by the presence of certain data structures.

Overall, the data structures introduced in this thesis helped formulate the framework (see Research Question 2) and were sufficient for characterising different social media platforms according to their affordances which helped inform the appropriate network

representation.

Towards the end of Chapter 3, the investigations featured within the thesis were discussed in relation to their respective data structure and network representation. This laid out the foundations for the thesis and exposed the feasibility of possible future work. Table 7.5 recalls the format of this thesis and provides scope for future work based upon popular trends within society and the supporting literature. This section discusses how this framework could be used as a general approach for classifying different social media platforms and capturing behaviour more generally and does not necessarily have to involve disruption.

		Data Structure			
		Community	Message	Collaborative	Feed
Network	Transitional	-	-	See 4.4 (Wikipedia)	See 4.5 (Reddit)
	User-to-user	-	See 5.4, 5.5 (Twitter, Reddit)	-	-
	User Association	See 6.4 (Reddit)	-	-	-

Table 7.5: The original table featured in Chapter 3, used to outline the work completed in this thesis and suggest potential research for future work.

This thesis focused on three platforms (Wikipedia, Reddit and Twitter) that were introduced in Chapter 3 as "platforms of interest". While these three platforms adequately utilised all four data structures, it is important to acknowledge that these can be applied to other social media platforms. This section offers suggestions for future research using the framework and methods developed in this thesis.

Community Data Structure

As mentioned in Chapter 3, many social media platforms provide a mechanism for allowing users to integrate with other like-minded users through community structures [84]. In this thesis, this structure was observed in Chapter 6 using Reddit where posting activity was analysed by modelling the many-to-many mapping between users and

subreddits. The benefit of this data structure is that the *Interaction* field can be used to model alternative types of interaction which go beyond simple actions like posting to a subreddit, for example.

Alternatively, this approach could also be used to model other different types of interaction of Reddit including commenting and replying to posts within a community. Furthermore, this applies directly to other platforms such as Facebook through the concept of Facebook Groups. Similarly, the community data structure can be used to capture interactions such as posting and commenting, could also include joining and subscribing to certain topics.

Message Data Structure

Following on from the community data structure, the message data structure is another fundamental component for categorising platforms as communication is an essential feature to all social media. Within the thesis, this was primarily demonstrated in Chapter 5 through Twitter and Reddit by modelling the message data structure against discussion threads on Reddit and through quote retweets, mentions and replies on Twitter. Much like the community data structure, the message structure can also be used to model multiple types of interaction using the *Interaction* field, as demonstrated with the investigation on Twitter. However, this model is focused exclusively on exchanges which take place between users in a pairwise fashion.

In addition to the investigations provided in this thesis, the use of the message data structure has broader implications for social media but also includes other services and websites where users can participate in a conversation through the use of a comment section. An appropriate example of this is Disqus[1], an online service allowing users to embed a comment section on their website to drive user engagement and facilitate conversations. Furthermore, the message structure applies equality to other online

[1]Disqus: `https://disqus.com/`

services such instant messaging platforms including WhatsApp, Facebook Messenger, Telegram and WeChat to list a few.

Collaborative Data Structure

The collaborative data structure is provided to include platforms whereby users collectively contribute to the delivery end quality of content. The most obvious example of this can be found in Chapter 4 using Wikipedia which was focus of this thesis. Using Wikipedia, the collaborative structure described the relationship between a user and an article based upon an action (amend, remove, revert e.t.c.) at a point in time. The benefit of this approach is that it provided a record of all actions which took place over time for a given article.

Beyond Wikipedia, all Wiki-based applications more broadly serve as ideal examples of the collaborative data structure in use. While the work of this thesis focused exclusively on Wikipedia, there is potential for this structure to be used in other more informal settings. For example, a similar approach could be used to model a user's contributions to code-based repositories such as GitHub or online pinboard generators such as Pinterest to allow users to create and remove pins on a virtual board.

Feed Data Structure

Most social media platforms provide a way of displaying information tailored to a user's interests in the form of a news feed. This was observed on platforms such as Reddit but equally apply to Twitter as well. As seen in Chapter 4, a user's profile displayed the most recent activity in the form of a news feed and was used to observe changes in activity over time to observe common switching patterns. The benefit of this data structure is that it can be used to model different types of information such as a single user's activity or an aggregation of activities from multiple accounts (e.g. the default news feed on Twitter).

As mentioned previously in Chapter 3, the use of news feeds and the algorithms used to present tailored information and has been the subject of discussion within the literature and has the potential to develop echo chambers and filter bubbles [116, 361, 350, 46]. More specifically, platform such as YouTube have received criticism for the lack of transparency regarding how recommendations are formed on a user's news feed [48, 73]. The framework provided in this thesis could offer a potential solution for future work and also aligns to the theme of disruptive activity.

Transitional Networks

Transitional networks provide a convent approach for modelling switching behaviour which occurs over time. The novelty of this representation is that temporal sequential information is embedded in the form of a static network making it possible to understand user activity through latent substructures. In Chapter 4, transitional networks were used to understand the switching behaviour of users from two perspectives (content-oriented and user-oriented) using both Wikipedia and Reddit to investigate the hypothesis.

As concluded in Chapter 4, it is clear that transitional networks perform well in a content-orientated context but does not transfer to a user-oriented approach. This was due to several factors concerning the high volume of subreddits and duplicated switches. Ultimately, the solution provided does not scale well with time and alternative solutions are needed in order to assess the true potential of user-oriented transitional networks using platforms beyond Reddit. Given that this approach is somewhat of a novel concept, further research and experimentation is needed.

User-To-User Networks

User-to-user networks offer a versatile representation for capturing different types of interaction which take place between pairs of users. As demonstrated in Chapter 5, Twitter was used to investigate the hypothesis by demonstrating how user-to-user

network representations can be used to encompass multiple types of interaction (quote retweets, mentions and replies) in a single representation. The advantage of this approach is that a single network representation can be used to model many different types of interactions without having to depend on other alternative network structures / representations.

While user-to-user networks lend themselves to capturing behaviour in message-based interactions, this approach can be extended to model other, less explicit types of interaction such as "sharing" and "liking" user content. Furthermore, methodological improvements can be made to further enhance the reliability and accuracy of this representation. For example, the user-to-user networks as introduced in this thesis are static meaning that there is no consideration for dynamic interactions as they evolve over time. As discussed earlier, this is especially important considering that all data structures and social media platforms introduced in this thesis provide temporal features in the form of a timestamp.

User Association Networks

Finally, user association networks use bipartite networks to model relationships between a user and community based upon explicit interaction such as posting and commenting. This was demonstrated in Chapter 6 using Reddit to map users against subreddits they have posted in previously. As a result, the benefit of this approach is that it facilitates the construct and analysis of cross-community interactions at scale, centred around a particular user or community of interest.

In view of the results of Chapter 6, role of user association networks serve as the basis for recommendation algorithms [408, 413] and community detection based upon bipartite network structures [244]. Very little research considers how these two applications either contribute to or are affected by disruptive behaviour. For example, could coordinate malicious posting behaviour within a collection of subreddit distort the results of community detection algorithms among "normal" activity? From a methodological per-

spective, much like user-to-user networks, user association networks could be adapted to feature temporal edges with the potential to understand how users migrate between communities over time.

7.3 Research Impact

The research performed within this thesis contributes to new knowledge, has an influence on the surrounding literature and provides confidence in supporting the hypothesis in multiple ways. As a consequence of investigating the hypothesis and supporting research questions, the findings have implications in three main areas: novelty, application and overall impact on society. These are discussed further as follows:

7.3.1 Novelty

To begin, the techniques set out in this thesis provide a means of understanding how social media platforms operate and provide a universal approach for capturing user behaviour according to high-level HCI principles for understanding how users use computers though affordances. As introduced in Chapter 3, the concept of data structures provides a convent framework for both modelling and comparing different social media platforms based upon the presence of certain features, a topic which is seldom used within the literature with respect to social media. In doing so, this approach helps navigate across the space of different social media platforms for future research and supports the selection of an appropriate network representation for behavioural analysis (see Research Question 1) and for capturing diverse affordances (see Research Question 3).

Secondly, the three network representations introduced in this thesis were explored using a combination of simple network-based features that are widely used among social network analysis and utilised techniques such network motif analysis and sub-

graph counting [263, 264, 320] for generating feature vectors to be used as part of a classification task (see Research Question 4). Very few studies consider using network representations for classification exclusively without the need for considering additional information such as text and other metadata. As a result, these network representations can be used as part of an ensemble of networks for detecting disruption (see Research Question 2).

Finally, perhaps the biggest implication is that the methods developed in this thesis are language-agnostic, meaning that analysis can be performed independent of spoken language (see Research Question 4). This is particularly important considering the widespread usage of social media globally whereby many spoken languages are used as a result. Many studies within the supporting literature which seek to detect disruptive behaviour though the use of NLP and only consider the English language (e.g. [115, 88, 260]) for analysis, which not representative of all social media activity.

The advantage of this approach is that there is no need to consider text as this is not included as part of the model as the focus of this thesis is on network analysis. As a result, the concept of using networks for representing users and or communities is not a novel concept within the literature, but analysis performed independent of other information (e.g. text) demonstrates the future potential of network representations within the domain of detecting disruptive behaviour.

While the results of this thesis demonstrate that it is possible to circumvent NLP for detecting disruptive activity, it is important to note that it can also be used alongside these methods to further improve reliability and accuracy. For example, a network-based approach can be used to detect disruption at scale which can then be verified using a combination of NLP and human-in-the-loop involvement to ensure results are as accurate as possible.

7.3.2 Application

Overall, the insights obtained from this thesis demonstrate how different network representations can be used to detect different forms of disruptive behaviour using simple metrics derived from social network analysis. As a result, the ability to detect and identify disruption occurring on social media platforms is advantageous and has implications for advancing techniques for content moderation and understanding more about how disruptive users mobilise on different social media platforms. Additionally, as mentioned previously, the solutions developed in this thesis are language-agnostic which is highly desirable considering the scale and widespread adoption of social media internationally [1, 2, 179].

Of the social media platforms featured in this investigation, all depend on some sort of community-driven moderation in the form of volunteers (Reddit) and administrators (Wikipedia) or as staff hired by the platform's parent company (Twitter). The techniques featured in this thesis could be used as part of a semi-automated solution using the human-in-the-loop model for detecting disruptive activity in a real time as part of a hybrid approach similar to Botometer [111], Hoaxy [336] and CoVaxxy [115]. It is possible to speculate that the methods developed in this thesis could, upon further improvement, be used to flag instances of disruption at a much faster pace by comparison.

7.3.3 Impact on Society

Given the ubiquitous support and continued growth of social media [308], platforms, such as Twitter and Facebook, started off as a novelty and are now considered essential in order to communicate with others. As a consequence, a by-product of its popularity has resulted in disruptive behaviour such trolling and misinformation becoming more mainstream and increasing discourse within society [171, 243]. This thesis alludes to specific instances of disruptive behaviour through the investigations presented in

Chapters 4, 5 and 6.

More specifically, this thesis addresses a range of relevant and timely issues which are likely to have an impact offline. The work of Chapters 5 and 6 addressed COVID-19 misinformation (see Chapter 2, Section 2.1.4) which can become a threat to public health. Chapters 4 and 5 focused on political trolls (see Section 2.2.2) resulting in polarisation. Finally, Chapter 4 studied controversial articles which are likely to attract users seeking to cause disruption.

Due to the adoption of the internet [1, 2, 179], this is not a problem that is localised to a particular location or region but is instead a global problem. As a result, urgent action is needed to help combat the propagation of disruptive activity on social media. Through the use of social network analysis, this thesis offers a language-agnostic solution which can be scaled accordingly to adjust to the growing demands of social media and the internet more broadly. While the efforts of this thesis focus on online environments, offline movements are also important but go beyond the scope of this thesis.

7.4 Final Conclusions

***Hypothesis**: Anomalous activity related to conflict or disruption in social media can be detected through the construction and analysis of networks representing different types of user behaviour and interaction, based on alternative affordances provided by social media.*

To conclude, this thesis answers the hypothesis by providing substantial evidence that a network-based approach can be used to capture disruptive activity on online social media platforms without the need to consider additional information such as spoken language (Research Question 4). This is demonstrated as part of a framework (Research Question 2) through the use of three network representations which define behavioural networks from social media (Research Question 1) as a way of capturing different behavioural signals and diverse affordances (Research Question 3) from users on a

collective and individual basis. As a result, this thesis rigorously assessed the utility of these network representations by assessing multiple social media platforms by using Wikipedia, Reddit and Twitter to explore the hypothesis.

Through the use of high-level global metrics and local subgraph analysis, these network representations were examined using various classification techniques to detect the presence of disruptive activity and to assess their predictive utility. As a result, the work set out in this thesis demonstrates that this can be achieved with relative success. Consequently, this has broader implications for society to help combat issues such as trolling and misinformation using methods which are language-agnostic and scalable to adjust to the growing demand for the internet.

Appendices

E Chapter 5

E.1 Case Study 1

Term	Total	Mean	Std
#covidvaccineispoison	19	3.8	0.4
#vaccineskill	18	3.6	0.8
#vaccinesarepoison	18	3.6	0.8
#billgatesbioterrorist	17	3.4	0.49
#syringeslaughter	16	3.2	1.6
#vaccinemicrochips	16	3.2	0.4
#fuckvaccines	16	3.2	0.75
#depopulation	16	3.2	0.75
#vaccineharm	16	3.2	0.75
#vaccinedamage	15	3.0	1.1
#covid19hoax	15	3.0	1.1
#arrestfauci	15	3.0	0.89
#arrestbillgates	15	3.0	0.89
#coronavirusfraud	15	3.0	0.89
#vaccinefraud	14	2.8	0.98

chinese virus	14	2.8	1.17
#chinesevirus	14	2.8	1.17
#scamdemic	14	2.8	0.98
#scaredemic	14	2.8	0.98
#donttakethemicrochip	14	2.8	1.47
#kungflu	13	2.6	0.8
#antivaccine	13	2.6	0.8
#novaccine	13	2.6	0.8
#populationcontrol	13	2.6	1.02
#saynotovaccines	13	2.6	1.36
#vaccinefailure	13	2.6	0.8
#notovaccines	13	2.6	0.8
#justsaynotovaccines	13	2.6	0.8
#microchip	13	2.6	1.02
#billgatesisevil	13	2.6	0.8
#vaccineinjuries	13	2.6	1.02
#sonotoneedles	13	2.6	1.02
#noforcedvaccines	12	2.4	0.49
#nocovidvaccine	12	2.4	1.2
#boycottcorona	12	2.4	0.8
#vaccineinjury	12	2.4	0.8
kung flu	12	2.4	1.02
#vaccinesarenottheanswer	12	2.4	1.62
#cdcfraud	12	2.4	0.49
#novax	12	2.4	1.02
#bigpharmakills	12	2.4	0.8
#thegreatreset	12	2.4	1.02
#novaccineforme	12	2.4	0.8

#vaccineagenda	12	2.4	1.02
#stopmandatoryvaccination	12	2.4	1.02
#exposebillgates	12	2.4	1.5
#abolishbigpharma	12	2.4	0.8
#billgatesevil	12	2.4	1.36
#covidhoax	12	2.4	0.8
#microchipping	12	2.4	1.36
#plandemic	12	2.4	1.36
#unvaccinatedlivesmatter	12	2.4	1.02
#parentsoverpharma	11	2.2	0.4
#forcedvaccines	11	2.2	1.17
#wuhanvirus	11	2.2	1.17
#educateb4uvax	11	2.2	1.17
#nocoronavaccine	11	2.2	1.17
#vaccinescause	11	2.2	1.33
#wedonotconsent	11	2.2	0.98
#covid1984	11	2.2	0.75
#nomandatoryvaccinations	11	2.2	0.98
#nomasktoday	11	2.2	0.4
#notomandatoryvaccinations	11	2.2	0.98
#bigpharmafia	11	2.2	1.47
#notomandatoryvaccines	11	2.2	0.4
#cdcwhistleblower	10	2.0	0.63
#noforcedflushots	10	2.0	0.63
#cdctruth	10	2.0	1.1
covidiots	10	2.0	0.63
#covidiots	10	2.0	0.63
#thegreatawakening	10	2.0	1.1

#infertility	10	2.0	1.41
#fuckthesystem	10	2.0	1.67
wuhan virus	10	2.0	0.89
#fightforyourrights	9	1.8	0.75
#wakeupworld	9	1.8	0.4
#novaccinemandates	9	1.8	0.75
#fuckcorona	9	1.8	1.33
#billgatesisnotadoctor	9	1.8	0.75
#momsofunvaccinatedchildren	9	1.8	1.17
#nonewnormal	8	1.6	1.02
#vaccinationchoice	8	1.6	0.49
#doctorsspeakup	8	1.6	0.8
#vaccinemandate	8	1.6	1.02
#medicalfreedom	8	1.6	0.8
#standupforyourrights	8	1.6	0.8
#idonotconsent	8	1.6	0.8
#medicalfreedomofchoice	7	1.4	0.49
#mybodymychoice	7	1.4	1.02
#unvaccinated	7	1.4	1.02
#propaganda	7	1.4	0.8
#billgatesvaccine	7	1.4	1.5
#pharmalobby	7	1.4	0.8
#v4vglobaldemo	7	1.4	1.36
#bigpharma	7	1.4	1.02
#idonotcomply	7	1.4	0.8
#wakeupcall	6	1.2	0.98
#ncov2019	6	1.2	1.47
#nocivilwar	6	1.2	0.75

#freepeople	6	1.2	0.4
#coronawarriors	6	1.2	0.4
ncov2019	6	1.2	1.47
#informedconsent	6	1.2	0.75
#parentalrights	6	1.2	0.98
ncov	5	1.0	1.1
#betweenmeandmydoctor	5	1.0	0.89
herd immunity	5	1.0	1.26
flattening the curve	5	1.0	0.63
#mainstreammedia	5	1.0	0.63
#learntherisk	5	1.0	0.89
#wakeup	5	1.0	0.89
#flatteningthecurve	5	1.0	0.63
#herdimmunity	5	1.0	0.89
flatten the curve	5	1.0	0.63
#flattenthecurve	5	1.0	0.63
2019-ncov	4	0.8	1.17
#faceshields	4	0.8	0.75
#freedomofspeech	4	0.8	0.75
#masks4all	4	0.8	0.75
#faceshield	4	0.8	0.75
wear a mask	3	0.6	0.49
#vaxxed	3	0.6	0.49
#billgates	3	0.6	1.2
#stayhome	3	0.6	0.49
wearamask	3	0.6	0.49
#lockdown	3	0.6	0.49
#makethisgoviral	3	0.6	0.49

#socialdistancing	3	0.6	0.49
#humanrights	3	0.6	0.49
#stayathome	3	0.6	0.49
social distancing	3	0.6	0.49
#quarantine	3	0.6	0.49
#ncov	3	0.6	0.8
#2019-ncov	2	0.4	0.49
face shields	2	0.4	0.49
#yeht	2	0.4	0.8
#selfisolating	2	0.4	0.49
#wearamask	2	0.4	0.49
#homeschooling	2	0.4	0.49
#2019ncov	2	0.4	0.49
sars cov 2	2	0.4	0.8
#hometasking	2	0.4	0.49
lockdown	2	0.4	0.49
#homeschool	2	0.4	0.49
self isolating	2	0.4	0.49
quarantine	2	0.4	0.49
#stayhomestaysafe	2	0.4	0.49
#vaccines	1	0.2	0.4
#workingfromhome	1	0.2	0.4
#frontlineheroes	1	0.2	0.4
#workfromhome	1	0.2	0.4
work from home	1	0.2	0.4
#echo	1	0.2	0.4
#wfh	1	0.2	0.4
n95	1	0.2	0.4

#healthworkers	1	0.2	0.4
2019ncov	1	0.2	0.4
sars cov2	1	0.2	0.4
#coronavaccine	1	0.2	0.4
#covid-19	1	0.2	0.4
#vaccine	1	0.2	0.4
#coronavaccines	1	0.2	0.4
ppe	1	0.2	0.4
vaccines	1	0.2	0.4
#ppe	1	0.2	0.4
corona vaccines	1	0.2	0.4
vaccine	1	0.2	0.4
corona vaccine	1	0.2	0.4
#n95	1	0.2	0.4
sarscov2	1	0.2	0.4
#sarscov2	1	0.2	0.4
face shield	1	0.2	0.4
working from home	1	0.2	0.4
#handsanitizer	0	0.0	0.0
health worker	0	0.0	0.0
wash ur hands	0	0.0	0.0
pneumonia	0	0.0	0.0
#coronaupdate	0	0.0	0.0
covid	0	0.0	0.0
hand sanitizer	0	0.0	0.0
#community	0	0.0	0.0
#pandemic	0	0.0	0.0
#washurhands	0	0.0	0.0

covid-19	0	0.0	0.0
wash your hands	0	0.0	0.0
#corona	0	0.0	0.0
coronavirus	0	0.0	0.0
health workers	0	0.0	0.0
#washyourhands	0	0.0	0.0
#covid	0	0.0	0.0
#healthworker	0	0.0	0.0
#coronavirus	0	0.0	0.0
pandemic	0	0.0	0.0
#covid_19	0	0.0	0.0
#covid19	0	0.0	0.0
corona	0	0.0	0.0
covid19	0	0.0	0.0
covid_19	0	0.0	0.0
#pneumonia	0	0.0	0.0

Table 7.7: The aggregated results of the Likert scale of $N = 5$ participants reporting the score total, mean and standard deviation of each term.

F Chapter 6

F.1 Case Study 1

Subreddits Used

- **PFM**: COVID19, ncovshills, cvnews, epidemic, Real_Coronavirus, CoronaVirusFFA, Wuhan_Flu, ID_News, VirusOutbreak, CoronavirusUncensored, corovirusdata, novel_coronavirus, CoronavirusUK, CoronavirusGLOBAL, Covid2019,

COVID19_support, 2020WuhanVirus, China_Flu, nCoronaVirus, 2019COVID, Coronavirus, CoronavirusFOS

- **Ask**: AskWomen, AskCulinary, AskDrugs, AskFeminists, AskLiteraryStudies, TrueAskReddit, AskMusic, AskModerators, AskStatistics, AskEngineers, AskDad, AskComputerScience, AskHistorians, AskLosAngeles, AskSeddit, AskAcademia, AskPhotography, AskScience, AskReddit, AskSciTech, AskUk, AskMen, AskTransgender, AskGSM, AskElectronics, AskArt, Ask_OfReddit, AskPhilosophy, AskSocialScience

- **New**: OnlyFun4U, BBC4BBWS, breadboysyt, bfatIRL, SpecialHumor, sims2help, Adultcontentcreators, NativePlantGardening, YourWellnessNerd, TradeAnalyzerFF, JuliaBayonetta, SheismichaelaNSFW, Jord627_, AllSaintsStreet, moreplatesmoredates, HUEstation, KiryuCoco, WallStreetbetsELITE, Cartooncat, USAHotGirls, VALORANT, delta8, ImaginaryAnthro, TopPops, skamtebord, InfluencergossipDK, IndianStreetBets, onlyfansbros, CatfishMePlease, TheWildAtHeart, onlyfanschicks, RedditMasterClasses, Dodocodes, oldhagfashion, SRGroup, MeatoSubincision, SatoshiBets, Promote_Your_YouTube, IPTVresell, onlyfansgirls101, WKHS, naughtychicks, Wallstreetbetsnew, Mya_For_The_Queen_, HeroWarsFB, Spudmode, quarantineactivities, ACVillager, assettopirate, Tifax-Aerith, LegendofthePhoenix, TheYouShow, PokeMeow, MLFBprospringfootball, BigBoobsAndAssess, EquityResearchIndia, OnlyfansXXX, DankExchange, AMDLaptops, USTravelBan, xxxycelebs, VictoriasecretGW, AmateurGoneWildPlus, CruelSummer, TgirlHUB, yeagerbomb, onlyfans_get_noticed, Naveljunkies, OnlyFans_Amateurs, ExtremelyHairyWomen, Desihub, mummytummies, Cross_Trading_Roblox, OnlyFansAsstastic, RedditPregunta, Epicentr, Teenpussyx, TLAUNCHER, CaliforniaJobsForAll, Helltaker, buksebule, confidentlyincorrect, AlabamaJobs, Life360, MedicineCommunity, CPTSDFightMode, PPPLoans, BlackOnlyFun, AdoptmyACNLvillagers, Bugsnax, CODWarzone, exfds, DirtySocialMedia

- **Random**: bodybuilding, drunk, summonerswar, googlehome, rails, oscp, brave-

frontier, Doom, Buddhism, germany, WeWantPlates, lepin, harrystyles, 3amjokes, hearthstone, Bedbugs, BravoRealHousewives, PUBGMobile, massachusetts, csgomarketforum, Sat, NoPoo, Vinesauce, talesfromcallcenters, shameless, datascience, NelkFilmz, Citrix, btc, Swimming, Buffalo, VitaPiracy, statistics, uofm, Shitty_Car_Mods, Cloud9, NFL_Draft, 8bitdo, Warhammer, outwardgame, minnesota, BoneAppleTea, greentext, zoombackgrounds, MordekaiserMains, girlsfrontline, TNOmod, JRPG, SuperMegaBaseball, M1Finance, botw, deadbydaylight, APUSH, funimation, canadacordcutters, queensuniversity, PHP, drums, mtgfinance, Miata, silenthill, TurkeyJerky, bangtan, TheMidnightGospel, manhwa, RATS, Choices, uruguay, Switzerland, rpghorrorstories, italy, findapath, 23andme, virginvschad, XboxSeriesX, zelda, InternetStars, JohnMayer, PiratedGTA, Nikon, Lexus, Soundbars, TheMonkeysPaw, stimuluscheck, lastimages, TrueCrime, SuperModelIndia, Machinists, Galaxy_S20, fixit, whichbike, IllegallySmolCats, SkyGame, Suomi, ElectricSkateboarding, starwarsrebels, nuzlocke, thedavidpakmanshow, TryingForABaby

Prediction Results

Origin	Terms
[29]	#novax, #infertility, #sonotoneedles, #notomandatoryvaccinations, #arrestfauci, #scamdemic, #plandemic, #vaccinemandate, #justsaynotovaccines, #nocivilwar, #vaccinefraud, #microchip, #community, #vaccinemicrochips, #nocoronavaccine, #exposebillgates, #covidhoax, #covid19hoax, #nocovidvaccine, #idonotcomply, #idonotconsent, #thegreatawakening, #novaccinemandates, #wakeupcall, #wedonotconsent, #wakeupworld, #nomandatoryvaccinations, #novaccine, #donttakethemicrochip, #makethisgoviral, #nonewnormal, #echo, #coronavirusfraud, #microchipping, #wakeup, #notovaccines, #scaredemic, #covid1984, #populationcontrol, #arrestbillgates
[274]	#bigpharmafia, #learntherisk, #momsofunvaccinatedchildren, #billgatesvaccine, #vaccineharm, #billgatesisnotadoctor, #cdctruth, #vaxxed, #vaccinefraud, #forcedvaccines, #cdcwhistleblower, #notomandatoryvaccines, #syringeslaughter, #informedconsent, #fuckvaccines, #novaccineforme, #abolishbigpharma, #exposebillgates, #billgatesbioterrorist, #vaccineagenda, #vaccineskill, #unvaccinated, #betweenmeandmydoctor, #idonotconsent, #cdcfraud, #billgatesisevil, #antivaccine, #novaccinemandates, #vaccinesarepoison, #mybodymychoice, #billgatesevil, #bigpharmakills, #saynotovaccines, #vaccinescause, #novaccine, #covidvaccineispoison, #noforcedvaccines, #vaccinesarenottheanswer, #depopulation, #stopmandatoryvaccination, #vaccinefailure, #noforcedflushots, #v4vglobaldemo, #yeht, #medicalfreedomofchoice, #vaccineinjuries, #parentsoverpharma, #parentalrights, #vaccinationchoice, #medicalfreedom, #educateb4uvax, #vaccinedamage, #vaccineinjury, #doctorsspeakup, #arrestbillgates
[232]	#ncov, lockdown, #workingfromhome, #homeschooling, #corona, sars cov 2, ncov2019, 2019ncov, #frontlineheroes, sarscov2, #wfh, social distancing, wearamask, face shield, working from home, #washyourhands, wuhan virus, #chinesevirus, covid19, #coronavaccine, #socialdistancing, #washurhands, corona vaccines, #herdimmunity, covid, #healthworkers, #covid, #kungflu, #pneumonia, n95, herd immunity, #n95, #sarscov2, self isolating, #wuhanvirus, vaccine, covid-19, ppe, face shields, kung flu, #coronavirus, #lockdown, work from home, corona, #coronaupdate, #2019-ncov, corona vaccine, sars cov2, 2019-ncov, wear a mask, #coronawarriors, #hometasking, #stayathome, #flattenthecurve, health worker, #2019ncov, #masks4all, #stayhomestaysafe, wash ur hands, #covidiots, coronavirus, #homeschool, flattening the curve, #faceshield, #stayhome, health workers, covidiots, hand sanitizer, #ncov2019, #workfromhome, quarantine, ncov, vaccines, pandemic, #vaccine, #flatteningthecurve, covid_19, #faceshields, #healthworker, #selfisolating, #coronavaccines, chinese virus, #covid-19, #quarantine, flatten the curve, #covid19, #covid_19, #pandemic, wash your hands, #handsanitizer, #vaccines, pneumonia, #ppe, #wearamask
Custom	#boycottcorona, #bigpharma, #thegreatreset, #billgates, #fuckthesystem, #vaxxed, #freepeople, #standupforyourrights, #fightforyourrights, #nomasktoday, #unvaccinatedlivesmatter, #fuckcorona, #freedomofspeech, #nonewnormal, #pharmalobby, #mainstreammedia, #wakeup, #humanrights, #covid1984, #propaganda

Table 7.6: Complete list of all hashtags / keywords used as part of the investigation..

Label	Terms
Controversial	#learntherisk, #thegreatawakening, #saynotovaccines, #plandemic, #justsaynotovaccines, #wakeup, #ncov2019, #vaccinesarenottheanswer, #nomasktoday, #billgatesbioterrorist, #vaccinescause, #unvaccinatedlivesmatter, #nomandatoryvaccinations, #billgatesisevil, #vaccinationchoice, #vaccinemicrochips, #covidiots, #vaccinesarepoison, #covidvaccineispoison, #bigpharmafia, #nonewnormal, ncov, #fuckcorona, #microchip, #wedonotconsent, #informedconsent, herd immunity, #notomandatoryvaccinations, #freepeople, #vaccinefraud, #nocoronavaccine, #covidhoax, flatten the curve, #coronavirusfraud, #covid1984, #cdctruth, #fuckthesystem, #antivaccine, #infertility, #medicalfreedom, #fightforyourrights, #bigpharma, #thegreatreset, #momsofunvaccinatedchildren, #wakeupcall, #bigpharmakills, #notomandatoryvaccines, #populationcontrol, #idonotcomply, #scamdemic, #donttakethemicrochip, #arrestfauci, #wuhanvirus, #flatteningthecurve, #coronawarriors, #vaccinefailure, #cdcfraud, #propaganda, flattening the curve, #billgatesevil, #vaccineskill, kung flu, #idonotconsent, #doctorsspeakup, #billgatesisnotadoctor, #vaccineinjury, #chinesevirus, #boycottcorona, #vaccinedamage, #arrestbillgates, #microchipping, #parentalrights, #syringeslaughter, #standupforyourrights, #exposebillgates, #flattenthecurve, #parentsoverpharma, #v4vglobaldemo, #vaccineharm, #novaccineforme, #nocivilwar, #novax, ncov2019, #vaccineagenda, #noforcedflushots, #wakeupworld, #betweenmeandmydoctor, #pharmalobby, chinese virus, #herdimmunity, #mybodymychoice, #vaccinemandate, #kungflu, #novaccine, #nocovidvaccine, #depopulation, #medicalfreedomofchoice, wuhan virus, #vaccineinjuries, #fuckvaccines, #notovaccines, #covid19hoax, #abolishbigpharma, #noforcedvaccines, #mainstreammedia, #stopmandatoryvaccination, #forcedvaccines, #educateb4uvax, #novaccinemandates, #billgatesvaccine, #cdcwhistleblower, #unvaccinated, #sonotoneedles, #scaredemic, covidiots
Non-controversial	coronavirus, n95, covid_19, #humanrights, #faceshields, #ncov, #ppe, working from home, #covid19, health workers, sars cov 2, #yeht, #washyourhands, #echo, #workingfromhome, #billgates, #covid_19, face shield, #stayathome, sarscov2, covid19, sars cov2, lockdown, corona, work from home, #healthworkers, #n95, #coronavaccine, vaccine, #stayhomestaysafe, wash ur hands, #2019ncov, covid-19, #socialdistancing, #lockdown, #homeschooling, #stayhome, wash your hands, #masks4all, #coronaupdate, covid, #sarscov2, #vaccines, #quarantine, #covid-19, #coronavaccines, #2019-ncov, #maketh-isgoviral, vaccines, #frontlineheroes, #selfisolating, #community, #vaccine, hand sanitizer, wear a mask, pandemic, #healthworker, pneumonia, #pneumonia, #faceshield, wearamask, #handsanitizer, ppe, #hometasking, self isolating, corona vaccines, face shields, #homeschool, #covid, #vaxxed, #freedomofspeech, quarantine, #workfromhome, #washurhands, 2019ncov, corona vaccine, #wfh, #wearamask, #coronavirus, 2019-ncov, #corona, #pandemic, social distancing, health worker

Table 7.8: Complete list of terms grouped by classification label.

Metric	BLR	SVM	RFC
Mean Accuracy	0.677347	0.792449	0.864286
SD Accuracy	0.066987	0.030552	0.041260
Mean F1 Score	0.460567	0.745211	0.853024
SD F1 Score	0.176783	0.031276	0.043656
Mean Precision	0.831587	0.836975	0.833129
SD Precision	0.185315	0.065113	0.052828
Mean Recall	0.330909	0.673636	0.875455
SD Recall	0.147160	0.024377	0.047508
Mean Sensitivity	0.959630	0.889259	0.855185
SD Sensitivity	0.029674	0.051452	0.050538
Mean Specificity	0.330909	0.673636	0.875455
SD Specificity	0.147160	0.024377	0.047508
Mean Positive Predictive Value (PPV)	NaN	0.836975	0.833129
SD Positive Predictive Value (PPV)	NaN	0.065113	0.052828
Mean Negative Predictive Value (NPV)	0.641535	0.769555	0.894412
SD Negative Predictive Value (NPV)	0.050188	0.016824	0.038926

Table 7.9: Graphlet prediction results for PFM subreddits.

Metric	BLR	SVM	RFC
Mean Accuracy	0.573103	0.594483	0.677069
SD Accuracy	0.060866	0.053835	0.055326
Mean F1 Score	0.577603	0.662721	0.687415
SD F1 Score	0.059650	0.049826	0.057733
Mean Precision	0.572328	0.568504	0.665524
SD Precision	0.058992	0.041630	0.052569
Mean Recall	0.584483	0.802759	0.713103
SD Recall	0.067109	0.102534	0.074578
Mean Sensitivity	0.561724	0.386207	0.641034
SD Sensitivity	0.075279	0.112162	0.063788
Mean Specificity	0.584483	0.802759	0.713103
SD Specificity	0.067109	0.102534	0.074578
Mean Positive Predictive Value (PPV)	0.572328	0.568504	0.665524
SD Positive Predictive Value (PPV)	0.058992	0.041630	0.052569
Mean Negative Predictive Value (NPV)	0.574485	0.678635	0.693155
SD Negative Predictive Value (NPV)	0.064786	0.118185	0.063627

Table 7.10: Graphlet prediction results for Ask subreddits.

Metric	BLR	SVM	RFC
Mean Accuracy	0.893050	0.923150	0.961100
SD Accuracy	0.010557	0.012097	0.007334
Mean F1 Score	0.885511	0.926093	0.962094
SD F1 Score	0.011644	0.012178	0.006946
Mean Precision	0.952498	0.891871	0.939142
SD Precision	0.013654	0.019278	0.012450
Mean Recall	0.827500	0.964000	0.986300
SD Recall	0.015516	0.027532	0.005226
Mean Sensitivity	0.958600	0.882300	0.935900
SD Sensitivity	0.012330	0.024934	0.013935
Mean Specificity	0.827500	0.964000	0.986300
SD Specificity	0.015516	0.027532	0.005226
Mean Positive Predictive Value (PPV)	0.952498	0.891871	0.939142
SD Positive Predictive Value (PPV)	0.013654	0.019278	0.012450
Mean Negative Predictive Value (NPV)	0.847619	0.962011	0.985603
SD Negative Predictive Value (NPV)	0.011891	0.026861	0.005404

Table 7.11: Graphlet prediction results for New subreddits.

Metric	BLR	SVM	RFC
Mean Accuracy	7.755102e-01	8.302041e-01	0.882245
SD Accuracy	3.330669e-16	1.629589e-02	0.020379
Mean F1 Score	6.841905e-01	7.832142e-01	0.868471
SD F1 Score	5.167490e-03	1.599712e-02	0.023026
Mean Precision	9.292308e-01	9.212255e-01	0.870916
SD Precision	2.086871e-02	4.315329e-02	0.023729
Mean Recall	5.418182e-01	6.818182e-01	0.866364
SD Recall	1.233151e-02	2.220446e-16	0.027887
Mean Sensitivity	9.659259e-01	9.511111e-01	0.895185
SD Sensitivity	1.004790e-02	2.957402e-02	0.020323
Mean Specificity	5.418182e-01	6.818182e-01	0.866364
SD Specificity	1.233151e-02	2.220446e-16	0.027887
Mean Positive Predictive Value (PPV)	9.292308e-01	9.212255e-01	0.870916
SD Positive Predictive Value (PPV)	2.086871e-02	4.315329e-02	0.023729
Mean Negative Predictive Value (NPV)	7.212865e-01	7.856704e-01	0.891793
SD Negative Predictive Value (NPV)	3.173020e-03	5.375983e-03	0.021321

Table 7.12: Graphlet prediction results for PFM vs Ask subreddits.

Metric	BLR	SVM	RFC
Mean Accuracy	0.666939	0.793469	0.847551
SD Accuracy	0.049813	0.049427	0.026524
Mean F1 Score	0.540839	0.790091	0.822059
SD F1 Score	0.086942	0.051439	0.029424
Mean Precision	0.703240	0.731279	0.865912
SD Precision	0.080590	0.063025	0.043399
Mean Recall	0.444091	0.870455	0.783636
SD Recall	0.093122	0.098202	0.032853
Mean Sensitivity	0.848519	0.730741	0.899630
SD Sensitivity	0.044474	0.091453	0.036400
Mean Specificity	0.444091	0.870455	0.783636
SD Specificity	0.093122	0.098202	0.032853
Mean Positive Predictive Value (PPV)	0.703240	0.731279	0.865912
SD Positive Predictive Value (PPV)	0.080590	0.063025	0.043399
Mean Negative Predictive Value (NPV)	0.653767	0.883523	0.836411
SD Negative Predictive Value (NPV)	0.041225	0.076711	0.023180

Table 7.13: Graphlet prediction results for PFM vs New subreddits.

Metric	BLR	SVM	RFC
Mean Accuracy	0.910345	0.947414	0.938621
SD Accuracy	0.033343	0.008577	0.024422
Mean F1 Score	0.916716	0.944772	0.937064
SD F1 Score	0.028057	0.009047	0.024812
Mean Precision	0.867796	0.994276	0.961653
SD Precision	0.054369	0.014777	0.029025
Mean Recall	0.974138	0.900345	0.914138
SD Recall	0.015708	0.018178	0.028641
Mean Sensitivity	0.846552	0.994483	0.963103
SD Sensitivity	0.073372	0.014400	0.028542
Mean Specificity	0.974138	0.900345	0.914138
SD Specificity	0.015708	0.018178	0.028641
Mean Positive Predictive Value (PPV)	0.867796	0.994276	0.961653
SD Positive Predictive Value (PPV)	0.054369	0.014777	0.029025
Mean Negative Predictive Value (NPV)	0.971285	0.909261	0.918486
SD Negative Predictive Value (NPV)	0.017669	0.015529	0.026614

Table 7.14: Graphlet prediction results for Ask vs New subreddits.

Metric	BLR	SVM	RFC
Mean Accuracy	0.654227	0.643126	0.851116
SD Accuracy	0.047716	0.054630	0.041896
Mean F1 Score	0.660762	0.647285	0.854024
SD F1 Score	0.036657	0.042211	0.041231
Mean Precision	0.651952	0.646075	0.838903
SD Precision	0.054786	0.066341	0.050351
Mean Recall	0.671818	0.653636	0.871818
SD Recall	0.029162	0.050526	0.049950
Mean Sensitivity	0.636700	0.632787	0.830395
SD Sensitivity	0.083369	0.107988	0.060411
Mean Specificity	0.671818	0.653636	0.871818
SD Specificity	0.029162	0.050526	0.049950
Mean Positive Predictive Value (PPV)	0.651952	0.646075	0.838903
SD Positive Predictive Value (PPV)	0.054786	0.066341	0.050351
Mean Negative Predictive Value (NPV)	0.658219	0.644607	0.868011
SD Negative Predictive Value (NPV)	0.042705	0.049602	0.044486

Table 7.15: Global feature prediction results for PFM subreddits.

Metric	BLR	SVM	RFC
Mean Accuracy	0.752320	0.729994	0.750444
SD Accuracy	0.039178	0.044819	0.050748
Mean F1 Score	0.781529	0.770427	0.764137
SD F1 Score	0.030536	0.030722	0.041858
Mean Precision	0.700392	0.673178	0.727952
SD Precision	0.040198	0.046165	0.057747
Mean Recall	0.885172	0.903103	0.806207
SD Recall	0.024892	0.019320	0.035428
Mean Sensitivity	0.620023	0.557724	0.694931
SD Sensitivity	0.066762	0.089088	0.083976
Mean Specificity	0.885172	0.903103	0.806207
SD Specificity	0.024892	0.019320	0.035428
Mean Positive Predictive Value (PPV)	0.700392	0.673178	0.727952
SD Positive Predictive Value (PPV)	0.040198	0.046165	0.057747
Mean Negative Predictive Value (NPV)	0.843439	0.850940	0.781313
SD Negative Predictive Value (NPV)	0.034998	0.031749	0.044294

Table 7.16: Global feature prediction results for Ask subreddits.

Metric	BLR	SVM	RFC
Mean Accuracy	0.856791	0.904920	0.919840
SD Accuracy	0.019584	0.012780	0.010216
Mean F1 Score	0.855728	0.908995	0.924333
SD F1 Score	0.019649	0.011571	0.009393
Mean Precision	0.866938	0.876939	0.879953
SD Precision	0.020933	0.018413	0.013673
Mean Recall	0.844894	0.943723	0.973617
SD Recall	0.020307	0.010125	0.011700
Mean Sensitivity	0.868817	0.865699	0.865484
SD Sensitivity	0.021398	0.022733	0.017438
Mean Specificity	0.844894	0.943723	0.973617
SD Specificity	0.020307	0.010125	0.011700
Mean Positive Predictive Value (PPV)	0.866938	0.876939	0.879953
SD Positive Predictive Value (PPV)	0.020933	0.018413	0.013673
Mean Negative Predictive Value (NPV)	0.847188	0.938402	0.970279
SD Negative Predictive Value (NPV)	0.019566	0.010567	0.012685

Table 7.17: Global feature prediction results for New subreddits.

Metric	BLR	SVM	RFC
Mean Accuracy	0.844624	0.850121	0.886028
SD Accuracy	0.039900	0.039865	0.038112
Mean F1 Score	0.884763	0.878343	0.912165
SD F1 Score	0.028106	0.032561	0.029950
Mean Precision	0.846590	0.915432	0.898462
SD Precision	0.045386	0.048242	0.040680
Mean Recall	0.929545	0.846364	0.927273
SD Recall	0.042336	0.043016	0.030829
Mean Sensitivity	0.679384	0.849499	0.810676
SD Sensitivity	0.155103	0.123040	0.082326
Mean Specificity	0.929545	0.846364	0.927273
SD Specificity	0.042336	0.043016	0.030829
Mean Positive Predictive Value (PPV)	0.846590	0.915432	0.898462
SD Positive Predictive Value (PPV)	0.045386	0.048242	0.040680
Mean Negative Predictive Value (NPV)	NaN	0.751909	0.863737
SD Negative Predictive Value (NPV)	NaN	0.088869	0.049020

Table 7.18: Global feature prediction results for PFM vs Ask subreddits.

Metric	BLR	SVM	RFC
Mean Accuracy	0.684572	0.664609	0.908104
SD Accuracy	0.063011	0.067679	0.036129
Mean F1 Score	0.765947	0.761987	0.918894
SD F1 Score	0.034403	0.035630	0.031361
Mean Precision	0.656142	0.633654	0.900343
SD Precision	0.049897	0.052448	0.042637
Mean Recall	0.925909	0.961364	0.939545
SD Recall	0.048810	0.033633	0.035241
Mean Sensitivity	0.376018	0.290885	0.868509
SD Sensitivity	0.187424	0.182232	0.065086
Mean Specificity	0.925909	0.961364	0.939545
SD Specificity	0.048810	0.033633	0.035241
Mean Positive Predictive Value (PPV)	0.656142	0.633654	0.900343
SD Positive Predictive Value (PPV)	0.049897	0.052448	0.042637
Mean Negative Predictive Value (NPV)	NaN	NaN	0.921435
SD Negative Predictive Value (NPV)	NaN	NaN	0.043658

Table 7.19: Global feature prediction results for PFM vs New subreddits.

Metric	BLR	SVM	RFC
Mean Accuracy	9.195431e-01	0.948298	0.937001
SD Accuracy	2.425640e-02	0.020217	0.022099
Mean F1 Score	9.318403e-01	0.954787	0.943523
SD F1 Score	1.850276e-02	0.017036	0.019267
Mean Precision	9.010704e-01	0.949901	0.958464
SD Precision	3.441313e-02	0.030859	0.027908
Mean Recall	9.655172e-01	0.960345	0.929655
SD Recall	1.110223e-16	0.014113	0.023851
Mean Sensitivity	8.588389e-01	0.932327	0.946216
SD Sensitivity	5.798061e-02	0.043500	0.037929
Mean Specificity	9.655172e-01	0.960345	0.929655
SD Specificity	1.110223e-16	0.014113	0.023851
Mean Positive Predictive Value (PPV)	9.010704e-01	0.949901	0.958464
SD Positive Predictive Value (PPV)	3.441313e-02	0.030859	0.027908
Mean Negative Predictive Value (NPV)	9.499837e-01	0.947613	0.912277
SD Negative Predictive Value (NPV)	5.837108e-03	0.018189	0.028947

Table 7.20: Global feature prediction results for Ask vs New subreddits.

Bibliography

[1] Individuals using the Internet (% of population) | Data. URL https://data.worldbank.org/indicator/it.net.user.zs.

[2] Most used social media 2021. URL https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

[3] Number of social media users 2025. URL https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/.

[4] Daron Acemoglu, Asuman Ozdaglar, and Ali ParandehGheibi. Spread of (mis)information in social networks. *Games and Economic Behavior*, 70(2):194–227, 2010. ISSN 0899-8256.

[5] Lada A. Adamic, Jun Zhang, Eytan Bakshy, and Mark S. Ackerman. Knowledge sharing and yahoo answers: Everyone knows something. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, page 665674, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605580852. doi: 10.1145/1367497.1367587. URL https://doi.org/10.1145/1367497.1367587.

[6] B Thomas Adler, Luca De Alfaro, Santiago M Mola-Velasco, Paolo Rosso, and Andrew G West. Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 277–288. Springer, 2011.

[7] Ronak Agrawal and Dilip Kumar Sharma. A survey on video-based fake news detection techniques. In *2021 8th International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 663–669. IEEE, 2021.

[8] Wasim Ahmed, Josep Vidal-Alaball, Joseph Downing, and Francesc López Seguí. Covid-19 and the 5g conspiracy theory: social network analysis of twitter data. *Journal of Medical Internet Research*, 22(5):e19458, 2020.

[9] Hyerim Ahn and Ji-Hong Park. The structural effects of sharing function on twitter networks: Focusing on the retweet function. *Journal of Information Science*, 41(3):354–365, 2015.

[10] Badreya Al-Jenaibi. The nature of arab public discourse: Social media and the arab spring. *Journal of Applied Journalism & Media Studies*, 3(2):241–260, 2014.

[11] Faisal Alatawi, Lu Cheng, Anique Tahir, Mansooreh Karami, Bohan Jiang, Tyler Black, and Huan Liu. A survey on echo chambers on social media: Description, detection and mitigation. *arXiv preprint arXiv:2112.05084*, 2021.

[12] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[13] Kholoud Khalil Aldous, Jisun An, and Bernard J Jansen. View, like, comment, post: Analyzing user engagement by topic at 4 levels across 5 social media platforms for 53 news organizations. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 47–57, 2019.

[14] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–36, 2017.

[15] Hunt Allcott, Matthew Gentzkow, and Chuan Yu. Trends in the diffusion of misinformation on social media. *Research & Politics*, 6(2):2053168019848554, 2019.

[16] Jennifer Allen, Baird Howland, Markus Mobius, David Rothschild, and Duncan J Watts. Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, 6(14):eaay3539, 2020. Publisher: American Association for the Advancement of Science.

[17] Dimosthenis Antypas, Jose Camacho-Collados, Alun Preece, and David Rogers. Covid-19 and misinformation: A large-scale lexical analysis on twitter. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 119–126, 2021.

[18] Sofía Aparicio, Javier Villazón-Terrazas, and Gonzalo Álvarez. A model for scale-free networks: application to twitter. *Entropy*, 17(8):5848–5867, 2015.

[19] Gil Appel, Lauren Grewal, Rhonda Hadi, and Andrew T. Stephen. The future of social media in marketing. *Journal of the Academy of Marketing Science*, 48 (1):79–95, January 2020. ISSN 1552-7824. doi: 10.1007/s11747-019-00695-1. URL `https://doi.org/10.1007/s11747-019-00695-1`.

[20] Valerio Arnaboldi, Marco Conti, Andrea Passarella, and Robin IM Dunbar. Online social networks and information diffusion: The role of ego networks. *Online Social Networks and Media*, 1:44–55, 2017.

[21] Yael Artzy-Randrup. Comment on" network motifs: Simple building. *science*, 1099334(1107c):305, 2004.

[22] James Ashford, Liam Turner, Roger Whitaker, Alun Preece, Diane Felmlee, and Don Towsley. Understanding the signature of controversial wikipedia articles through motifs in editor revision networks. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 1180–1187, 2019.

[23] James R Ashford, Liam D Turner, Roger M Whitaker, Alun Preece, and Diane Felmlee. Assessing temporal and spatial features in detecting disruptive users on reddit. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 892–896. IEEE, 2020.

[24] James R Ashford, Liam D Turner, Roger M Whitaker, Alun Preece, and Diane Felmlee. Understanding the characteristics of covid-19 misinformation communities through graphlet analysis. *Online Social Networks and Media*, page 100178, 2021.

[25] Dennis Assenmacher, Lena Clever, Janina Susanne Pohl, Heike Trautmann, and Christian Grimme. A two-phase framework for detecting manipulation campaigns in social media. In *International Conference on Human-Computer Interaction*, pages 201–214. Springer, 2020.

[26] Ngo Xuan Bach, Nguyen Do Hai, and Tu Minh Phuong. Personalized recommendation of stories for commenting in forum-based social media. *Information Sciences*, 352:48–60, 2016.

[27] Chongyang Bai, Srijan Kumar, Jure Leskovec, Miriam Metzger, Jay Nunamaker, and VS Subrahmanian. Predicting the visual focus of attention in multi-person discussion videos. In *IJCAI 2019*. International Joint Conferences on Artificial Intelligence, 2019.

[28] Timothy L Bailey, Charles Elkan, et al. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. 1994.

[29] Annalise Baines, Muhammad Ittefaq, and Mauryne Abwao. # scamdemic,# plandemic, or# scaredemic: what parler social media platform tells us about covid-19 vaccine. *Vaccines*, 9(5):421, 2021.

[30] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, pages 519–528, 2012.

[31] Harvir S Bansal, Shirley F Taylor, and Yannik St. James. migrating to new service providers: Toward a unifying framework of consumers switching behaviors. *Journal of the Academy of Marketing Science*, 33(1):96–115, 2005.

[32] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[33] Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature reviews genetics*, 5(2):101–113, 2004.

[34] Pablo Barberá, John T Jost, Jonathan Nagler, Joshua A Tucker, and Richard Bonneau. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science*, 26(10):1531–1542, 2015.

[35] Pablo Barberá, Ning Wang, Richard Bonneau, John T Jost, Jonathan Nagler, Joshua Tucker, and Sandra González-Bailón. The critical periphery in the growth of social protests. *PloS one*, 10(11):e0143611, 2015.

[36] Oana Barbu. Advertising, microtargeting and social media. *Procedia-Social and Behavioral Sciences*, 163:44–49, 2014.

[37] Zapan Barua, Sajib Barua, Salma Aktar, Najma Kabir, and Mingze Li. Effects of misinformation on covid-19 individual responses and recommendations for resilience of disastrous consequences of misinformation. *Progress in Disaster Science*, page 100119, 2020. ISSN 2590-0617. doi: https://doi.org/10.1016/j.pdisas.2020.100119. URL http://www.sciencedirect.com/science/article/pii/S2590061720300569.

[38] Marco Bastos and Dan Mercea. The public accountability of social platforms: Lessons from a study on bots and trolls in the brexit campaign. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2128):20180003, 2018.

[39] Vladimir Batagelj and Andrej Mrvar. A subquadratic triad census algorithm for large sparse networks with small maximum degree. *Social Networks*, 23(3):237–243, July 2001. ISSN 03788733. doi: 10.1016/S0378-8733(01)00035-1. URL http://linkinghub.elsevier.com/retrieve/pii/S0378873301000351.

[40] Natalya N Bazarova, Yoon Hyung Choi, Victoria Schwanda Sosik, Dan Cosley, and Janis Whitlock. Social sharing of emotions on facebook: Channel differences, satisfaction, and replies. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pages 154–164, 2015.

[41] Nesserine Benchettara, Rushed Kanawati, and Celine Rouveirol. Supervised machine learning applied to link prediction in bipartite social networks. In *2010 international conference on advances in social networks analysis and mining*, pages 326–330. IEEE, 2010.

[42] Kelly Bergstrom. dont feed the troll: Shutting down debate about community expectations on reddit. com. *First Monday*, 16(8), 2011.

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

[44] Alessandro Bessi, Fabiana Zollo, Michela Del Vicario, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Trend of narratives in the age of misinformation. *PloS one*, 10(8):e0134641, 2015.

[45] Z. Bin, Z. Gang, F. Yunbo, Z. Xiaolu, J. Weiqiang, D. Jing, and G. Jiafeng. Behavior analysis based sms spammer detection in mobile communication networks. In *2016 IEEE First International Conference on Data Science in Cyberspace (DSC)*, pages 538–543, 2016. doi: 10.1109/DSC.2016.48.

[46] Andreas Birkbak and Hjalmar Carlsen. The world of edgerank: Rhetorical justifications of facebook's news feed algorithm. *Computational Culture (5), Special Issue on Rhetoric and Computation*, 2016.

[47] Halil Bisgin, Nitin Agarwal, and Xiaowei Xu. A study of homophily on social media. *World Wide Web*, 15(2):213–232, 2012.

[48] Sophie Bishop. Anxiety, panic and self-optimization: Inequalities and the youtube algorithm. *Convergence*, 24(1):69–84, 2018.

[49] Ph Blanchard and Tyll Krüger. The cameo principle and the origin of scale-free graphs in social networks. *Journal of statistical physics*, 114(5):1399–1416, 2004.

[50] Per Block. Reciprocity, transitivity, and the mysterious three-cycle. *Social Networks*, 40:163–173, 2015.

[51] Leticia Bode. Political news in the news feed: Learning politics from social media. *Mass communication and society*, 19(1):24–48, 2016.

[52] Leticia Bode. Pruning the news feed: Unfriending and unfollowing political content on social media. *Research & Politics*, 3(3):2053168016661873, 2016.

[53] Ruth N. Bolton, A. Parasuraman, Ankie Hoefnagels, Nanne Migchels, Sertan Kabadayi, Thorsten Gruber, Yuliya Komarova Loureiro, and David Solnet. Understanding Generation Y and their use of social media: a review and research agenda. *Journal of Service Management*, 24(3):245–267, January 2013. ISSN 1757-5818. doi: 10.1108/09564231311326987. URL `https://doi.org/10.1108/09564231311326987`. Publisher: Emerald Group Publishing Limited.

[54] Stephen P Borgatti, Ajay Mehra, Daniel J Brass, and Giuseppe Labianca. Network analysis in the social sciences. *science*, 323(5916):892–895, 2009.

[55] Rosa Borge Bravo and Marc Esteve Del Valle. Opinion leadership in parliamentary twitter networks: A matter of layers of interaction? *Journal of Information Technology & Politics*, 14(3):263–276, 2017.

[56] Michael Bossetta. The digital architectures of social media: Comparing political campaigning on facebook, twitter, instagram, and snapchat in the 2016 us election. *Journalism & mass communication quarterly*, 95(2):471–496, 2018.

[57] Andrei Boutyline and Robb Willer. The Social Structure of Political Echo Chambers: Variation in Ideological Homophily in Online Networks: Political Echo Chambers. *Political Psychology*, 38(3):551–569, June 2017. ISSN 0162895X. doi: 10.1111/pops.12337. URL http://doi.wiley.com/10.1111/pops.12337.

[58] Paul Boyle, Halfacree Keith, et al. *Exploring contemporary migration*. Routledge, 2014.

[59] Daren C Brabham. Crowdsourcing as a Model for Problem Solving: An Introduction and Cases. *Convergence*, 14(1):75–90, 2008.

[60] Samantha Bradshaw and Philip Howard. Troops, trolls and troublemakers: A global inventory of organized social media manipulation. 2017.

[61] Dave Braines, Diane Felmlee, Don Towsley, Kun Tu, Roger M Whitaker, and Liam D Turner. The role of motifs in understanding behavior in social and engineered networks. In *Next-Generation Analyst VI*, volume 10653, page 106530W. International Society for Optics and Photonics, 2018.

[62] Ulrik Brandes, Patrick Kenis, Jürgen Lerner, and Denise Van Raaij. Network analysis of collaboration structure in wikipedia. In *Proceedings of the 18th international conference on World wide web*, pages 731–740. ACM, 2009.

[63] Verena K Brändle, Charlotte Galpin, and Hans-Jörg Trenz. Brexit as politics of division: Social media campaigning after the referendum. *Social Movement Studies*, 21(1-2):234–253, 2022.

[64] Daniel J Brass, Kenneth D Butterfield, and Bruce C Skaggs. Relationships and unethical behavior: A social network perspective. *Academy of management review*, 23(1):14–31, 1998.

[65] Marco Bressan, Flavio Chierichetti, Ravi Kumar, Stefano Leucci, and Alessandro Panconesi. Counting graphlets: Space vs time. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, page 557566, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450346757. doi: 10.1145/3018661.3018732. URL https://doi.org/10.1145/3018661.3018732.

[66] Marco Bressan, Flavio Chierichetti, Ravi Kumar, Stefano Leucci, and Alessandro Panconesi. Motif Counting Beyond Five Nodes. *ACM Transactions on Knowledge Discovery from Data*, 20(2):1–25, April 2018. ISSN 15564681. doi: 10.1145/3186586. URL http://dl.acm.org/citation.cfm?doid=3208362.3186586.

[67] Aengus Bridgman, Eric Merkley, Peter John Loewen, Taylor Owen, Derek Ruths, Lisa Teichmann, and Oleg Zhilin. The causes and consequences of covid-19 misperceptions: Understanding the role of news and social media. *Harvard Kennedy School Misinformation Review*, 1(3), 2020.

[68] Jonathan Bright, Nahema Marchal, Bharath Ganesh, and Stevan Rudinac. How do individuals in a radical echo chamber react to opposing views? evidence from a content analysis of stormfront. *Human Communication Research*, 48(1): 116–145, 2022.

[69] Duraisamy Brindha, R Jayaseelan, and S Kadeswaran. Social media reigned by information or misinformation about covid-19: a phenomenological study. 2020.

[70] Axel Bruns. Echo Chamber? What Echo Chamber? Reviewing the Evidence. page 12.

[71] Axel Bruns, Tim Highfield, and Jean Burgess. The arab spring and its social media audiences: English and arabic twitter users and their networks. In *Cyberactivism on the participatory web*, pages 96–128. Routledge, 2014.

[72] Axel Bruns, Brenda Moon, Felix Münch, and Troy Sadkowsky. The australian twittersphere in 2016: Mapping the follower/followee network. *Social Media+ Society*, 3(4):2056305117748162, 2017.

[73] Lauren Valentino Bryant. The youtube algorithm and the alt-right filter bubble. *Open Information Science*, 4(1):85–90, 2020.

[74] Taina Bucher and Anne Helmond. The affordances of social media platforms. 2017.

[75] Cody Buntain and Jennifer Golbeck. Identifying social roles in reddit using network structure. In *Proceedings of the 23rd international conference on world wide web*, pages 615–620, 2014.

[76] Anthony G Burton and Dimitri Koehorst. Research note: The spread of political misinformation on online subcultural platforms. *HKS Misinfo Rev*, 1(6): 10–37016, 2020.

[77] Jason Vincent A Cabañes. The imaginative dimension of digital disinformation: Fake news, political trolling, and the entwined crises of covid-19 and inter-asian racism in a postcolonial city. *International Journal of Cultural Studies*, page 13678779211068533, 2022.

[78] Guido Caldarelli, Rocco De Nicola, Marinella Petrocchi, Manuel Pratelli, and Fabio Saracco. Flow of online misinformation during the peak of the covid-19 pandemic in italy. *EPJ data science*, 10(1):34, 2021.

[79] Juan Cao, Peng Qi, Qiang Sheng, Tianyun Yang, Junbo Guo, and Jintao Li. Exploring the role of visual content in fake news detection. *Disinformation, Misinformation, and Fake News in Social Media*, pages 141–161, 2020.

[80] Qiang Cao, Michael Sirivianos, Xiaowei Yang, and Tiago Pregueiro. Aiding the detection of fake accounts in large scale social online services. In *Presented as part of the 9th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 12)*, pages 197–210, 2012.

[81] Qilin Cao, Yong Lu, Dayong Dong, Zongming Tang, and Yongqiang Li. The roles of bridging and bonding in social media communities. *Journal of the American Society for Information Science and Technology*, 64(8):1671–1681, 2013.

[82] Mark Carman, Mark Koerber, Jiuyong Li, Kim-Kwang Raymond Choo, and Helen Ashman. Manipulating visibility of political and apolitical threads on reddit via score boosting. In *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, pages 184–190. IEEE, 2018.

[83] Juan Antonio Carrasco, Bernie Hogan, Barry Wellman, and Eric J Miller. Collecting social network data to study social activity-travel behavior: an egocentric approach. *Environment and Planning B: Planning and Design*, 35(6):961–980, 2008.

[84] Ruei-Yuan Chang, Sheng-Lung Peng, Guanling Lee, and Chia-Jung Chang. Comparing group characteristics to explain community structures in social media networks. , 16(6):957–962, 2015.

[85] Mingxuan Chen, Xinqiao Chu, and KP Subbalakshmi. Mmcovar: multimodal covid-19 vaccine focused data repository for fake news detection and a baseline architecture for classification. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 31–38, 2021.

[86] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. Antisocial Behavior in Online Discussion Communities. page 10.

[87] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. Antisocial behavior in online discussion communities. *arXiv preprint arXiv:1504.00680*, 2015.

[88] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. page 14, 2017.

[89] Mingxi Cheng, Chenzhong Yin, Shahin Nazarian, and Paul Bogdan. Deciphering the laws of social network-transcendent covid-19 misinformation dynamics and implications for combating misinformation phenomena. *Scientific Reports*, 11 (1):1–14, 2021.

[90] Shuang Cheng, Sang-Joon Lee, and Beomjin Choi. An empirical investigation of users voluntary switching intention for mobile personal cloud storage services based on the push-pull-mooring framework. *Computers in Human Behavior*, 92: 198–215, 2019.

[91] Darko Cherepnalkoski and Igor Mozetič. Retweet networks of the european parliament: Evaluation of the community structure. *Applied network science*, 1 (1):1–20, 2016.

[92] Paul-Alexandru Chirita, Jörg Diederich, and Wolfgang Nejdl. Mailrank: using ranking for spam detection. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 373–380, 2005.

[93] Eugene Ch'ng. Local interactions and the emergence of a twitter small-world network. *arXiv preprint arXiv:1508.03594*, 2015.

[94] Harshita Chopra, Aniket Vashishtha, Ridam Pal, Ananya Tyagi, Tavpritesh Sethi, et al. Mining trends of covid-19 vaccine beliefs on twitter with lexical embeddings. *arXiv preprint arXiv:2104.01131*, 2021.

[95] Giovanni Luca Ciampaglia. Fighting fake news: a role for computational social science in the fight against digital misinformation. *Journal of Computational Social Science*, 1(1):147–153, 2018.

[96] Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. The covid-19 social media infodemic. *Scientific reports*, 10 (1):1–10, 2020.

[97] Elanor Colleoni, Alessandro Rozza, and Adam Arvidsson. Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of communication*, 64(2):317–332, 2014.

[98] Botambu Collins, Dinh Tuyen Hoang, Ngoc Thanh Nguyen, and Dosam Hwang. Trends in combating fake news on social media–a survey. *Journal of Information and Telecommunication*, 5(2):247–266, 2021.

[99] European Commission, Directorate-General for Research, and Innovation. *Tackling complexity in science : general integration of the application of complexity in science*. Publications Office, 2007.

[100] Francesca Comunello and Giuseppe Anzera. Will the revolution be tweeted? a conceptual framework for understanding the social media and the arab spring. *Islam and Christian–Muslim Relations*, 23(4):453–470, 2012.

[101] Noshir S Contractor and Eric M Eisenberg. Communication networks and new media in organizations. 1990.

[102] Stephen A Cook. The complexity of theorem-proving procedures. In *Proceedings of the third annual ACM symposium on Theory of computing*, pages 151–158, 1971.

[103] Alceu Ferraz Costa, Yuto Yamaguchi, Agma Juci Machado Traina, Caetano Traina Jr, and Christos Faloutsos. Modeling temporal activity to detect anomalous behavior in social media. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 11(4):1–23, 2017.

[104] Liviu-Adrian Cotfas, Camelia Delcea, Ioan Roxin, Corina Ioanăş, Dana Simona Gherai, and Federico Tajariol. The longest month: analyzing covid-19 vaccination opinions dynamics from tweets in the month following the first vaccine announcement. *Ieee Access*, 9:33203–33223, 2021.

[105] Jose Yunam Cuan-Baltazar, Maria José Muñoz-Perez, Carolina Robledo-Vega, Maria Fernanda Pérez-Zepeda, and Elena Soto-Vega. Misinformation of COVID-19 on the Internet: Infodemiology Study. *JMIR Public Health and Surveillance*, 6(2):e18444, 2020. doi: 10.2196/18444. URL `https://publichealth.jmir.org/2020/2/e18444/`. Company: JMIR Public Health and Surveillance Distributor: JMIR Public Health and Surveillance Institution: JMIR Public Health and Surveillance Label: JMIR Public Health and Surveillance Publisher: JMIR Publications Inc., Toronto, Canada.

[106] Limeng Cui and Dongwon Lee. Coaid: Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885*, 2020.

[107] Limeng Cui, Haeseung Seo, Maryam Tabar, Fenglong Ma, Suhang Wang, and Dongwon Lee. Deterrent: Knowledge guided graph attention network for detecting healthcare misinformation. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 492–502, 2020.

[108] Pádraig Cunningham, Martin Harrigan, Guangyu Wu, and Derrk O'Callaghan. Characterizing ego-networks using motifs. *Network Science*, 1(2):170190, 2013. doi: 10.1017/nws.2013.12.

[109] Philipp Darius and Michael Urquhart. Disinformed social movements: A large-scale mapping of conspiracy narratives as online harms during the covid-19 pandemic. *Online Social Networks and Media*, 26:100174, 2021.

[110] Cai Davies, James Ashford, Luis Espinosa-Anke, Alun Preece, Liam Turner, Roger Whitaker, Mudhakar Srivatsa, and Diane Felmlee. Multi-scale user migration on reddit. AAAI, 2021.

[111] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. BotOrNot: A System to Evaluate Social Bots. In *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion*, pages 273–274, Montréal, Québec, Canada, 2016. ACM Press. ISBN 978-1-4503-4144-8. doi: 10.1145/2872518.2889302. URL http://dl.acm.org/citation.cfm?doid=2872518.2889302.

[112] Nicollas R de Oliveira, Pedro S Pisa, Martin Andreoni Lopez, Dianne Scherly V de Medeiros, and Diogo MF Mattos. Identifying fake news on social networks based on natural language processing: trends and challenges. *Information*, 12(1): 38, 2021.

[113] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559, January 2016. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1517441113. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1517441113.

[114] Sarah Jane Delany, Mark Buckley, and Derek Greene. Sms spam filtering: Methods and data. *Expert Systems with Applications*, 39(10):9899–9908, 2012.

[115] Matthew R DeVerna, Francesco Pierri, Bao Tran Truong, John Bollenbacher, David Axelrod, Niklas Loynes, Christopher Torres-Lugo, Kai-Cheng Yang, Filippo Menczer, and John Bryden. Covaxxy: A collection of english-language twitter posts about covid-19 vaccines. In *Proceedings of the AAAI international conference on web and social media (ICWSM)*, 2021.

[116] Michael A DeVito. From editors to algorithms: A values-based approach to understanding story selection in the facebook news feed. *Digital journalism*, 5 (6):753–773, 2017.

[117] Riccardo Di Clemente, Miguel Luengo-Oroz, Matias Travizano, Sharon Xu, Bapu Vaitla, and Marta C González. Sequences of purchases in credit card data reveal lifestyles in urban populations. *Nature communications*, 9(1):1–8, 2018.

[118] Anhai Doan, Raghu Ramakrishnan, and Alon Y. Halevy. Crowdsourcing systems on the world-wide web. *Commun. ACM*, 54(4):86–96, April 2011. ISSN 0001-0782.

[119] Diyana Dobreva, Daniel Grinnell, and Martin Innes. Prophets and loss: How soft facts on social media influenced the brexit campaign and social reactions to the murder of jo cox mp. *Policy & Internet*, 12(2):144–164, 2020. doi:

10.1002/poi3.203. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/poi3.203.

[120] Charmaine Du Plessis. The role of content marketing in social media content communities. *South African Journal of Information Management*, 19(1):1–7, 2017.

[121] Elizabeth Dubois and Grant Blank. The echo chamber is overstated: the moderating effect of political interest and diverse media. *Information, communication & society*, 21(5):729–745, 2018.

[122] Anamaria Dutceac Segesten, Michael Bossetta, Nils Holmberg, and Diederick Niehorster. The cueing power of comments on social media: how disagreement in facebook comments affects user engagement with news. *Information, Communication & Society*, pages 1–20, 2020.

[123] William H Dutton, Bianca Reisdorf, Elizabeth Dubois, and Grant Blank. Social shaping of the politics of internet search and networking: Moving beyond filter bubbles, echo chambers, and fake news. 2017.

[124] David Easley and Jon Kleinberg. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge university press, 2010.

[125] Reem El-Deeb, Fatma El-Zahraa El-Gamal, Nehal Sakr, Sara Elhishi, and Sara El-Metwally. Unlocking the public perception of covid-19 vaccination process on social media. In *2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS)*, pages 327–334. IEEE, 2021.

[126] Nicole B Ellison, Jessica Vitak, Rebecca Gray, and Cliff Lampe. Cultivating social resources on social network sites: Facebook relationship maintenance behaviors and their role in social capital processes. *Journal of Computer-Mediated Communication*, 19(4):855–870, 2014.

[127] Gunn Enli. Twitter as arena for the authentic outsider: exploring the social media campaigns of trump and clinton in the 2016 us presidential election. *European journal of communication*, 32(1):50–61, 2017.

[128] David Eppstein. Subgraph isomorphism in planar graphs and related problems. In *Graph Algorithms and Applications I*, pages 283–309. World Scientific, 2002.

[129] Motahhare Eslami, Amirhossein Aleyasen, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. Feedvis: A path for exploring news feed curation algorithms. In *Proceedings of the 18th acm conference companion on computer supported cooperative work & social computing*, pages 65–68, 2015.

[130] Ilham Esslimani, Armelle Brun, and Anne Boyer. From social networks to behavioral networks in recommender systems. In *2009 International Conference*

on Advances in Social Network Analysis and Mining, pages 143–148. IEEE, 2009.

[131] Ilham Esslimani, Armelle Brun, and Anne Boyer. Detecting leaders in behavioral networks. In *2010 International Conference on Advances in Social Networks Analysis and Mining*, pages 281–285. IEEE, 2010.

[132] Raymond A Eve, Raymond A Eve, Sara Horsfall, and Mary E Lee. *Chaos, complexity, and sociology: Myths, models, and theories*. Sage, 1997.

[133] Courtney Falk. Detecting twitter trolls using natural language processing techniques trained on message bodies, 2019.

[134] Lizhou Fan, Huizi Yu, and Zhanyuan Yin. Stigmatization in social media: Documenting and analyzing hate speech for covid-19 on twitter. *Proceedings of the Association for Information Science and Technology*, 57(1):e313, 2020.

[135] Davide Di Fatta, Francesco Caputo, Federica Evangelista, and Gandolfo Dominici. Small world theory and the world wide web: linking small world properties and website centrality. *International Journal of Markets and Business Systems*, 2 (2):126–140, 2016.

[136] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. The rise of social bots. *Communications of the ACM*, 59(7):96–104, June 2016. ISSN 00010782. doi: 10.1145/2818717. URL `http://dl.acm.org/citation.cfm?doid=2963119.2818717`.

[137] Alceu Ferraz Costa, Yuto Yamaguchi, Agma Juci Machado Traina, Caetano Traina, Jr., and Christos Faloutsos. RSC: Mining and Modeling Temporal Activity in Social Media. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 269–278, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3664-2. doi: 10.1145/2783258.2783294. URL `http://doi.acm.org/10.1145/2783258.2783294`. event-place: Sydney, NSW, Australia.

[138] Pnina Fichman and Samantha Sharp. Successful trolling on reddit: A comparison across subreddits in entertainment, health, politics, and religion. *Proceedings of the Association for Information Science and Technology*, 57(1):e333, 2020.

[139] Eileen Fischer and A. Rebecca Reuber. Social interaction via new social media: (How) can interactions on Twitter affect effectual thinking and behavior? *Journal of Business Venturing*, 26(1):1–18, January 2011. ISSN 08839026. doi: 10.1016/j.jbusvent.2010.09.002. URL `https://linkinghub.elsevier.com/retrieve/pii/S0883902610000856`.

[140] D. Fisher. Using egocentric networks to understand communication. *IEEE Internet Computing*, 9(5):20–28, 2005. doi: 10.1109/MIC.2005.114.

[141] Claudia Flores-Saviaga, Brian C Keegan, and Saiph Savage. Mobilizing the Trump Train:Understanding Collective Action in a Political Trolling Community. page 10.

[142] Peter L Francia. Free media and twitter in the 2016 presidential election: The unconventional campaign of donald trump. *Social Science Computer Review*, 36 (4):440–455, 2018.

[143] Sheera Frenkel. The rise and fall of the stop the stealfacebook group. *New York Times*, 5, 2020.

[144] Miguel Gallegos, Viviane de Castro Pecanha, and Tomás Caycho-Rodríguez. Anti-vax: the history of a scientific problem. *Journal of Public Health*, 2022.

[145] Venkata Rama Kiran Garimella and Ingmar Weber. A long-term analysis of polarization on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 528–531, 2017.

[146] Robert W. Gehl and Maria Bakardjieva. *Socialbots and Their Friends: Digital Media and the Automation of Sociality*. Taylor & Francis, December 2016. ISBN 978-1-317-26739-3. Google-Books-ID: ujklDwAAQBAJ.

[147] Natalie Gerhart and Mehrdad Koohikamali. Social network migration and anonymity expectations: What anonymous social network apps offer. *Computers in Human Behavior*, 95:101–113, 2019.

[148] Federico Germani and Nikola Biller-Andorno. The anti-vaccination infodemic on social media: A behavioral analysis. *PloS one*, 16(3):e0247642, 2021.

[149] James J Gibson. The theory of affordances. *Hilldale, USA*, 1(2):67–82, 1977.

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

[151] Maria Glenski and Tim Weninger. Predicting user-interactions on reddit. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 609–612, 2017.

[152] Amit Goldenberg, James Gross, and David Garcia. Emotional sharing on social media: How twitter replies contribute to increased emotional intensity.

[153] Vicenç Gómez, Andreas Kaltenbrunner, and Vicente López. Statistical analysis of the social network and discussion threads in slashdot. In *Proceedings of the 17th international conference on World Wide Web*, pages 645–654, 2008.

[154] Bruno Gonçalves, Nicola Perra, and Alessandro Vespignani. Modeling users' activity on twitter networks: Validation of dunbar's number. *PloS one*, 6(8): e22656, 2011.

[155] Yuriy Gorodnichenko, Tho Pham, and Oleksandr Talavera. Social media, sentiment and public opinions: Evidence from# brexit and# uselection. *European Economic Review*, 136:103772, 2021.

[156] Nicolò Gozzi, Michele Tizzani, Michele Starnini, Fabio Ciulla, Daniela Paolotti, André Panisson, and Nicola Perra. Collective response to the media coverage of COVID-19 Pandemic on Reddit and Wikipedia. *arXiv:2006.06446 [physics]*, June 2020. URL `http://arxiv.org/abs/2006.06446`. arXiv: 2006.06446.

[157] Mark S Granovetter. The strength of weak ties. *American journal of sociology*, 78(6):1360–1380, 1973.

[158] Siobhán Grayson and Derek Greene. Temporal analysis of reddit networks via role embeddings. *arXiv preprint arXiv:1908.05192*, 2019.

[159] Christian Grimme, Mike Preuss, Lena Adam, and Heike Trautmann. Social bots: Human-like by means of human control? *Big data*, 5(4):279–293, 2017.

[160] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. Fake news on twitter during the 2016 us presidential election. *Science*, 363(6425):374–378, 2019.

[161] Jacob Groshek and Karolina Koc-Michalska. Helping populism win? social media use, filter bubbles, and support for populist presidential candidates in the 2016 us election campaign. *Information, Communication & Society*, 20(9): 1389–1407, 2017.

[162] Nabil Guelzim, Samuele Bottani, Paul Bourgine, and François Képès. Topological and causal structure of the yeast transcriptional regulatory network. *Nature genetics*, 31(1):60, 2002.

[163] Andrew Guess, Jonathan Nagler, and Joshua Tucker. Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science advances*, 5(1):eaau4586, 2019.

[164] Ella Guest. (Anti-)Echo Chamber Participation: Examining Contributor Activity Beyond the Chamber. In *Proceedings of the 9th International Conference on Social Media and Society - SMSociety '18*, pages 301–304, Copenhagen, Denmark, 2018. ACM Press. ISBN 978-1-4503-6334-1. doi: 10.1145/3217804.3217933. URL `http://dl.acm.org/citation.cfm?doid=3217804.3217933`.

[165] Viorel Guliciuc. Complexity and social media. *Procedia-social and behavioral sciences*, 149:371–375, 2014.

[166] Sharath Chandra Guntuku, Pratik Narang, and Chittaranjan Hota. Real-time Peer-to-Peer Botnet Detection Framework based on Bayesian Regularized Neural Network. page 18.

[167] Andrew J Guydish, J Trevor DArcey, and Jean E Fox Tree. Reciprocity in conversation. *Language and Speech*, 64(4):859–872, 2021.

[168] Jeffrey A Hall. When is social media use social interaction? Defining mediated social interaction. *New Media & Society*, 20(1):162–179, January 2018. ISSN 1461-4448. doi: 10.1177/1461444816660782. URL https://doi.org/10.1177/1461444816660782. Publisher: SAGE Publications.

[169] Abdullah Hamid, Nasrullah Shiekh, Naina Said, Kashif Ahmad, Asma Gul, Laiq Hassan, and Ala Al-Fuqaha. Fake news detection in social media using graph neural networks and nlp techniques: A covid-19 use-case. *arXiv preprint arXiv:2012.07517*, 2020.

[170] William L Hamilton, Justine Zhang, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. Loyalty in online communities. In *Proceedings of the... International AAAI Conference on Weblogs and Social Media. International AAAI Conference on Weblogs and Social Media*, volume 2017, page 540. NIH Public Access, 2017.

[171] Jason Hannan. Trolling ourselves to death? social media and post-truth politics. *European Journal of Communication*, 33(2):214–226, 2018.

[172] Max Hänska and Stefan Bauchowitz. Tweeting for brexit: How social media shaped the referendum campaign. *Brexit, Trump and the Media*, pages 27–31, 2017.

[173] Folker Hanusch and Edson C Tandoc Jr. Comments, analytics, and social media: The impact of audience feedback on journalists market orientation. *Journalism*, 20(6):695–713, 2019.

[174] Summer Harlow. It was a facebook revolution": Exploring the meme-like spread of narratives during the egyptian protest. *Revista de comunicación*, (12):59–82, 2013.

[175] Martin Harrigan, Daniel Archambault, Pádraig Cunningham, and Neil Hurley. Egonav: Exploring networks through egocentric spatializations. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pages 563–570, 2012.

[176] Shlomo Havlin, Dror Y Kenett, Eshel Ben-Jacob, Armin Bunde, Reuven Cohen, H Hermann, JW Kantelhardt, J Kertész, S Kirkpatrick, Jürgen Kurths, et al. Challenges in network science: Applications to infrastructures, climate, social systems and economics. *The European Physical Journal Special Topics*, 214(1): 273–293, 2012.

[177] Bing He, Caleb Ziems, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, and Srijan Kumar. Racism is a virus: anti-asian hate and counterspeech in social media

during the covid-19 crisis. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 90–94, 2021.

[178] Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009. doi: 10.1109/TKDE.2008.239.

[179] S Hema et al. Changing face of india by rising face of internet usage. *SAARJ Journal on Banking & Insurance Research*, 8(2):28–35, 2019.

[180] Jeff Hemsley and Robert M Mason. Knowledge and knowledge management in the social media age. *Journal of Organizational Computing and Electronic Commerce*, 23(1-2):138–167, 2013.

[181] Aliaksandr Herasimenka, Jonathan Bright, Aleksi Knuutila, and Philip N. Howard. Misinformation and professional news on largely unmoderated platforms: the case of telegram. *Journal of Information Technology & Politics*, 0(0):1–15, 2022. doi: 10.1080/19331681.2022.2076272. URL `https://doi.org/10.1080/19331681.2022.2076272`.

[182] Maximilian Höller. The human component in social media and fake news: the performance of uk opinion leaders on twitter during the brexit campaign. *European Journal of English Studies*, 25(1):80–95, 2021.

[183] Tamanna Hossain, Robert L Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. Covidlies: Detecting covid-19 misinformation on social media. 2020.

[184] Yubo Hou, Dan Xiong, Tonglin Jiang, Lily Song, and Qi Wang. Social media addiction: Its impact, mediation, and intervention. *Cyberpsychology: Journal of psychosocial research on cyberspace*, 13(1), 2019.

[185] Hadrien Hours, Eric Fleury, and Márton Karsai. Link prediction in the twitter mention network: impacts of local structure and similarity of interest. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 454–461. IEEE, 2016.

[186] Philip N Howard, Aiden Duffy, Deen Freelon, Muzammil M Hussain, Will Mari, and Marwa Maziad. Opening closed regimes: what was the role of social media during the arab spring? *Available at SSRN 2595096*, 2011.

[187] Philip N Howard, Samuel Woolley, and Ryan Calo. Algorithms, bots, and political communication in the us 2016 election: The challenge of automated political communication for election law and administration. *Journal of information technology & politics*, 15(2):81–93, 2018.

[188] Lee Howell et al. Digital wildfires in a hyperconnected world. *WEF report*, 3 (2013):15–94, 2013.

[189] Toma Hoevar and Janez Demar. A combinatorial approach to graphlet counting. *Bioinformatics*, 30(4):559–565, 12 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btt717. URL https://doi.org/10.1093/bioinformatics/btt717.

[190] Man Hung, Evelyn Lauren, Eric S Hon, Wendy C Birmingham, Julie Xu, Sharon Su, Shirley D Hon, Jungweon Park, Peter Dang, Martin S Lipsky, et al. Social network analysis of covid-19 sentiments: Application of artificial intelligence. *Journal of medical Internet research*, 22(8):e22590, 2020.

[191] Sofia Hurtado, Poushali Ray, and Radu Marculescu. Bot Detection in Reddit Political Discussion. In *Proceedings of the Fourth International Workshop on Social Sensing - SocialSense'19*, pages 30–35, Montreal, QC, Canada, 2019. ACM Press. ISBN 978-1-4503-6706-6. doi: 10.1145/3313294.3313386. URL http://dl.acm.org/citation.cfm?doid=3313294.3313386.

[192] Takashi Iba, Keiichi Nemoto, Bernd Peters, and Peter A Gloor. Analyzing the creative editing behavior of wikipedia editors: Through dynamic social network analysis. *Procedia-Social and Behavioral Sciences*, 2(4):6441–6456, 2010.

[193] Martin Innes. Techniques of disinformation: Constructing and communicating soft facts after terrorism. *The British Journal of Sociology*, 71(2):284–299, 2020. ISSN 1468-4446. doi: 10.1111/1468-4446.12735. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-4446.12735. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-4446.12735.

[194] Martin Innes, Diyana Dobreva, and Helen Innes. Disinformation and digital influencing after terrorism: spoofing, truthing and social proofing. *Contemporary Social Science*, pages 1–15, 2019.

[195] José Luis Iribarren and Esteban Moro. Impact of human activity patterns on the dynamics of information diffusion. *Physical review letters*, 103(3):038702, 2009.

[196] Royi Itzhack, Yelena Mogilevski, and Yoram Louzoun. An optimal algorithm for counting network motifs. *Physica A: Statistical Mechanics and its Applications*, 381:482–490, July 2007. ISSN 03784371. doi: 10.1016/j.physa.2007.02.102. URL https://linkinghub.elsevier.com/retrieve/pii/S0378437107002257.

[197] Amelia Jamison, David A Broniatowski, Michael C Smith, Kajal S Parikh, Adeena Malik, Mark Dredze, and Sandra C Quinn. Adapting and extending a typology to identify vaccine misinformation on twitter. *American Journal of Public Health*, 110(S3):S331–S339, 2020.

[198] Jeannette Janssen, Matt Hurshman, and Nauzer Kalyaniwalla. Model selection for social networks using graphlets. *Internet Mathematics*, 8(4):338–363, 2012.

[199] Lorien Jasny, Joseph Waggle, and Dana R. Fisher. An empirical examination of echo chambers in US climate policy networks. *Nature Climate Change*, 5(8):782–786, August 2015. ISSN 1758-678X, 1758-6798. doi: 10.1038/nclimate2666. URL http://www.nature.com/articles/nclimate2666.

[200] Jiwan Jeong, Jeong-han Kang, and Sue Moon. Identifying and quantifying coordinated manipulation of upvotes and downvotes in naver news comments. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 303–314, 2020.

[201] Jiaojiao Jiang, Sheng Wen, Shui Yu, Yang Xiang, Wanlei Zhou, and Houcine Hassan. The structure of communities in scale-free networks. *Concurrency and Computation: Practice and Experience*, 29(14):e4040, 2017.

[202] Zhi-Qiang Jiang, Wen-Jie Xie, Ming-Xia Li, Boris Podobnik, Wei-Xing Zhou, and H Eugene Stanley. Calling patterns in human communication dynamics. *Proceedings of the National Academy of Sciences*, 110(5):1600–1605, 2013.

[203] Ehsan Jokar and Mohammad Mosleh. Community detection in social networks based on improved label propagation algorithm and balanced link density. *Physics Letters A*, 383(8):718–727, 2019.

[204] David Jurgens and Tsai-Ching Lu. Temporal motifs reveal the dynamics of editor interactions in wikipedia. In *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[205] Shaidah Jusoh. A study on nlp applications and ambiguity problems. *Journal of Theoretical & Applied Information Technology*, 96(6), 2018.

[206] Elaheh Kamaliha, Fatemeh Riahi, Vahed Qazvinian, and Jafar Adibi. Characterizing network motifs to identify spam comments. In *2008 IEEE international conference on data mining workshops*, pages 919–928. IEEE, 2008.

[207] Gerald C Kane and Sam Ransbotham. Research notecontent and collaboration: an affiliation network approach to information quality in online peer production communities. *Information Systems Research*, 27(2):424–439, 2016.

[208] Daekook Kang, Bomi Song, Byoungun Yoon, Youngjo Lee, and Yongtae Park. Diffusion pattern analysis for social networking sites using small-world network multiple influence model. *Technological Forecasting and Social Change*, 95: 73–86, 2015.

[209] Gloria J Kang, Sinclair R Ewing-Nelson, Lauren Mackey, James T Schlitt, Achla Marathe, Kaja M Abbas, and Samarth Swarup. Semantic network analysis of vaccine sentiment in online social media. *Vaccine*, 35(29):3621–3638, 2017.

[210] Kawaljeet Kaur Kapoor, Kuttimani Tamilmani, Nripendra P. Rana, Pushp Patil, Yogesh K. Dwivedi, and Sridhar Nerur. Advances in Social Media Research: Past, Present and Future. *Information Systems Frontiers*, 20(3):531–558, June 2018. ISSN 1572-9419. doi: 10.1007/s10796-017-9810-y. URL `https://doi.org/10.1007/s10796-017-9810-y`.

[211] Alireza Karduni, Isaac Cho, Ryan Wesslen, Sashank Santhanam, Svitlana Volkova, Dustin L Arendt, Samira Shaikh, and Wenwen Dou. Vulnerable to misinformation? verifi! In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 312–323, 2019.

[212] Zahra Razaghi Moghadam Kashani, Hayedeh Ahrabian, Elahe Elahi, Abbas Nowzari-Dalini, Elnaz Saberi Ansari, Sahar Asadi, Shahin Mohammadi, Falk Schreiber, and Ali Masoudi-Nejad. Kavosh: a new algorithm for finding network motifs. *BMC bioinformatics*, 10(1):1–12, 2009.

[213] Shoko Kato, Akihiro Koide, Takayasu Fushimi, Kazumi Saito, and Hiroshi Motoda. Network analysis of three twitter functions: favorite, follow and mention. In *Pacific Rim Knowledge Acquisition Workshop*, pages 298–312. Springer, 2012.

[214] Brian Keegan, Darren Gergle, and Noshir Contractor. Staying in the loop: structure and dynamics of wikipedia's breaking news collaborations. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, page 1. ACM, 2012.

[215] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*, pages 2915–2921, 2019.

[216] Young Mie Kim, Jordan Hsu, David Neiman, Colin Kou, Levi Bankston, Soo Yun Kim, Richard Heinrich, Robyn Baragwanath, and Garvesh Raskutti. The stealth media? groups and targets behind divisive issue campaigns on facebook. *Political Communication*, 35(4):515–541, 2018.

[217] Aniket Kittur, Bongwon Suh, Bryan A Pendleton, and Ed H Chi. He says, she says: conflict and coordination in wikipedia. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 453–462. ACM, 2007.

[218] Magnus A Knustad. Get lost, troll: How accusations of trolling in newspaper comment sections affect the debate. 2020.

[219] Aleksi Knuutila, Lisa-Maria Neudert, and Philip N Howard. Who is afraid of fake news? modeling risk perceptions of misinformation in 142 countries.

[220] Paraskevas Koukaras, Christos Tjortjis, and Dimitrios Rousidis. Social media types: introducing a data driven taxonomy. *Computing*, 102(1):295–340, 2020.

[221] Ramez Kouzy, Joseph Abi Jaoude, Afif Kraitem, Molly B El Alam, Basil Karam, Elio Adib, Jabra Zarka, Cindy Traboulsi, Elie Akl, and Khalil Baddour. Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter. *Cureus*, March 2020. ISSN 2168-8184. doi: 10.7759/cureus.7255.

[222] Lauri Kovanen, Márton Karsai, Kimmo Kaski, János Kertész, and Jari Saramäki. Temporal motifs in time-dependent networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(11):P11005, 2011.

[223] Lauri Kovanen, Kimmo Kaski, János Kertész, and Jari Saramäki. Temporal motifs reveal homophily, gender-specific patterns, and group talk in call sequences. *Proceedings of the National Academy of Sciences*, 110(45):18070–18075, 2013.

[224] Navin Kumar, Isabel Corpus, Meher Hans, Nikhil Harle, Nan Yang, Curtis McDonald, Shinpei Nakamura Sakai, Kamila Janmohamed, Keyu Chen, Frederick L Altice, et al. Covid-19 vaccine perceptions in the initial phases of us vaccine roll-out: an observational study on reddit. *BMC Public Health*, 22(1):1–14, 2022.

[225] Srijan Kumar, Robert West, and Jure Leskovec. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th international conference on World Wide Web*, pages 591–602. International World Wide Web Conferences Steering Committee, 2016.

[226] Srijan Kumar, William L Hamilton, Jure Leskovec, and Dan Jurafsky. Community interaction and conflict on the web. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 933–943. International World Wide Web Conferences Steering Committee, 2018.

[227] Srijan Kumar, Xikun Zhang, and Jure Leskovec. Predicting dynamic embedding trajectory in temporal interaction networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1269–1278. ACM, 2019.

[228] Srijan Kumar, Chongyang Bai, VS Subrahmanian, and Jure Leskovec. Deception detection in group video conversations using dynamic interaction networks. In *ICWSM 2021*. International AAAI Conference on Web and Social Media, 2021.

[229] Vincent Lacroix, Cristina G Fernandes, and Marie-France Sagot. Motif search in graphs: application to metabolic networks. *IEEE/ACM transactions on computational biology and bioinformatics*, 3(4):360–368, 2006.

[230] Brenden M Lake and Gregory L Murphy. Word meaning in minds and machines. *Psychological review*, 2021.

[231] Himabindu Lakkaraju, Julian McAuley, and Jure Leskovec. What's in a name? understanding the interplay between titles, content, and communities in social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7, pages 311–320, 2013.

[232] Rabindra Lamsal. Design and analysis of a large-scale covid-19 tweets dataset. *Applied Intelligence*, 51(5):2790–2804, 2021.

[233] David Laniado, Riccardo Tasso, Yana Volkovich, and Andreas Kaltenbrunner. When the wikipedians talk: Network and tree structure of wikipedia discussion pages. In *Sixth International AAAI Conference on Weblogs and Social Media*, pages 177–184, 2011.

[234] Austin E Lee. Coronavirus misinformation on reddit. 2022.

[235] Eunjin Lee, James Ashford, Malgorzata Turalska, Liam Turner, Vera Liao, Rachel Bellamy, Geeth de Mel, and Roger Whitaker. An exploratory analysis of suspicious reddit user accounts based on sentiment and interactions, 2019.

[236] Kristina Lerman and Rumi Ghosh. Information contagion: An empirical study of the spread of news on digg and twitter social networks. In *Fourth international AAAI conference on weblogs and social media*, 2010.

[237] Jure Leskovec and Julian Mcauley. Learning to discover social circles in ego networks. *Advances in neural information processing systems*, 25, 2012.

[238] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2–es, 2007.

[239] Qing Li, Jia Wang, Yuanzhu Peter Chen, and Zhangxi Lin. User comments for news recommendation in forum-based social media. *Information Sciences*, 180 (24):4929–4939, 2010.

[240] Yachao Li, Sylvia Twersky, Kelsey Ignace, Mei Zhao, Radhika Purandare, Breeda Bennett-Jones, and Scott R Weaver. Constructing and communicating covid-19 stigma on twitter: a content analysis of tweets during the early stage of the covid-19 outbreak. *International Journal of Environmental Research and Public Health*, 17(18):6847, 2020.

[241] Federica Liberini, Michela Redoano, Antonio Russo, Angel Cuevas, and Ruben Cuevas. Politics in the facebook era-evidence from the 2016 us presidential elections. 2020.

[242] Guanfeng Liu, Yan Wang, and Mehmet A Orgun. Trust transitivity in complex social networks. In *twenty-fifth AAAI conference on artificial intelligence*, 2011.

[243] Qiang Liu, Feng Yu, Shu Wu, and Liang Wang. Mining significant microblogs for misinformation identification: an attention-based approach. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 9(5):1–20, 2018.

[244] Xin Liu and Tsuyoshi Murata. Community detection in large-scale bipartite networks. *Transactions of the Japanese Society for Artificial Intelligence*, 25(1): 16–24, 2010.

[245] Giseli Rabello Lopes, Mirella M Moro, Leandro Krug Wives, and José Palazzo Moreira De Oliveira. Collaboration recommendation on academic social networks. In *International conference on conceptual modeling*, pages 190–199. Springer, 2010.

[246] Ann Majchrzak, Samer Faraj, Gerald C. Kane, and Bijan Azad. The Contradictory Influence of Social Media Affordances on Online Communal Knowledge Sharing. *Journal of Computer-Mediated Communication*, 19(1):38–55, October 2013. ISSN 10836101. doi: 10.1111/jcc4.12030. URL `https://academic.oup.com/jcmc/article/19/1/38-55/4067499`.

[247] Parul Malik and Seungyoon Lee. Follow me too: Determinants of transitive tie formation on twitter. *Social Media+ Society*, 6(3):2056305120939248, 2020.

[248] Osama Mansour. *Share with Social Media: the case of a Wiki*. PhD thesis, School of Computer Science, Physics and Mathematics, Linnaeus University, 2011.

[249] Osama Mansour, Mustafa Abu Salah, and Linda Askenäs. Wiki collaboration in organizations: an exploratory study. In *19th European Conference in Information Systems, Helsinki, Finland 9-11 June, 2011., 2011*. European Conference on Information Systems, 2011.

[250] Jacob Marx. Twitter and the 2016 presidential election. *Critique: A Worldwide*, pages 17–37, 2017.

[251] Sergei Maslov and Kim Sneppen. Specificity and stability in topology of protein networks. *Science*, 296(5569):910–913, 2002.

[252] Joan Massachs, Corrado Monti, Gianmarco De Francisci Morales, and Francesco Bonchi. Roots of trumpism: Homophily and social feedback in donald trump support on reddit. In *12th ACM Conference on Web Science*, pages 49–58, 2020.

[253] Brendan D. McKay and Adolfo Piperno. Practical graph isomorphism, ii. *Journal of Symbolic Computation*, 60:94112, Jan 2014. ISSN 0747-7171. doi: 10.1016/j.jsc.2013.09.003. URL `http://dx.doi.org/10.1016/j.jsc.2013.09.003`.

[254] Cassie McMillan, Diane Felmlee, and James R Ashford. Reciprocity, transitivity, and skew: Comparing local structure in 40 positive and negative social networks. *Plos one*, 17(5):e0267886, 2022.

[255] Juan Carlos Medina Serrano, Orestis Papakyriakopoulos, and Simon Hegelich. NLP-based feature extraction for the detection of COVID-19 misinformation videos on YouTube. In *Proceedings of the 1st Workshop on NLP for COVID-19 at*

ACL 2020, Online, July 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.nlpcovid19-acl.17.

[256] Alexey N Medvedev, Jean-Charles Delvenne, and Renaud Lambiotte. Modelling structure and predicting dynamics of discussion threads in online boards. *Journal of Complex Networks*, 7(1):67–82, 2019.

[257] David Melamed. Community structures in bipartite networks: A dual-projection approach. *PloS one*, 9(5):e97823, 2014.

[258] Chad Melton, Olufunto A Olusanya, and Arash Shaban-Nejad. Network analysis of covid-19 vaccine misinformation on social media. *Stud Health Technol Inform*, 287:165–166, 2021.

[259] Paul Mena, Danielle Barbe, and Sylvia Chan-Olmsted. Misinformation on instagram: The impact of trusted endorsements on message credibility. *Social Media+ Society*, 6(2):2056305120935102, 2020.

[260] Emily Merritt. An Analysis of the Discourse of Internet Trolling: A Case Study of Reddit.com. page 135.

[261] Emily Merritt. *An analysis of the discourse of Internet trolling: A case study of Reddit. com*. PhD thesis, 2012.

[262] Elena Milani, Emma Weitkamp, and Peter Webb. The visual vaccine debate on twitter: A social network analysis. *Media and Communication*, 8(2):364–375, 2020.

[263] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[264] Ron Milo, Shalev Itzkovitz, Nadav Kashtan, Reuven Levitt, Shai Shen-Orr, Inbal Ayzenshtat, Michal Sheffer, and Uri Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–1542, 2004.

[265] Kunihiro Miyazaki, Takayuki Uchiba, Kenji Tanaka, and Kazutoshi Sasahara. Characterizing the anti-vaxxers reply behavior on social media. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 83–89, 2021.

[266] Kunihiro Miyazaki, Takayuki Uchiba, Kenji Tanaka, and Kazutoshi Sasahara. The strategy behind anti-vaxxers' reply behavior on social media. *arXiv preprint arXiv:2105.10319*, 2021.

[267] Ewa Młynarska, Derek Greene, and Padraig Cunningham. Time series clustering of moodle activity data. In *24th Irish Conference on Artificial Intelligence and Cognitive Science (AICS'16), University College Dublin, Dublin, Ireland, 20-21 September 2016*, 2016.

[268] Delia Mocanu, Luca Rossi, Qian Zhang, Marton Karsai, and Walter Quattrociocchi. Collective attention in the age of (mis)information. *Computers in Human Behavior*, 51:1198–1204, 2015. ISSN 0747-5632.

[269] Sina Mohseni and Eric Ragan. Combating fake news with interpretable news feed algorithms. *arXiv preprint arXiv:1811.12349*, 2018.

[270] Logan Molyneux and Rachel R Mourão. Political journalists normalization of twitter: Interaction and new affordances. *Journalism Studies*, 20(2):248–266, 2019.

[271] Peter R Monge and Noshir S Contractor. Emergence of communication networks. *The new handbook of organizational communication: Advances in theory, research, and methods*, pages 440–502, 2001.

[272] Michal Monselise, Chia-Hsuan Chang, Gustavo Ferreira, Rita Yang, Christopher C Yang, et al. Topics and sentiments of public concerns regarding covid-19 vaccines: social media trend analysis. *Journal of Medical Internet Research*, 23 (10):e30765, 2021.

[273] Ahmet Anil Müngen and Mehmet Kaya. Influence analysis of posts in social networks by using quad-motifs. In *2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*, pages 1–5. IEEE, 2017.

[274] Goran Muric, Yusong Wu, Emilio Ferrara, et al. Covid-19 vaccine hesitancy on social media: building a public twitter data set of antivaccine content, vaccine misinformation, and conspiracies. *JMIR public health and surveillance*, 7(11): e30642, 2021.

[275] Seth A Myers, Aneesh Sharma, Pankaj Gupta, and Jimmy Lin. Information network or social network? the structure of the twitter follow graph. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 493–498, 2014.

[276] Maziar Nekovee, Yamir Moreno, Ginestra Bianconi, and Matteo Marsili. Theory of rumour spreading in complex social networks. *Physica A: Statistical Mechanics and its Applications*, 374(1):457–470, 2007.

[277] Edward Newell, David Jurgens, Haji Mohammad Saleem, Hardik Vala, Jad Sassine, Caitrin Armstrong, and Derek Ruths. User migration in online social networks: A case study on reddit during a period of community unrest. In *ICWSM*, pages 279–288, 2016.

[278] M. E. J. Newman. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Phys. Rev. E*, 64:016132, Jun 2001. doi: 10.1103/PhysRevE.64.016132. URL https://link.aps.org/doi/10.1103/PhysRevE.64.016132.

[279] M. E. J. Newman. Scientific collaboration networks. i. network construction and fundamental results. *Phys. Rev. E*, 64:016131, Jun 2001. doi: 10.1103/PhysRevE.64.016131. URL https://link.aps.org/doi/10.1103/PhysRevE.64.016131.

[280] Mark Newman. Networks, 2nd edn oxford. *UK: Oxford University Press.[Google Scholar]*, 2018.

[281] Mark EJ Newman. The structure of scientific collaboration networks. *Proceedings of the national academy of sciences*, 98(2):404–409, 2001.

[282] Mark EJ Newman. Detecting community structure in networks. *The European physical journal B*, 38(2):321–330, 2004.

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

[284] Mark EJ Newman, Steven H Strogatz, and Duncan J Watts. Random graphs with arbitrary degree distributions and their applications. *Physical review E*, 64(2): 026118, 2001.

[285] Rishab Nithyanand, Brian Schaffner, and Phillipa Gill. Online Political Discourse in the Trump Era. *arXiv:1711.05303 [cs]*, November 2017. URL http://arxiv.org/abs/1711.05303. arXiv: 1711.05303.

[286] Arlind Nocaj, Mark Ortmann, and Ulrik Brandes. Untangling the hairballs of multi-centered, small-world online social media networks. *Journal of Graph Algorithms and Applications: JGAA*, 19(2):595–618, 2015.

[287] Pippa Norris. Political mobilization and social networks. the example of the arab spring. *Electronic democracy*, pages 53–76, 2012.

[288] Martin A Nowak and Karl Sigmund. Evolution of indirect reciprocity. *Nature*, 437(7063):1291–1298, 2005.

[289] Jonathan A Obar and Steven S Wildman. Social media definition and the governance challenge-an introduction to the special issue. *Obar, JA and Wildman, S.(2015). Social media definition and the governance challenge: An introduction to the special issue. Telecommunications policy*, 39(9):745–750, 2015.

[290] Derek O'Callaghan, Martin Harrigan, Joe Carthy, and Pádraig Cunningham. Network analysis of recurring youtube spam campaigns. In *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[291] Derek O'Callaghan, Martin Harrigan, Joe Carthy, and Pádraig Cunningham. Identifying discriminating network motifs in youtube spam. *arXiv preprint arXiv:1202.5216*, 2012.

[292] Daire O'Doherty, Salim Jouili, and Peter Van Roy. Towards trust inference from bipartite social networks. In *Proceedings of the 2nd ACM SIGMOD Workshop on Databases and Social Networks*, pages 13–18, 2012.

[293] El E Omran and Jacob van Etten. Spatial-data sharing: Applying social-network analysis to study individual and collective behaviour. *International journal of geographical information science*, 21(6):699–714, 2007.

[294] Daniele Orso, Nicola Federici, Roberto Copetti, Luigi Vetrugno, and Tiziana Bove. Infodemic and the spread of fake news in the covid-19-era. *European Journal of Emergency Medicine*, 2020.

[295] Ray Oshikawa, Jing Qian, and William Yang Wang. A survey on natural language processing for fake news detection. *arXiv preprint arXiv:1811.00770*, 2018.

[296] Mariam El Ouirdi, Asma El Ouirdi, Jesse Segers, and Erik Henderickx. Social media conceptualization and taxonomy: A lasswellian framework. *Journal of Creative Communications*, 9(2):107–126, 2014.

[297] Tim Owen, Wayne Noble, and Faye Christabel Speed. Trolling, the ugly face of the social network. In *New Perspectives on Cybercrime*, pages 113–139. Springer, 2017.

[298] Cathal OConnor and Michelle Murphy. Going viral: doctors must tackle fake news in the covid-19 pandemic. *bmj*, 24(369):m1587, 2020.

[299] Jordi Palau, Miquel Montaner, Beatriz López, and Josep Lluís De La Rosa. Collaboration analysis in recommender systems using social networks. In *International Workshop on Cooperative Information Agents*, pages 137–151. Springer, 2004.

[300] Pietro Panzarasa, Tore Opsahl, and Kathleen M Carley. Patterns and dynamics of users' behavior and interaction: Network analysis of an online community. *Journal of the American Society for Information Science and Technology*, 60(5): 911–932, 2009.

[301] Ashwin Paranjape, Austin R Benson, and Jure Leskovec. Motifs in temporal networks. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 601–610. ACM, 2017.

[302] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200, 2001.

[303] Xin Pei and Deval Mehta. # coronavirus or# chinesevirus?!: Understanding the negative sentiment reflected in tweets with racist hashtags across the development of covid-19. *arXiv preprint arXiv:2005.08224*, 2020.

[304] Xixian Peng, Yuxiang Chris Zhao, and Qinghua Zhu. Investigating user switching intention for mobile instant messaging application: Taking wechat as an example. *Computers in Human Behavior*, 64:206–216, 2016.

[305] Gordon Pennycook and David G Rand. The psychology of fake news. *Trends in cognitive sciences*, 25(5):388–402, 2021.

[306] Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G Lu, and David G Rand. Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention. page 11.

[307] Gordon Pennycook, Jabin Binnendyk, Christie Newton, and David G Rand. A practical guide to doing behavioral research on fake news and misinformation. *Collabra: Psychology*, 7(1):25293, 2021.

[308] Andrew Perrin. Social media usage. *Pew research center*, 125:52–68, 2015.

[309] Chee Wei Phang, Chenghong Zhang, and Juliana Sutanto. The influence of user interaction and participation in social media on the consumption intention of niche products. *Information & Management*, 50(8):661–672, December 2013. ISSN 03787206. doi: 10.1016/j.im.2013.07.001. URL `https://linkinghub.elsevier.com/retrieve/pii/S0378720613000736`.

[310] Christopher Phethean, Thanassis Tiropanis, and Lisa Harris. Engaging with Charities on Social Media: Comparing Interaction on Facebook and Twitter. In Thanassis Tiropanis, Athena Vakali, Laura Sartori, and Pete Burnap, editors, *Internet Science*, Lecture Notes in Computer Science, pages 15–29, Cham, 2015. Springer International Publishing. ISBN 978-3-319-18609-2. doi: 10.1007/978-3-319-18609-2_2.

[311] Martin Potthast, Benno Stein, and Robert Gerling. Automatic vandalism detection in wikipedia. In *European conference on information retrieval*, pages 663–668. Springer, 2008.

[312] Brandon Price. Lie machines: how to save democracy from troll armies, deceitful robots, junk news operations, and political operatives: authored by philip n. howard, yale university press, new haven, ct, $26.00usd, hardback, (isbn13 : 978 - 0300250206), 2020.240pp., 2021.$

[313] Peng Qi, Juan Cao, Tianyun Yang, Junbo Guo, and Jintao Li. Exploiting multi-domain visual information for fake news detection. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 518–527. IEEE, 2019.

[314] Xiangju Qin and Pádraig Cunningham. Assessing the quality of wikipedia pages using edit longevity and contributor centrality. *arXiv preprint arXiv:1206.2517*, 2012.

[315] Xiangju Qin, Pádraig Cunningham, and Michael Salter-Townshend. The influence of network structures of wikipedia discussion pages on the efficiency of wikiprojects. *Social Networks*, 43:1–15, 2015.

[316] Hoda Sepehri Rad and Denilson Barbosa. Identifying controversial articles in wikipedia: A comparative study. In *Proceedings of the eighth annual international symposium on wikis and open collaboration*, page 7. ACM, 2012.

[317] Raquel Recuero, Ricardo Araujo, and Gabriela Zago. How does social capital affect retweets? In *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[318] Ulrike Reisach. The responsibility of social media in times of societal and political manipulation. *European Journal of Operational Research*, 291(3):906–917, 2021.

[319] Filipe N Ribeiro, Koustuv Saha, Mahmoudreza Babaei, Lucas Henrique, Johnnatan Messias, Fabricio Benevenuto, Oana Goga, Krishna P Gummadi, and Elissa M Redmiles. On microtargeting socially divisive ads: A case study of russia-linked ad campaigns on facebook. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 140–149, 2019.

[320] Pedro Ribeiro and Fernando Silva. G-tries: a data structure for storing and finding subgraphs. *Data Mining and Knowledge Discovery*, 28(2):337–377, 2014.

[321] Garry Robins and Malcolm Alexander. Small worlds among interlocking directors: Network structure and distance in bipartite graphs. *Computational & Mathematical Organization Theory*, 10(1):69–94, 2004.

[322] Ylva Rodny-Gumede. *Fake It till You Make It: The Role, Impact and Consequences of Fake News*, pages 203–219. Springer International Publishing, Cham, 2018. ISBN 978-3-319-62057-2. doi: 10.1007/978-3-319-62057-2_13. URL https://doi.org/10.1007/978-3-319-62057-2_13.

[323] Daniel M Romero, Brian Uzzi, and Jon Kleinberg. Social networks under stress: Specialized team roles and their communication structure. *ACM Transactions on the Web (TWEB)*, 13(1):1–24, 2019.

[324] Miguel Romero, Camilo Rocha, and Jorge Finke. Spectral evolution of twitter mention networks. In *International Conference on Complex Networks and Their Applications*, pages 532–542. Springer, 2019.

[325] Luca Rossi and Matteo Magnani. Conversation practices and network structure in twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 6, pages 563–566, 2012.

[326] Rahmtin Rotabi, Krishna Kamath, Jon Kleinberg, and Aneesh Sharma. Detecting strong ties using network motifs. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 983–992, 2017.

[327] Jeffrey M Rzeszotarski, Emma S Spiro, Jorge Nathan Matias, Andrés Monroy-Hernández, and Meredith Ringel Morris. Is anyone out there? unpacking q&a hashtags on twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2755–2758, 2014.

[328] Fadi Safieddine and Rawad Hammad. Fake news: Origins and political impact. In *Handbook of Research on Recent Developments in Internet Activism and Political Participation*, pages 103–121. IGI global, 2020.

[329] Monique A Sager, Aditya M Kashyap, Mila Tamminga, Sadhana Ravoori, Christopher Callison-Burch, and Jules B Lipoff. Identifying and responding to health misinformation on reddit dermatology forums with artificially intelligent bots using natural language processing: design and evaluation study. *JMIR Dermatology*, 4(2):e20975, 2021.

[330] Mattia Samory and Tanushree Mitra. Conspiracies online: User discussions in a conspiracy community following dramatic events. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[331] Dietram A. Scheufele and Nicole M. Krause. Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences*, 116(16):7662–7669, 2019. ISSN 0027-8424. doi: 10.1073/pnas.1805871115. URL `https://www.pnas.org/content/116/16/7662`. Publisher: National Academy of Sciences _eprint: https://www.pnas.org/content/116/16/7662.full.pdf.

[332] Falk Schreiber and Henning Schwöbbermeyer. Mavisto: a tool for the exploration of network motifs. *Bioinformatics*, 21(17):3572–3574, 2005.

[333] Dina Sebastião and Susana Borges. Should we stay or should we go: Eu input legitimacy under threat? social media and brexit. *Transforming Government: People, Process and Policy*, 2021.

[334] Hoda Sepehri-Rad and Denilson Barbosa. Identifying controversial wikipedia articles using editor collaboration networks. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6(1):5, 2015.

[335] Gautam Kishore Shahi, Anne Dirkson, and Tim A Majchrzak. An exploratory study of covid-19 misinformation on twitter. *Online Social Networks and Media*, 22:100104, 2021.

[336] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th international conference companion on world wide web*, pages 745–750, 2016.

[337] Karishma Sharma, Yizhou Zhang, and Yan Liu. Covid-19 vaccine misinformation campaigns and social media narratives. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 920–931, 2022.

[338] Zeqian Shen and Neel Sundaresan. ebay: an e-commerce marketplace as a complex network. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 655–664, 2011.

[339] Shai S Shen-Orr, Ron Milo, Shmoolik Mangan, and Uri Alon. Network motifs in the transcriptional regulation network of escherichia coli. *Nature genetics*, 31(1):64, 2002.

[340] Jieun Shin, Lian Jian, Kevin Driscoll, and François Bar. The diffusion of misinformation on social media: Temporal pattern, message, and source. *Computers in Human Behavior*, 83:278–287, 2018.

[341] Michelle Shumate, Andrew Pilny, Yannick Catouba, Jinseok Kim, Macarena Peña-y Lillo, Katherine Rcooper, Ariann Sahagun, and Sijia Yang. A taxonomy of communication networks. *Annals of the International Communication Association*, 37(1):95–123, 2013.

[342] Philipp Singer, Fabian Flöck, Clemens Meinhart, Elias Zeitfogel, and Markus Strohmaier. Evolution of reddit: from the front page of the internet to a self-referential community? In *Proceedings of the 23rd international conference on world wide web*, pages 517–522, 2014.

[343] Lisa Singh, Shweta Bansal, Leticia Bode, Ceren Budak, Guangqing Chi, Kornraphop Kawintiranon, Colton Padden, Rebecca Vanarsdall, Emily Vraga, and Yanchen Wang. A first look at COVID-19 information and misinformation sharing on Twitter. *arXiv:2003.13907 [cs]*, March 2020. URL `http://arxiv.org/abs/2003.13907`. arXiv: 2003.13907.

[344] Neil J Smelser. Problematics of sociology. In *Problematics of Sociology*. University of California Press, 2020.

[345] Adam Smidi and Saif Shahin. Social media and social mobilisation in the middle east: A survey of research on the arab spring. *India Quarterly*, 73(2):196–209, 2017.

[346] Ahmed Soliman, Jan Hafer, and Florian Lemmerich. A characterization of political communities on reddit. In *Proceedings of the 30th ACM conference on hypertext and Social Media*, pages 259–263, 2019.

[347] Jihye Song, Minjoon Kim, Haksung Kim, and Kwanghoon Kim. A framework: Workflow-based social network discovery and analysis. In *2010 13th IEEE International Conference on Computational Science and Engineering*, pages 421–426. IEEE, 2010.

[348] Xingyi Song, Johann Petrak, Ye Jiang, Iknoor Singh, Diana Maynard, and Kalina Bontcheva. Classification Aware Neural Topic Model and its Application on a New COVID-19 Disinformation Corpus. *arXiv:2006.03354 [cs, stat]*, June 2020. URL `http://arxiv.org/abs/2006.03354`. arXiv: 2006.03354.

[349] Daniel Sousa, Luís Sarmento, and Eduarda Mendes Rodrigues. Characterization of the twitter @replies network: are user ties social or topical? In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 63–70, 2010.

[350] Dominic Spohr. Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business Information Review*, 34(3):150–160, 2017.

[351] Tiziano Squartini, Francesco Picciolo, Franco Ruzzenenti, and Diego Garlaschelli. Reciprocity of weighted networks. *Scientific reports*, 3:2729, 2013.

[352] Karsten Steinhaeuser and Nitesh V Chawla. Community detection in a large real-world social network. In *Social computing, behavioral modeling, and prediction*, pages 168–175. Springer, 2008.

[353] Ekaterina Stepanova. The role of information communication technologies in the arab spring. *Ponars Eurasia*, 15(1):1–6, 2011.

[354] Leo G Stewart, Ahmer Arif, and Kate Starbird. Examining trolls and polarization with a retweet network. In *Proc. ACM WSDM, workshop on misinformation and misbehavior mining on the web*, 2018.

[355] Qi Su, Mingyu Wan, Xiaoqian Liu, Chu-Ren Huang, et al. Motivations, methods and metrics of misinformation detection: an nlp perspective. *Natural Language Processing Research*, 1(1-2):1–13, 2020.

[356] VS Subrahmanian, Amos Azaria, Skylar Durst, Vadim Kagan, Aram Galstyan, Kristina Lerman, Linhong Zhu, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. The darpa twitter bot challenge. *Computer*, 49(6):38–46, 2016.

[357] Kumar Subramani, Alexander Velkov, Irene Ntoutsi, Peer Kroger, and Hans-Peter Kriegel. Density-based community detection in social networks. In *2011 IEEE 5th International Conference on Internet Multimedia Systems Architecture and Application*, pages 1–8. IEEE, 2011.

[358] YOON Sunmoo, Michelle Odlum, Peter Broadwell, Nicole Davis, CHO Hwayoug, DENG Nanyi, Maria Patrao, Deborah Schauer, Michael E Bales, and Carmela Alcantara. Application of social network analysis of covid-19 related tweets mentioning cannabis and opioids to gain insights for drug abuse research. *Studies in health technology and informatics*, 272:5, 2020.

[359] James Surowiecki. *The Wisdom of Crowds*. Anchor, 2005. ISBN 0385721706.

[360] James Surowiecki and James. *The wisdom of crowds*. Anchor Books, 2005. ISBN 0385721706. URL `https://dl.acm.org/citation.cfm?id=1095645`.

[361] Joëlle Swart. Experiencing algorithms: How young people understand, feel about, and engage with algorithmic news selection on social media. *Social media+ society*, 7(2): 20563051211008828, 2021.

[362] Samia Tasnim, Md Mahbub Hossain, and Hoimonty Mazumder. Impact of rumors and misinformation on covid-19 in social media. *Journal of preventive medicine and public health*, 53(3):171–174, 2020.

[363] Huu Dat Tran. *Make A-meme-rica great again!: Studying the memers among Trump supporters in the 2020 US presidential election on Twitter via hashtags# maga and# trump2020*. PhD thesis, 2021.

[364] Jeffrey W. Treem and Paul M. Leonardi. Social Media Use in Organizations: Exploring the Affordances of Visibility, Editability, Persistence, and Association. *Annals of the International Communication Association*, 36(1):143–189, January 2013. ISSN 2380-8985. doi: 10.1080/23808985.2013.11679130. URL `https://doi.org/10.1080/23808985.2013.11679130`. Publisher: Routledge _eprint: https://doi.org/10.1080/23808985.2013.11679130.

[365] Bao Tran Truong, Oliver Melbourne Allen, and Filippo Menczer. News sharing networks expose information polluters on social media. *arXiv preprint arXiv:2202.00094*, 2022.

[366] Maksim Tsvetovat and Alexander Kouznetsov. *Social Network Analysis for Startups: Finding connections on the social web*. " O'Reilly Media, Inc.", 2011.

[367] K. Tu, J. Li, D. Towsley, D. Braines, and L. D. Turner. gl2vec: Learning feature representation using graphlets for directed networks. In *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 216–221, 2019. doi: 10.1145/3341161.3342908.

[368] Kun Tu, Jian Li, Don Towsley, Dave Braines, and Liam Turner. Classifying Types of Network Communities Using Motifs. page 1.

[369] Kun Tu, Jian Li, Don Towsley, Dave Braines, and Liam D Turner. Network classification in temporal networks using motifs. *arXiv preprint arXiv:1807.03733*, 2018.

[370] Liam D Turner, Roger M Whitaker, Stuart M Allen, David EJ Linden, Kun Tu, Jian Li, and Don Towsley. Evidence to support common application switching behaviour on smartphones. *Royal Society Open Science*, 6(3):190018, 2019.

[371] Petter Törnberg. Echo chambers and viral misinformation: Modeling fake news as complex contagion. *PLoS one*, 13(9):e0203958, 2018. Publisher: Public Library of Science San Francisco, CA USA.

[372] Jay J Van Bavel, Katherine Baicker, Paulo S Boggio, Valerio Capraro, Aleksandra Cichocka, Mina Cikara, Molly J Crockett, Alia J Crum, Karen M Douglas, James N Druckman, et al. Using social and behavioural science to support covid-19 pandemic response. *Nature Human Behaviour*, pages 1–12, 2020.

[373] Wil MP Van Der Aalst, Hajo A Reijers, and Minseok Song. Discovering social networks from event logs. *Computer Supported Cooperative Work (CSCW)*, 14(6):549–593, 2005.

[374] Sander van der Linden. Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature Medicine*, 28(3):460–467, 2022.

[375] Richard Van Noorden. Online collaboration: Scientists and the social network. *Nature news*, 512(7513):126, 2014.

[376] Nathalie Van Raemdonck. The echo chamber of anti-vaccination conspiracies: mechanisms of radicalization on facebook and reddit. *Institute for Policy, Advocacy and Governance (IPAG) Knowledge Series, Forthcoming*, 2019.

[377] M Del Vicario, A Bessi, F Zollo, F Petroni, A Scala, and G Caldarelli. The spreading of misinformation online. *Proc National Academy of Sciences*, 113(3), 2016.

[378] Lauren Vogel. Viral misinformation threatens public health, 2017.

[379] Wouter Vollenbroek, Sjoerd De Vries, Efthymios Constantinides, and Piet Kommers. Identification of influence in social media communities. *International Journal of Web Based Communities 4*, 10(3):280–297, 2014.

[380] Vilma Vuori and Jari Jussila. The 5c categorization of social media tools. In *Proceedings of the 20th international academic mindtrek conference*, pages 26–33, 2016.

[381] Byron C Wallace. Computational irony: A survey and new perspectives. *Artificial intelligence review*, 43(4):467–483, 2015.

[382] Chuang Wang, Matthew KO Lee, and Zhongsheng Hua. A theory of social media dependence: Evidence from microblog users. *Decision support systems*, 69:40–49, 2015.

[383] Yu Wang, Yuncheng Li, and Jiebo Luo. Deciphering the 2016 us presidential campaign in the twitter sphere: A comparison of the trumpists and clintonists. In *Tenth International AAAI Conference on Web and Social Media*, 2016.

[384] Yuxi Wang, Martin McKee, Aleksandra Torbica, and David Stuckler. Systematic literature review on the spread of health-related misinformation on social media. *Social science & medicine*, 240:112552, 2019.

[385] Ken Ward. Social networks, the 2016 us presidential election, and kantian ethics: applying the categorical imperative to cambridge analyticas behavioral microtargeting. *Journal of media ethics*, 33(3):133–148, 2018.

[386] Echo L Warner, Juliana L Barbati, Kaylin L Duncan, Kun Yan, and Stephen A Rains. Vaccine misinformation types and properties in russian troll tweets. *Vaccine*, 40(6): 953–960, 2022.

[387] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

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

[389] Helena Webb, Marina Jirotka, Bernd Carsten Stahl, William Housley, Adam Edwards, Matthew Williams, Rob Procter, Omer Rana, and Pete Burnap. Digital wildfires: hyper-connectivity, havoc and a global ethos to govern social media. *ACM SIGCAS Computers and Society*, 45(3):193–201, 2016.

[390] Jen Weedon, William Nuland, and Alex Stamos. Information operations and facebook. *Retrieved from Facebook: https://fbnewsroomus. files. wordpress. com/2017/04/Facebook-and-information-operations-v1. pdf*, 2017.

[391] Tim Weninger. An exploration of submissions and discussions in social news: Mining collective intelligence of reddit. *Social Network Analysis and Mining*, 4(1):173, 2014.

[392] Tim Weninger, Xihao Avi Zhu, and Jiawei Han. An exploration of discussion threads in social news sites: A case study of the reddit community. In *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, pages 579–583. IEEE, 2013.

[393] Sebastian Wernicke and Florian Rasche. Fanmod: a tool for fast network motif detection. *Bioinformatics*, 22(9):1152–1153, 2006.

[394] Anita Whiting and David Williams. Why people use social media: a uses and gratifications approach. *Qualitative Market Research: An International Journal*, 16(4): 362–369, January 2013. ISSN 1352-2752. doi: 10.1108/QMR-06-2013-0041. URL `https://doi.org/10.1108/QMR-06-2013-0041`. Publisher: Emerald Group Publishing Limited.

[395] Gadi Wolfsfeld, Elad Segev, and Tamir Sheafer. Social media and the arab spring: Politics comes first. *The International Journal of Press/Politics*, 18(2):115–137, 2013.

[396] Samuel C. Woolley and Philip N. Howard. *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media*. Oxford University Press, November 2018. ISBN 978-0-19-093140-7.

[397] Guangyu Wu and Pádraig Cunningham. Integration of multiple network views in wikipedia. *Knowledge and Information Systems*, 45(2):473–490, 2015.

[398] Guangyu Wu, Martin Harrigan, and Pádraig Cunningham. Characterizing wikipedia pages using edit network motif profiles. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 45–52, 2011.

[399] Guangyu Wu, Martin Harrigan, and Pádraig Cunningham. Characterizing Wikipedia pages using edit network motif profiles. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents - SMUC '11*, page 45, Glasgow, Scotland, UK, 2011. ACM Press. ISBN 978-1-4503-0949-3. doi: 10.1145/2065023.2065036. URL http://dl.acm.org/citation.cfm?doid=2065023.2065036.

[400] Guangyu Wu, Martin Harrigan, and Pádraig Cunningham. Classifying wikipedia articles using network motif counts and ratios. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, pages 1–10, 2012.

[401] Guangyu Wu, Martin Harrigan, and Pádraig Cunningham. Classifying wikipedia articles using network motif counts and ratios. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, WikiSym '12, pages 12:1–12:10, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1605-7. doi: 10.1145/2462932.2462948. URL http://doi.acm.org/10.1145/2462932.2462948.

[402] Zi Yang, Jingyi Guo, Keke Cai, Jie Tang, Juanzi Li, Li Zhang, and Zhong Su. Understanding retweeting behaviors in social networks. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1633–1636, 2010.

[403] Lindsay E Young, Emily Sidnam-Mauch, Marlon Twyman, Liyuan Wang, Jackie Jingyi Xu, Matthew Sargent, Thomas W Valente, Emilio Ferrara, Janet Fulk, and Peter Monge. Disrupting the covid-19 misinfodemic with network interventions: Network solutions for network problems. *American journal of public health*, 111(3):514–519, 2021.

[404] Sixie Yu, Yevgeniy Vorobeychik, and Scott Alfeld. Adversarial classification on social networks. *arXiv preprint arXiv:1801.08159*, 2018.

[405] M N Yudina, V N Zadorozhnyi, and E B Yudin. Mixed Random Sampling of Frames method for counting number of motifs. page 8.

[406] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 218–226, San Francisco USA, May 2019. ACM. ISBN 978-1-4503-6675-5. doi: 10.1145/3308560.3316495. URL https://dl.acm.org/doi/10.1145/3308560.3316495.

[407] Savvas Zannettou, Tristan Caulfield, William Setzer, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. Who let the trolls out? towards understanding state-sponsored trolls. In *Proceedings of the 10th acm conference on web science*, pages 353–362, 2019.

[408] Fuguo Zhang, Shumei Qi, Qihua Liu, Mingsong Mao, and An Zeng. Alleviating the data sparsity problem of recommender systems by clustering nodes in bipartite networks. *Expert Systems with Applications*, 149:113346, 2020.

[409] Jason Shuo Zhang, Brian C. Keegan, Qin Lv, and Chenhao Tan. A Tale of Two Communities: Characterizing Reddit Response to COVID-19 through /r/China_flu and /r/Coronavirus. *arXiv:2006.04816 [cs]*, June 2020. URL `http://arxiv.org/abs/2006.04816`. arXiv: 2006.04816.

[410] Junlong Zhang and Yu Luo. Degree centrality, betweenness centrality, and closeness centrality in social network. In *2017 2nd international conference on modelling, simulation and applied mathematics (MSAM2017)*, pages 300–303. Atlantis Press, 2017.

[411] Kang Zhao, Xi Wang, Mo Yu, and Bo Gao. User recommendations in reciprocal and bipartite social networks–an online dating case study. *IEEE intelligent systems*, 29(2): 27–35, 2013.

[412] Kou Zhongbao and Zhang Changshui. Reply networks on a bulletin board system. *Phys. Rev. E*, 67:036117, Mar 2003. doi: 10.1103/PhysRevE.67.036117. URL `https://link.aps.org/doi/10.1103/PhysRevE.67.036117`.

[413] Tao Zhou, Jie Ren, Matúš Medo, and Yi-Cheng Zhang. Bipartite network projection and personal recommendation. *Physical review E*, 76(4):046115, 2007.

[414] Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40, 2020.

[415] Hongqiang Zhu. Countering covid-19-related anti-chinese racism with translanguaged swearing on social media. *Multilingua*, 39(5):607–616, 2020.

[416] Zhiguo Zhu, Jingqin Su, and Liping Kong. Measuring influence in online social network based on the user-content bipartite graph. *Computers in Human Behavior*, 52:184–189, 2015.

[417] Caleb Ziems, Bing He, Sandeep Soni, and Srijan Kumar. Racism is a virus: Anti-asian hate and counterhate in social media during the covid-19 crisis. *arXiv preprint arXiv:2005.12423*, 2020.

[418] Vinko Zlatić, Miran Božičević, Hrvoje Štefančić, and Mladen Domazet. Wikipedias: Collaborative web-based encyclopedias as complex networks. *Physical Review E*, 74 (1):016115, 2006.