The Wikipedia News Network: Understanding Collective Response to Current Events Through the Internet's Encyclopaedia

Thesis submitted in partial fulfilment of the requirements for the degree of DPhil in Information, Communication and the Social Sciences in the Oxford Internet Institute at the University of Oxford

Patrick Gildersleve
Oxford Internet Institute & Green Templeton College, University of Oxford

Supervisors: Prof Taha Yasseri and Prof Renaud Lambiotte

Trinity Term 2021

46,581 words

Abstract

Wikipedia is the primary authoritative information resource on the web for billions of people, and perhaps the most important reference work in human history. Its modernised notion of the encyclopaedia is a widely accessible and rapidly updatable record for both historical knowledge and current events. The trove of available trace data from its users' browsing patterns also means that the site acts as an appealing, representative barometer for wider patterns of collective attention.

Studies of news media are often concerned with news values—properties of events—that together define newsworthiness—how likely an event is to be chosen for news coverage. In essence; what makes an event news? Traditional study, however, has frequently focussed on data directly from news media, journalists, or even news sharing on social media rather than independent "extra-media" data. This raises concerns of selection biases, platform-specific network effects, and endogeneity. To counter these issues I turn to Wikipedia and the way its users access its information for an audience-centric perspective on current events, news values, and newsworthiness.

In this thesis I conduct three studies designed around understanding collective response to current events. In the first study, I explore how events are represented on Wikipedia and accessed by is audience. To do this I develop a temporal community detection approach towards identifying the topics primarily browsed by users. Secondly, by combining the extra-media data of Wikipedia with a matched news article database I address and reformulate foundational hypotheses of news values and newsworthiness theory. In the third study, I more directly analyse the peaks of collective attention that emerge when a subject is in the news. I develop a time series clustering algorithm to identify the characteristic shapes of these peaks and propose a forecasting model for their growth and decay.

This thesis makes significant contributions to research on online representations of current events, modern understandings of news value theory, and models of collective attention dynamics. By not directly studying news media itself, this 'altmetric' style approach to studying the impact of news events is an important practical and theoretical advancement. It also has deep implications for newsrooms employing editorial analytics and for how platforms, including Wikipedia, serve content on current events to users. More widely, since newsworthiness and attention influence what information is popular—even acceptable—online, studies from novel, independent settings such as this are crucial to informing future journalistic and online speech policies and regulations.

Acknowledgements

It's perhaps a strange way to begin an acknowledgements, but I would like to start with a short confession. Many people get into editing Wikipedia through some casual or professional interest in a subject, with a motivation to contribute towards a common good intellectual resource. As a teenager I made a rather different start, instead opting for Wikipedia vandalism, usually of articles about football players I didn't like. Eventually my interest in the free knowledge won out over what amounted to my 'rebellious' phase as an adolescent. I would like to apologise to the community and thank the editors who fixed my contributions. Hopefully this and future work of mine will make up for my mistakes. Many people have been responsible for setting me on the straight and narrow and fostering my personal and academic development. Its been a privilege and a pleasure to get to know all of you.

Firstly, thank you to my supervisors Taha Yasseri and Renaud Lambiotte for your patience, guidance, and thoughtful feedback. It has been a challenging but truly rewarding experience jumping between academic disciplines; you have kept me grounded and helped make sense of so many jumbled thoughts. Thank you also to Greg Taylor for stepping in to manage my project through the final stages. I'd like to thank my assessors over the course of this DPhil: Jonathan Bright, Brian Keegan, Nicola Perra, and Scott Hale. I requested you as assessors out of admiration for and being inspired by your work. Your feedback in getting the thesis to where it is now has been invaluable. I thoroughly enjoyed collaborating with Ryota Kobayashi and Takeaki Uno in my research and feel it's only appropriate to express my gratitude with (more) sake. I learnt a lot through my experience teaching with Bernie Hogan and the many MSc students who passed through the OII, who I would also like to thank. I also extend my thanks to the academic support staff who have shepherded so many of us through our time at the OII—particularly Laura Maynard and Victoria McDermott.

Having been at Oxford for rather too long I've seen many friends come and go. Rest assured that you have all been uniquely treasured and influential at different points over the years. You are all so much of the reason I have enjoyed my DPhil so much, and also the reason this has taken so long. Thank you to the Worcester boys—the biggest names and the best group of mates I could ask for. There are also those that have stuck around since undergrad, most notably 'The Crib': Julius, Jake, Flo, Stu, and Jess. There's no one else I'd rather be quarantined with on an island through a pandemic. A special and fond mention to Julius; if you're reading this, don't forget to put the bins out on Thursday evening. To everyone at WCAFC, thanks for all the incredible memories from the hallowed turf of Worcester College and Iffley Road. To those from GTC, thank you to Alex and Is for the friendship and food (I think in that order). Thank

you to Jake for the truly productive, madly puzzling, and deeply pointless quiz experiences? I'd also like to thank Sumin for insisting that I remove her from the 'OII friend' section of this acknowledgements.

Through the OII, I have met some brilliant minds and better friends. Thank you to Siân for the terrible dating advice, Kate for indulging my dudebro-est interests, Yung for always keeping me on my toes, Josh for facilitating some punhealthy habits, Julia and Cailean for taking my teaching advice and integrating me into their social networks, and Sanna for being such a dependable friend and colleague. There are so many more that have enriched my experience at the OII—you are all truly appreciated.

To Mum, Dad, and Kate, you're the reason I've got to where I am today. I can (theoretically) stop being a lazy student now and become a real doctor. Thank you for everything.

Finally, to Nayana. Your constant support has kept me going while doing this crazy project in crazier times. Thank you for believing in me through it all. Having our lives further converge in academia has been a true blessing (not just for the thesis copy editing). You continue to intrigue and inspire me. I could not have done this without you.

Contents

List of Figures

List of Tables

List of Abbreviations

- **AIC:** Akaike Information Criterion
- **AMI:** Adjusted Mutual Information
- **API:** Application Programming Interface
- **BIC:** Bayesian Information Criterion
- **HTML:** Hyper Text Markup Language
- **IDC:** Impulse Decay Chain
- **IQR:** Interquartile Range
- **KDE:** Kernel Density Estimate
- **KSC:** K-Spectral Centroid
- **LSTM:** Long Short-Term Memory Network
- **MAE:** Mean Absolute Error
- **MSE:** Mean Squared Error
- **NLP:** Natural Language Processing
- **OLS:** Ordinary Least Squares
- **RNN:** Recurrent Neural Network
- **RSS:** Really Simple Syndication
- **SD:** Standard Deviation
- **SIR:** Susceptible-Infected-Recovered
- **WKSC:** Weighted K-Spectral Centroid

Chapter 1

Introduction

Since its founding in 2001, Wikipedia has grown from a simple, often dismissed, 'Web 2.0' based dream of shared knowledge to the Web's primary authoritative information resource for billions of people. Its emergence has overhauled traditional conceptions of the encyclopaedia from a slow moving tome to an accessible, rapidly updatable real-time record and audience barometer for both historical knowledge and, crucially, current events. Wikipedia is thus a unique, intriguing site of study for audience reception of news events—reflective of collective audience trends, yet somewhat divorced from the the influence of journalistic and content delivery processes. This sets up the core of the thesis; an examination of how news events are represented on Wikipedia, and conversely what Wikipedia tells us about news and the audience's collective response to it.

News media is of course a well trodden research area, but rapidly changing geopolitical situations and dramatic technological advancements that affect the production and consumption of news mean it demands ever more academic scrutiny. It is thus important to understand how events attract news and public attention in our evolving infosphere. Online environments are increasingly influential in the news process yet also offer enhanced opportunities for studying data on how the news audience responds to current events. This also allows us to revisit core theory in news media studies, up to now often focussed on data on or from journalistic output. In this thesis I focus on news values—inherent properties of events—and how they contribute to and define newsworthiness—how likely an event is to be selected for news coverage. Naturally, this draws interest from and has applications and implications directly for journalism. However, the

issue of what makes news news is much wider. Amid fractured times for social interaction online "newsworthiness" has come to be a central part of policies for speech on the Internet. Facebook's explicit "newsworthiness exemption"—allowing newsworthy content, often from politicians, even if it breaches other community standards—is the clearest example (Kaplan & Osofsky, 2016; Clegg, 2019), although its future existence is not guaranteed (Heath, 2021). Related policies on noteworthy content are also employed by Twitter (*Defining public interest on Twitter*, 2019) and Google/YouTube (Overly, 2019), as well as in European Union 'right to be forgotten' cases (*General Data Protection Regulation*, 2016). If newsworthiness is to act as a qualifier for expression on the web, then we should not be relying on the opaque definitions of newsworthy content or on the limited data provided directly from the social media platforms that supposedly police it. This calls for renewed academic focus on how news on current events is represented online and crucially how the audience responds to it.

Much concern is directed towards how representative activity on social media and the rest of the web is of wider societal attitudes, particularly bearing in mind the effects of curation algorithms shaping what and how users encounter. In truth, activity on the platforms are important in and of themselves and that distinctions between online and 'real-life' are misguided—so much of 'real-life' occurs on the web. It is of course important when drawing conclusions, however, to be conscious of the biases and insufficiencies surrounding the data and platform being studied. This can be difficult when data is privately owned and controlled, and business secrets, black box algorithms. Wikipedia does have many biases, notably on gender (Reagle & Rhue, 2011; Graells-Garrido, Lalmas, & Menczer, 2015), geography (Graham, Hogan, Straumann, & Medhat, 2014; C. Osborne, Graham, & Dittus, 2021), and race (Adams, Brückner, & Naslund, 2019). However, unlike many of the other dominant web plaforms these are open and can be interrogated. In addition, any effect of the platform shaping the actions of its users is relatively minor. This positions Wikipedia as a "natural" setting for digital social research.

Wikipedia has been heralded as "last best place on the internet", "a welcome oddity on the modern internet", and "one of the internet's saving graces, lifting the online world out of

the hellish hole that it often slithers down" (Cooke, 2021; Chase, 2021; German, 2021)—indeed research is starting to explore why this contrasts to much of the rest of social media (Yasseri & Menczer, 2021). Its basic mission is quite simple, to provide free, widely accessible encylopaedic knowledge to the world. This has been achieved (and continues) through the monumental effort of volunteer editors, across over 56 million articles (6 million in English, the largest Wikipedia), in nearly 300 language editions (*List of Wikipedias*, 2021). Indeed, anyone can edit or create an entry on the website (even without an account), provided it follows the basic principles and notability criteria. Each article on the website is produced as a hypertext 'wiki' page where any individual can use simple markup language or increasingly user-friendly rich text editor to make a contribution. The encyclopaedia is searchable, and articles are typically structured in sections, with occasionally a short table of contents and/or infobox where key information is displayed. As a webpage, each article is also able to take advantage of images, audio, and video that can be directly embedded. More importantly however, the very structure of the encyclopaedia is built around hyperlinks that can take a user directly from one article to another (and many more beyond that, down encyclopaedic rabbitholes). Rather than a global hierarchical table of contents, with articles classified into individual categories and subcategories, the content is organised around the emergent network structure that arises from related subjects being more tightly linked (non-exhaustive, overlapping category tags do also exist, but are not a primary means of organisation, or at least of information access).

As an online platform, Wikipedia is hugely important and relatively unique as an open, standardised, evolving setting for the recording and access of knowledge. The open, quick, collaborative nature of editing, which naturally has resulted in criticism for the quality, accuracy and possible bias of content (Black, 2010; Rosenzweig, 2006; Das, Lavoie, & Magdon-Ismail, 2013), is exactly what is responsible for the website's growth and popularity, as users engage in both content production and consumption. The work of editors that collaborate and conflict in the recording of vast structured accounts of the world, together crucially with how the public view these articles means that research on Wikipedia has great reach beyond simple application

to the platform.

This rise to prominence of Wikipedia signals a shift in the long history of encyclopaedism. Where encyclopaedias have historically been scholarly reference works, reflecting relatively stable knowledge and updated over periods of years, digital technology—culminating in the participatory encyclopaedism of Wikipedia—has transformed them into fast moving knowledge bases that also rapidly reflect current events. On the other side of the same coin, the digital preservation of news and its integration with wider information bases on the web has added a permanence to public news records not previously seen. Wikipedia acts as a shared space for the public for the recording and revisiting of news events and/or the actors, places, actions involved (even if there is no record of the event itself on Wikipedia). Events themselves, and much of the online reaction to them, are often transient, ephemeral phenomena. However, the recording and remembrance of them on Wikipedia from the outset is a more permanent pursuit (individually as well as in their accumulation e.g. towards a person's public profile). Of course, editors can extend, revise, and revert the content on Wikipedia, but the intent is for a single, distilled, quasi-permanent record that can be revisited—an encyclopaedification of news. The duality of Wikipedia as the "encyclopaedia with current events" (Keegan, 2012) means it represents a convergence of these disciplines.

As already alluded to, whilst the way in which Wikipedia editors organise and collaborate in how current events are recorded is certainly of research interest and impact, Wikipedia's influence and importance is far greater when considering the hundreds of millions of monthly users who only view the content on the website. At the time of writing, Wikipedia is the world's 13th most popular website (*Alexa Top Sites*, 2021), yet in everyday life and academic research, other large online platforms receive far more scrutiny and recognition for the role they play in society, and as indicators of wider trends. Moreover, many of the other most popular websites are either social networks or search services, fulfilling similar roles to the average user. Wikipedia's role at the forefront of the web is relatively unique. Beyond its foreground role, plenty of sites rely on Wikipedia's content in powering their own services. For example, platforms such as

Facebook, Google, YouTube, Twitter, Amazon Alexa, and Apple Siri use Wikipedia in producing their own knowledge graphs, informing automated search results and infoboxes, verifying of notable persons, and directing their users to authoritative sources on issues of conspiracies and misinformation (Matsakis, 2018; Withers, 2018; Perez, 2020; TwitterInc., 2020; Vincent & Hecht, 2021)

Wikipedia has a number of principles and policies that guide how it is run by the parent organisation Wikimedia, and to how Wikipedia administrators and editors contribute to and manage the content on the site. The barrier to entry for editing is admittedly low (in order to encourage more editors), however there is nothing in the way of rules, policies, or prescribed behaviour for the vast majority of users who do not edit the site and simply view the content. Many of these users browse Wikipedia for its original 'intended' purpose as an encyclopaedic reference work, finding information on science, history, culture etc. However, this much larger audience also find utility in the site far beyond the intended purpose of a traditional encyclopaedia, instead often flocking to information on current events, sports, and entertainment articles in great numbers around the time of news breaking. This attention towards events tends to be sharply clustered around the moment in time the event occurs. On Wikipedia, this translates to surges in edits and page views, particularly if the event is significant.

Whereas previously news had to wait for the next week's/morning's press to be delivered, now many millions of people are able to find news reports and collected, contextualised supporting information almost instantaneously. The tension between the traditional encyclopaedic elements of Wikipedia and the tendency for editors to focus on current events is reflected in how the platform is primarily used by its non-editing audience. Where Keegan writes that "Our understanding of how Wikipedia works is profoundly incomplete because the vast majority of studies are grounded in analyses of the entire corpus of articles co-authored under conditions of relative stability" (Keegan, 2012) is only further stretched when considering the wider audience whose appetite for information on current events is as, if not more, extreme. We must understand the representation of current events on Wikipedia through the lens of how hundreds of

millions of regular users access the information.

A typical user is unlikely to see Wikipedia as a news source, but more of a "news backgrounder" (*How the Current events page works*, 2021): They may find out about some event from dedicated news sites or social media, or even broadcast media or word of mouth, and go on to further research the matter using the structured encyclopaedic knowledge of Wikipedia when the news report does not suffice. The "Twitter-to-Google-to-Wikipedia" routine serves users requiring deeper information on "authoritative source working to verify an important news development" (Seward, 2009). Wikipedia's integration into both the practices of recording and serving content current events and places it as an underrepresented barometer for audience reception to news. The reasons these events draw interest in the public and on Wikipedia run parallel with the way journalists cover an event based on its 'news values', which ultimately contribute to its 'newsworthiness'—how likely it is to be covered. Studies of journalism typically look directly to news coverage and how it is consumed by an audience in order to judge and measure the effect of news values. However, as identified by (Rosengren, 1970), this introduces a survivorship bias into the work and information from an independent resource, also identifying the events that do not necessarily make the news—"extra-media data"—is required. For news value studies, Wikipedia as an independent encyclopaedic record largely fulfils Rosengren's requirement for extra-media data. This allows one to draw conclusions beyond the platform to how news itself represents events in the eyes of its audience.

Naturally, there is great interest in the aforementioned flurries of activity and attention that grow and fade around the time of events. How quickly and strongly a news subject enters and leaves public consciousness is a question of collective attention. Research on peaks in collective attention is often based around the modelling of phenomena such as shared posts or viral trends on social media, rather than the response of a population to some external force such as the temporary relevance of a news event. Page views on Wikipedia are a measure of intentional, relatively costless action. Users typically seek out a Wikipedia article via search (or second-order clicks through continued browsing on the site), rather than have the content presented to them

by other accounts or the effect of some feed algorithm. In addition, unlike edits, page views are relatively cost-free—the user does not have to spend time producing content or sending public signals—the information is only a click away.

This clarification is important. Firstly in that the page views measure surveys a wider audience of information consumers (or 'lurkers'), rather than producers. However, possibly more importantly—and in-keeping with previously mentioned theme of Wikipedia's independence—we measure *conscious* attention towards a subject, and are less subject to how specific outlets/individuals output content or how the means by which the content is curated and delivered to users (through editorial or algorithmic means). 'Collective attention' as a catch-all term for very different kinds of engagement online is then rather clumsy. This is not to say that these kind of behaviours are not at all engaged with or measurable on other platforms. However, the oft-used aggregated engagement metrics that are visible to both users and (external) researchers are of a very different kind of collective attention. The stripped back measures for collective attention we can glean from Wikipedia are appealing both in their popularity and simplicity.

The three studies in this thesis are designed around measuring and characterising audience response to current events, as recorded on Wikipedia. In preparation for this, I first survey the relevant literature on Wikipedia, news, and collective attention, describe the datasets I use, before setting out to build on the field with three substantive chapters, each addressing a central research question, outlined as follows.

- **RQ1:** How are current events represented in the knowledge structures and access patterns of Wikipedia and its users?
- **RQ2:** How are traditional conceptions of news values and newsworthiness of events reflected in extra-media data?
- **RQ3:** How can we model and predict peaks of collective attention towards news events?

In the first study I set about the detection and sampling of collective online reactions to current events on Wikipedia. I develop a temporal community detection approach to identify

the groups of articles, and larger topics, that are concurrently accessed by users in response to events in the news. The process incorporates both long-term established hyperlink structures as well as the short-term dynamics of page views on a sample of one year of entries from the Wikipedia current events portal. The relative importance of page view dynamics vs links in this process reveals the topics on Wikipedia that are dominated by news reaction compared to the more established collective knowledge structures.

Secondly, I use the previously detected networked news reactions from the first study towards examining theory on news values and newsworthiness. I take advantage of Wikipedia's status as extra-media news data and extract features representing news values for each event from Wikipedia, combining them with a matched dataset of over 100,000 news articles. I test key relations between news values, between news values and news topics, and between news values and newsworthiness, addressing and reformulating foundational hypotheses in news value theory in the process.

In the final study, I more directly analyse the link between news events and peaks of collective attention. Understanding how attention towards events rises and falls is critical for news media in anticipating what stories attract interest and for how long. I first develop a time series clustering algorithm to identify the characteristic shapes of peaks in attention towards news events. I also propose a model for the growth and decay of these peaks when driven by some exogenous force, and how they may feed into longer term changes in attention levels. I adapt both of these approaches towards predicting the shape of peak decays, with results close in performance to a more complex neural network approach.

Taken together, these works build on existing literature on Wikipedia, news, and collective attention. I provide a much-needed timescale-sensitive approach to topic detection on Wikipedia, make important contributions to news value theory based on data from a truly novel setting, and develop widely applicable computational methods for understanding and forecasting collective attention. Ultimately, my work seeks to understand collective response to news events in a context detached from challenging issues of endogeneity one encounters in regular social and

news media.

Methods and results from this thesis may be applied towards understanding and improving how Wikipedia, and indeed other platforms, serve content on current events to best suit users' informational needs. The study of events through Wikipedia's independent knowledge base can also provide more advanced forms of 'altmetric' style editorial analytics in the newsroom. Journalists, and even those involved in popular events, may use the findings on collective attention to determine how to best respond to and harness the nature of sharp peaks of interest towards news subjects. This work also bears important implications for the future of research on events and news theory beyond direct studies of news media. Understanding how events are considered newsworthy is clearly critical to journalists, but also increasingly to technology policy makers and legal teams in setting out the bounds of news online, and consequently acceptable speech. Utilising extra-media trace data to this end, such as from Wikipedia, is a critical and fruitful frontier that must be tackled.

Chapter 2

The Record of Wikipedia

In this chapter I first introduce background work on the state and overall growth of Wikipedia as a network, before then charting the emergence and convergence encyclopaedism and news. I expand upon theory on news values and newsworthiness, as well as cover work that researches news events online, particularly Wikipedia. Next, I examine work on the dynamics of attention and navigation online. Finally, I detail how I will address limitations in the existing literature in my pursuit to understand the impact of current events.

2.1 The Wikipedia Network

2.1.1 The Modern Encyclopaedia

Wikipedia, the "sum of all human knowledge" according to its co-founder Jimmy Wales (*Wikipedia Founder Jimmy Wales Responds*, 2004), is the world's largest encyclopaedia and 13$^{\text{th}}$ most popular website (*Alexa Top Sites*, 2021), serving information to hundreds of millions of users each month (Summers, 2013) across ∼300 language editions containing a total of around 50 million articles (*Wikipedia Statistics*, 2021). Users participate in both content production, through the volunteer editing of articles (both in collaboration and conflict), and the consumption of this information. Unlike traditional encyclopaedias, information on Wikipedia is connected to other related content in the encyclopaedia via hyperlinks that users may traverse. This means that networks of information form around particular topics and that any article may be understood by the context of its neighbours. It is hard to overstate how remarkable a collective achievement

the vast knowledge base is, despite criticism relating to issues of quality, accuracy and possible bias of content (Black, 2010; Rosenzweig, 2006; Das et al., 2013). A wealth of data on Wikipedia is made available courtesy of the Wikimedia Foundation, its host charity, drawing significant research interest to this fascinating social system.

2.1.2 The Growth of Wikipedia

The current state and overall evolution of Wikipedia is well studied and modelled. In "Measuring Wikipedia", Voss (2005) provided an early work on the state and growth of several language Wikipedias. He finds that after a linear phase, Wikipedias grow exponentially, as measured by counts of words, articles, links, active users, and database size, with different rates per language and that article sizes are log-normally distributed with a median that grows linearly in time. Additionally, the network of articles is observed to have scale free distributions for incoming, outgoing and broken links. Another early research effort from Almeida, Mozafari, and Cho (2007) further considers the dynamics of the creation and update of articles on Wikipedia. They find that overall growth of Wikipedia is exponential due to the rapidly increasing user base, however, average productivity of users is decreasing. Finally, they observe that there exist two distinct groups of contributors; a small number of editors contribute to thousands of articles whereas the clear majority of editors contribute to far fewer than this. More recently however, Wikipedia's growth has slowed, as studied by Suh, Convertino, Chi, and Pirolli (2009) who identify increasing resistance to new articles from occasional editors, greater overheads for coordination and bureaucracy, lack of easy topics to write about, and quality of editing tools available to users as limiting factors in the growth of the website. This overall trend has continued, and is broadly replicated across different languages.

2.1.3 Knowledge in the Network

The nature of Wikipedia as a website, with hyperlinks linking related pages, means that by design it creates a structured network of articles where information on topics does not simply exist in isolation, but is contextualised as part of the wider world of information. Indeed, to the

historian Rosenzweig's chagrin "the problem of Wikipedian history is not that it disregards the facts but that it elevates them above everything else and spends too much time and energy (in the manner of many collectors) on organizing those facts into categories and lists" (2006). This representation of concepts, relations, and categories can be considered a form of ontology, more specifically a knowledge graph or knowledge network[1]. Knowledge networks encode various types of relations between entities, and are used across a wide range of fields (Miller, 1995; Ashburner et al., 2000; Lehmann et al., 2015). For the purposes of this thesis, I restrict the types of relations between entities (articles) to simply hyperlinks, by far the most evident relational feature to the regular Wikipedia user[2]. The way that Wikipedia's knowledge network is created, structured and interpreted has drawn academic interest, both in terms of its properties as well as how it can be utilised.

Work that covers Wikipedia as a complex system in more depth includes "Preferential attachment in the growth of social networks: the case of Wikipedia", Capocci et al. (2006) examined properties and growth of the 'bow tie'-like structure of Wikipedia. Growth of the network is successfully modelled with a preferential attachment based mechanism, qualitatively reproducing in/out degree distributions and neighbour degree correlations. Zlatić, Božičević, Štefančić, and Domazet (2006), in "Wikipedias: Collaborative web-based encyclopedias as complex networks" used network metrics for a cross-cultural comparison of the growth of the 11 largest language Wikipedias by undirected links. They show many network characteristics (degree distribution properties, growth, topology, reciprocity, clustering, assortativity, average shortest path lengths, and triad significance profiles) are common to the different Wikipedias, showing that generally the growth process of Wikipedias is universal. They do however observe some discrepancies between different language Wikipedias when administrative decisions are made on editing standard practices. The communities of editors are not only responsible for 'minor' issues of page content, but their aggregate behaviour, policies, and editing practices can have a global effect

[1] A prominent example not to be confused with is Google's proprietary Knowledge Graph: developers.google.com/knowledge-graph

[2] The sister project Wikidata, far more focussed on creating knowledge graph with many more kinds of relation between entities, has also experienced a surge in popularity among enthusiasts in recent years.

on the article network.

2.1.4 Applications of the Network

Milne and Witten (2008), in "Learning to link with Wikipedia", use this network on Wikipedia to develop a system to automatically cross reference documents. The machine learning algorithm extracts key concepts from plain, unstructured text, using Wikipedia articles as training data to decide what terms should be linked. The Wikipedia articles that are linked to are selected based on not only semantic features but also on the relatedness of ambiguous terms in the text to unambiguous terms in their surroundings. This selection is based on the hyperlinks between corresponding Wikipedia articles—the structured knowledge on Wikipedia. This work, in effect, may be used to further link existing concepts in Wikipedia's knowledge network. Ciampaglia, Shiralkar, et al. (2015) have approached knowledge networks in their paper "Computational Fact Checking from Knowledge Networks", leveraging relations between topics from the DBpedia project and the ease of navigating between them to fact check statements not expressly present in the network. The work demonstrates the power of the information revealed simply by the relations between items in a knowledge graph. In Wikipedia's case, the network of hyperlinks is much more than just a navigation system. The network itself, from individual links to larger scale structure, is a form of user generated implicit knowledge.

2.1.5 Network Editors

Wikipedia is a network built by editors who will vary in expertise, individual biases, and cultural background, amongst other traits. What is constructed by different groups, as well as how it is built, better informs us of the information itself. In "It's a Network, Not an Encyclopedia: A Social Network Perspective on Wikipedia Collaboration", Kane (2009) effectively considers networks of articles as linked by common editors and relates the degree and eigenvector centralities of articles to a prior ranking of article quality. The positive relation found between both these centrality measures and article quality indicates that both rich local collaborative environments (through degree centrality) as well as, more strongly, strong collaboration between collaborative

environments (through eigenvector centrality) produce high quality articles. Aragon, Laniado, Kaltenbrunner, and Volkovich (2012) have explored how structured knowledge in the form of networks of famous people is recorded by different language Wikipedias. Cultural differences are explored by comparing these networks. Whilst global social network measures are similar and common structures are present across all languages, the social networks in Wikipedias in languages from geographically or linguistically close communities are indeed to found to be more similar. This indicates that the way in which knowledge is collectively recorded and structured on Wikipedia reveals information about the cultures that themselves construct it.

Wikipedia's properties as a knowledge network, with important information encoded in its hyperlink structure, not just in article content, make it essential to study it as such. The knowledge network structure offers enhanced opportunities to study news media and collective attention online, which much existing research does not take advantage of.

2.2 The Twin Histories of Encyclopaedism and Journalism

2.2.1 Shifting the Boundaries

To study Wikipedia as a modern encyclopaedic news resource, we must understand how the formats of encyclopaedism and news have progressed to this point. Wikipedia is just the latest development in the use of encyclopaedia as a format for knowledge exchange spanning over 2000 years. The content, distribution, and organising principles of encyclopaedias have morphed through time, sensitive to the technological affordances and societal demands of the period and locale. Keegan (2012) appeals to the theoretical lens of "boundary work" and "boundary objects" for how "Journalism and encyclopedism did not inevitably collide because of some intrinsic similarities, nor did Wikipedia come about solely because of the invention of new technologies. Rather, the actors and institutions invested in the production of increasingly overlapping types of knowledge redefined the boundaries of their work and established new identities in the face of technological and social discontinuities that demanded they transform and adapt". Encyclopaedia editors have responded to changing times in strategically demarcat-

ing the realms of encyclopaedic knowledge and the encyclopaedia's cultural status. With this, the size and demographics of encyclopaedias' audiences, and contributors, have shifted. Where previously, encyclopaedias' authority was conferred to it by close relationships to powerful institutions, Wikipedia's emerges from the messier "procedural rhetoric" of community consensus (McGrady, 2020). Nevertheless, there are still clear commonalities between encyclopaedic production historically and on Wikipedia, which Loveland and Reagle (2013) explore through three different modes: compulsive collection, stigmergic accumulation, and corporate production. It is therefore instructive to trace how Wikipedia has built on the encyclopaedic tradition.

2.2.2 Western Encyclopaedic History

The roots of Western encyclopaedic history can be traced to ancient Greece, with partial records of the work of Speusippus from 4th century BC compiling his uncle Plato's teachings. Influential Roman approaches such as from Marcus Terentius Varro and subsequently the Naturalis Historia of Pliny the Elder in particular, which remained highly cited for the next 1,500 years. For the most part, works from the Middle Ages were still hand-copied and saw limited reach beyond the wealthy patrons and monastic scholars for whom they were written e.g. Bartholomaeus Anglicus' popular De proprietatibus rerum ("On the Properties of Things") (1535) intended for student friars (Seymour, 1992). The invention of the printing press meant that encyclopaedias in the Renaissance were able to be distributed more widely. The rapidly increasing quantity of information to be covered also forced editors to wrestle with opposing philopsophies of inclusionism and exclusionism for their content, the exclusionary philosophy bringing in a brief spell of popularity for 'encylopaedic dictionaries'. Further controversy over how editions were volumised and released (alphabetised or by category, and simultaneously or periodically) also haunted the format. By the 17th and 18th centuries the systematic compilation, organisation, and classification of knowledge was both a necessary and revered undertaking. Efforts such as the Encyclopédie of Denis Diderot and Jean le Rond d'Alembert (1754) and the Encyclopædia Britannica (Smellie, 1771) reaching a form similar to the modern print encyclopaedia. The form was refined and popularised through the 19th and 20th centuries with door-to-door sales tactics

and public libraries putting a range of digested knowledge in the hands of many more literate people than ever before.

2.2.3 The Wider Reach of Encyclopaedias

Encyclopaedia also developed outside of the Western tradition, with their own organising principles and readership demographics. Historic Arabic encyclopaedias from the 9th and 10th centuries could initially be characterised as either for means of expansion of cultural knowledge or as administrative tools. The influential Persian encyclopaedia Mafātīḥ al-'Ulūm ("Keys to the Sciences") primary partition was along the line of indigenous versus foreign knowledge (*History of Encyclopaedias*, 2021). In China, encyclopaedias grew out of their precursors 2000 years ago which were composed of literature anthologies and dictionary material. The "period of the encyclopaedists" spanning the tenth to seventeenth centuries was marked by the state sponsoring hundreds of scholars to consolidate knowledge in encyclopaedias (Murray, 2009). The largest and most notable of these is the Yongle Encyclopedia of 1408, which with almost 23,000 folio volumes was the largest encyclopaedia in history until it was surpassed by Wikipedia. English Wikipedia has since been overtaken by two more Chinese state-sponsored encyclopaedia—Baidu Baike and Baike.com. However, there are claims of foul play with regard to copyright violation of Wikipedia's content, which echoes other copying disputes between historic fledgling encyclopaedias (Nystedt, 2007). The content of individual encyclopaedias can rapidly spread to other knowledge bases, as is being found now with information from Wikipedia being found in other online encyclopaedia, smart assistants, scientific papers, and even Wikipedia hoax content sprouting false information elsewhere on the web (which then perversely are used as references for the original Wikipedia hoax) (Randall, 2014).

2.2.4 A Format Shift: Rise of the Digital Encyclopaedia

Nowadays, we exist firmly in the era of the digital encyclopaedia. Microsoft's landmark Encarta software, first available on CD-ROM in 1993, often was bundled with new PC purchases. Encyclopaedia Brittanica responded in kind the following year with its own CD version as well as

a subscription based online version which provided hyperlinks to external destinations on the World Wide Web. Established boundaries of the encyclopaedic format were beginning to be redrawn as knowledge became more widely, rapidly accessible and enmeshed in a wider information ecosystem. Further efforts exploring the capabilities of the World Wide Web include the Stanford Encyclopaedia of Philosophy which gave File Transfer Protocol abilities to its authors (Hammer & Zalta, 1997) for the anytime update of its content, or more direct collaborative precursors to Wikipedia such as Nupedia. In 2001 Wikipedia was launched, its freely and easily editable content, together with malleable customs and practices harnessed the enthusiasm of a web community seeking to make their own contribution to a digital record of the world.

Wikipedia's interlinked structure performed well in the rankings of early search engines (Keegan, 2013) as it grew to be a foundational information resource of the Web. Issues of ordering and categorisation are far less apparent in a searchable system with flexible category tags and no requirement for a global hierarchy of information on the site. The prominent, instantly accessible and editable website meant encyclopaedias have never been so dynamic. An early turning point that shaped Wikipedia's editorial policies occurred with the September 11th World Trade Center attacks. Keegan (2013) explores how this created the "encyclopaedia with current events" as editors wrestled with how to include such a fresh, evolving, yet clearly historic moment in an encyclopaedic record and set a precedent for future events. This turn meant, in addition to the steady accretion of historical knowledge, Wikipedia was provided with a constant flurry of new material and users seeking information on current events. Keegan argues how current events drive the bulk of Wikipedia traffic and edits and signal a fundamental shift of encyclopaedias away from the slowly changing and spreading content of even the modern printed encyclopaedia era.

2.2.5 Encyclopaedic Audiences

Critical to this thesis is the question of not just how content Wikipedia is produced and represented, but to how the information is accessed and received by its audience. Clearly the reach and influence of encyclopaedias over the years has increased from the handwritten and copied

manuscripts, through the invention of the printing press and proliferation of the format, to now digital distribution methods. The intended audience for the encyclopaedias has correspondingly changed. It is hard to compare the exact reach of the encyclopaedias over time, given complicating factors such as uncertainty in historic copy numbers and library access. However, as an example, the hugely popular Encyclopaedia Brittanica has sold more that 7 million print editions in its 244 year run, with the final edition in 2012 commanding a fee of $1,395 (Channick, 2012). This still pales in comparison to the reach of Wikipedia; freely available to anyone with an Internet connection, garnering hundreds of millions of unique visitors per month, national blocks and data limits notwithstanding (efforts to mitigate these obstacles, such as Tor access and zero-rating Wikipedia, have been made). Nowadays, the searchable Web means Wikipedia (and other digital encyclopaedias) are just one part of a vast information ecosystem. The barriers to entry are comparatively low, and the boundaries around what is encyclopaedic knowledge are increasingly blurred. The encyclopaedic tradition has produced many influential examples of distilled, codified knowledge over time, but it is not a stretch of the imagination to say that Wikipedia is the single most popular and influential reference work in history. The encyclopaedia as a format has transitioned from a diverse genre of information resource available to educated individuals to, for all intents and purposes, a single globally accessible knowledge base. Studying the content, its production, and its access is a window to a distilled, structured reflection of the topics of interest of a huge audience.

2.2.6 A History of News

Where encyclopaedism has historically focussed on the slow accretion and centralisation of information, the purpose of journalism has been to quickly communicated transiently relevant messages to a wide audience. Initial approximations to news materials was restricted to letters between officials and dignitaries, and occasionally government proclamations could provide a unified message for the wider public. However, prior to wide adoption of the printing press, as a rule, news as a truly social enterprise for the general public was restricted to folk news—the oral communication of new information by travellers, pilgrims, and town criers. Networks

for news communication closely mapped to, even shaped, existing postal or trade networks, which were closely tied to the establishment of political power in a region. The large distances and limited locomotive capacity that news runners faced in a pre-industrial period meant that while this form of communication was far quicker than the years it may take to compile and write an encyclopaedia-like work, it could still be a matter of days, weeks, even months for key information on current events to spread across a region.

Again, it is the development of the printing press, and the ability to quickly copy (and eventually distribute) 'news' documents that ushers in true development for news media. Weekly periodicals started to appear in Germany after the Relation aller Fuernemmen und gedenckwürdigen Historien was established in 1605, and governments of other European countries began printing their own weekly newsletters (*Understanding Media and Culture: An Introduction to Mass Communication*, 2016). Information for public consumption was initially tightly controlled, with acts limiting news publication to approved presses and imposing high taxes on the medium. Newspapers experienced a boom in the late 18th and 19th centuries with printing developments such as the high speed presses together with a more liberal, literate public, as well as later the development telegraph for high speed communication. This was exemplified by the proliferation of news media in America. Freedom of the press (both editorially and financially) coupled with the ventures of entrepreneurs drove a variety of publications, notably the emergence of 'yellow journalism' (akin to the UK's tabloid journalism) which countered traditional news with sensationalism, increased sports and entertainment coverage, and increased advertising prominence.

2.2.7 Modern News Communication

News in the 20th century came to be characterised by radio and eventually television broadcasts as the profession increasingly catered to the mass market and received further advertising support. The establishment of professional codes, schools, and organisations for journalists cemented its cultural significance and position. Round the clock coverage (and the eventual emergence of 24 hour dedicated news channels) meant near instantaneous broadcast of events as they happened for any person who cared to tune in. News moved from a periodical publication

to an as-it-happens commentary on the world. This trend only intensified with wider Internet adoption in the late 20th and early 21st century where the rise of social media has shaken up the boundaries between news producers and consumers through citizen journalism, as well as with greater emphasis on the layman's response to events through social media. Distanced individuals have become far more involved in the as-it-happens evolution, reaction to, and recording of world events.

2.2.8 Digital Convergence

Historic variation in news and encylopaedism is based on the discrepancy in information distribution methods. The Internet has narrowed, even brought about a convergence in the time, format, and audience for these two initially very different knowledge exchange media. Through Wikipedia the encyclopaedia as a format has come to be a highly responsive trove of information on current events, and readily accessible online news archives (occasionally from citizens' records) comprehensively chart the events of the 21st century. News and encyclopaedic recording, together with the public's experience of them, have never been more similar. A key distinction remains the in the primary content producers. Encyclopaedism, as represented by Wikipedia, has moved from a closed scholarly pursuit to a collaborative effort of unpaid (and unpaying) individuals merely interested in the content.

The current state of affairs is now evocative of the social sensemaking practices around current events in the early periods of folk news. Indeed knowledge bases that power the artificial intelligence of major technology companies that have relied on Wikipedia, could be seen as the next stage in the new world of encyclopaedism and the beginning of a new formalisation of privatised encyclopaedic knowledge and the practices around it. This would mirror the move to the formalisation of journalism. In contrast, news and journalism as a discipline has comparatively retained much of its established institutional practices, authority, and influence (although concerns about the future of the industry are growing). The solid hierarchy of journalism may well have pushed interested, but uncredentialed, individuals towards less formal practises such as writing blogs, producing podcasts, and of course editing Wikipedia. Citizen journalism and

commentary on social media have refined rather than revolutionised the discipline, blurring its edges rather than overturning its core workings. The user generated content on social media in response to news receives much attention, both in the the academic everyday spheres, as a barometer for the events of the day but this approval has rarely been extended to the equivalent information production and consumption quietly being done in the encyclopaedia space.

Wikipedia as a convergence of encyclopaedism and journalism is well studied (though perhaps often overlooked), particularly on the side of information production. However, given the reach and influence of the platform, unlike any encyclopaedic or journalistic outlet before it, it is essential that we also study how its audience responds to how events play out through the access patterns and changing content of the platform.

2.3 News' Value

2.3.1 Establishing News Value Theory

Much of the work in this thesis revolves around the question of newsworthiness—what makes an event news? "News values" are typically defined as a set of criteria and properties of events that together define an event's "newsworthiness"—how likely it is to be selected for news coverage. News values can be an inherent property of an event, decided by individual journalists in framing a story, or even the product of larger choices and pressures at the organisational level. Galtung and Ruge's (1965) formulation has formed the basis for numerous other studies and originally included the news values of: Frequency, Threshold, Unambiguity, Meaningfulness, Consonance, Unexpectedness, Continuity, Composition, Reference to Elite Nations, Reference to Elite People, Reference to Persons, and Reference to Something Negative. There have been several alternative proposals, notably by Harcup and O'Neill (2001, 2017) who form the modern starting point, which also cover news values of Magnitude, Relevance/Proximity, Prominence, Organisation Agenda, amongst others. Thorough examination of the various proposals shows significant overlap between news value systems and Galtung and Ruge's original list, even with many newly proposed news values (Caple & Bednarek, 2013). Over time this news value literature has

developed with updated news value taxonomies sensitive to how the journalism has changed—particularly with the advent of the Internet. An alternative school of thought led by Pamela Shoemaker, inspired by theories of biological and cultural evolution, propose a more general news value classification based on values of significance and deviance (Shoemaker, Chang, & Brendlinger, 1987).

2.3.2 Testing the Theory

In addition to news values themselves, Galtung and Ruge propose several hypotheses describing news values' relation to each other, and to an event's newsworthiness, which have formed the basis for future studies in newsworthiness. Firstly, the 'complementarity' hypothesis; that events low in a given news value should exhibit high values in other news values—we should "expect a negative correlation between two news factors for a defined universe of news items" (Sande, 1971). Secondly, the 'additivity' hypothesis; that events which exhibit high values in several news values are more newsworthy, i.e., are more likely to be selected for news coverage. Finally the 'exclusion' hypothesis; that events which exhibit low values in several news values are less newsworthy. In practice, the additivity and exclusion hypotheses are mirrors of each other. The authors test and confirm these hypotheses on data from four Norwegian newspapers on three international crises. They also go on to speculate as to the universality across topics and cultural contexts of their work.

Despite seemingly superficial disputes between exact descriptions and representations of news values, the standard work—quantifying news values and measuring their existence and effect—typically proceeds through some content analysis and manual coding of news articles (usually dichotomous variables), often from a single source or on a single topic, followed by some sort of correlation/regression analysis. A paradigmatic example is Kepplinger and Ehmig's (2006) two component theory of news selection where news values and news factors are defined separately (thus far glossed over). The authors present a linear model for newsworthiness where 'news values' are the coefficients that describe the strength of linearly independent variables—the 'news factors'. The study finds different news values for different types of news outlet, and

whilst general news values are predictive of newsworthiness, there is no significant improvement in accuracy when considering outlet-specific news values.

2.3.3 Extra-Media Data, Read All About It

A key issue in Galtung and Ruge's original study is identified by Karl Erik Rosengren (1970). In only studying data from news media, it is difficult to evaluate the impact of news values on generating news stories from the universe of possible events, many of which go under- or un-reported—"extra-media data" is required. Many follow-up studies, even those challenging Galtung and Ruge, fall into the same trap (Bergsma, 1978; Schwarz, 2006; Piotrkowicz, Dimitrova, & Markert, 2017; Potts, Bednarek, & Caple, 2015; Westerståhl & Johansson, 1994). Alternative representations of news events or are sparingly featured in the literature, and if they do not rely on some list of a single class of event (Chang, Shoemaker, & Brendlinger, 1987) or audience behaviour on a single news service (Schaudt & Carpenter, 2009), they usually require some imaginative workaround in methodological design such as audience surveys or controlled laboratory experiments (Kepplinger & Ehmig, 2006). Of course, there is good reason for this. Despite Rosengren's simple instructions for requiring objective "data about both 'reality' and the media picture of reality" (Rosengren, 1970), descriptions of events 'independent' from news media are almost by definition impossible to come by. News is the mechanism by which information on important events is generated and communicated. And whilst we have seen the rise of a social form of participatory journalism in the 21st century, with live-streaming, live-tweeting, and other democratised user generated content, this is still subject to journalistic choices made by individuals, as well as the power of more traditional news media in both shaping the individual's worldview, and spreading their content. A worldview independent from news media barely makes sense. Schulz (1976) offers a lifeline in a response to Rosengren's critique, arguing that only different representations of reality can be compared, rather than the news media perspective to some objective telling of events[3]. To Schulz, news factors are hypotheses that guide

[3]Taken to its extreme, this would align with agenda-setting theory (McCombs & Shaw, 1972), where media (attempt to) influence and shape the reality, or realities, of their audience.

journalists' perception of reality, in turn informing decisions on an event's newsworthiness to an audience. This leaves room for comparison to alternative records of events, rather than strict objective records which as argued are rare and not comprehensive.

2.3.4 The Emergence of Digital Extra-Media Trace Data

Much of the early core theoretical work around newsworthiness is from a pre-Internet age. Not only have news factors themselves received updates as journalism has evolved, but perhaps we must reconsider the theoretical boundaries previously established. In particular, the existence, accessibility, and nature of wide ranging audience-centric extra-media data. This is not quite a gold-standard objective telling of reality as Rosengren might crave. However, the aggregated, collective perspective of the online news audience, relatively divorced from the journalist (or at least single sources of news media), is certainly appealing. News audiences' online activity leaves trace data on browsing patterns, page views, and opinions, which becomes only more relevant in an increasingly digital world. Moreover, beyond passive data traces, there now also exist active constructions of repositories of knowledge on current events online, without the same degree of journalistic prominence (despite some lingering commonalities such as moderation, partisanship, and gatekeeping). We now have ample extra-media data from news audiences, whose collective perspective on reality, and the corresponding audience-centric news values can be measured.

I select Wikipedia as a prime example of this extra-media data. The formation of and user behaviours on Wikipedia's network of articles is representative of wider news topics, news values, newsworthiness, and collective attention. The nature of the data, distinct from solely journalistic output yet community maintained and always up to date, is truly intriguing. I pursue an understanding of news through this platform and elaborate on the work around news and Wikipedia in the next section.

2.4 Wikipedia: Yesterday's News?

2.4.1 The Challenge of Encyclopaedic News

Wikipedia is not a news website. I can confidently assert that the majority of users do not use it as a daily news source, despite the presence of a short 'In the news' section on the home page. What Wikipedia is used for, however, is as a secondary information resource, which individuals use to further research and contribute towards topics they have encountered through other news media. Singer et al.'s (2017) survey finds that 13% of readers visit the site directly because of current events, and a further 30% visit due to wider media coverage[4]). Former editor-in-chief of the Encyclopaedia Brittanica Dale Hoiberg has criticised Wikipedia for its users' undue focus on news—"People write on things they're interested in, and so many subjects don't get covered; and news events get covered in great detail" (Waldman, 2004). Wikipedia, in breaking with traditional historical and encyclopaedic conventions by (somewhat unwillingly, at least unintentionally[5]) deferring to present day events, acts as a first port of call for many people to access (and contribute towards) further, perceived neutral, information on news stories they have encountered elsewhere. Keegan (2020) gives a history of the "encyclopaedia with breaking news", and how from its beginnings it has "leveraged the supply and demand for information about breaking news and current events into strategies that continue to sustain this radical experiment in online peer production". Nowadays, with concerns abound around disinformation and fake news, whilst it would be naive to claim Wikipedia is free from errors, bias, or misinformation, Wikipedia has largely resisted the controversies that have shaken other major online platforms. In fact, perceptions around the reliability of Wikipedia have only improved since its early days, teachers' and lecturers' warnings have grown increasingly futile, with various studies confirming its overall reliability on a range of subjects (Giles, 2005; Devgan, Powe, Blakey, & Makary, 2007; Fallis, 2008; Messner & South, 2011; Messner & DiStaso, 2013). From bar wager to academic papers (Thompson & Hanley, 2018), its authority as the unofficial arbiter of social facts is

[4]No noted percentage for usage of Wikipedia as a news source.

[5]See the following for an overview of the issues of recentism and news coverage: https://en.wikipedia.org/wiki/Wikipedia:Recentism, en.wikipedia.org/wiki/Wikipedia:Wikipedia_is_not_a_newspaper.

undeniable. The collective telling of events from Wikipedia, and the aggregate user behaviour in browsing it, is emblematic of the kind of audience-centric extra-media data required for studying news media.

2.4.2 News to History

Research interest on Wikipedia extends well beyond its coverage of news events. Rosenzweig (2006) explores the prospect of Wikipedia as "open-source history" and how Wikipedia does not conform to traditional historical recording practices. Rosenzweig notes the "possessive individualism" of historical scholarship, at odds with the practically anonymous volunteer editors of Wikipedia, as well as how in contrast to the gatekeeping practices of academia, there is a "denigration of expertise"—practically anyone can edit any article[6]. The author even observes the relatively disproportionate focus on current events and the opportunities for review and reframing—"Wikipedia offers a first draft of history, but unlike journalism's draft, that history is subject to continuous revision". Further scholarship characterises Wikipedia not just as a historical record, but as a site of collective memory—the shared knowledge of a social group, subject to constant renegotiation (Halbwachs, 1950), and distinct from the gatekeeping practices of the historical record. Pentzold, whose work "Fixing the floating gap: The online encyclopaedia Wikipedia as a global memory place" (2009) makes the case for Wikipedia "is a global memory place where locally disconnected participants can express and debate divergent points of view and that this leads to the formation and ratification of shared knowledge that constitutes collective memory" (p. 263). Follow-up studies build on this work with analysis of the construction, negotiation, as well of remembrance of events such as the Vietnam War(Luyt, 2015), the 2011 Egyptian Revolution (Ferron & Massa, 2011), the Black Lives Matter movement (Twyman, Keegan, & Shaw, 2017), disasters (Kanhabua, Nguyen, & Niederée, 2014), and plane crashes (García-Gavilanes, Mollgaard, Tsvetkova, & Yasseri, 2017). Miz, Benzi, Ricaud, and Vandergheynst (2017) use the lens of collective memory to interpret communities obtained

[6] Especially true at the time of writing, however limited rules are now in place for some contentious or important articles.

from a graph-based dynamical pattern extraction model, based on correlated patterns of attention and remembrance. Their mission of detecting the emergent topics on Wikipedia (based on both network structure and page view patterns), is key to understanding what is covered and remembered on Wikipedia, even if it does not explicitly connect to specific events.

2.4.3 Metrics Inside and Outside the Newsroom

In focussing on news events in particular, one must consider how they translate to Wikipedia both in terms of patterns of editing and regular user access. Regular news editors decide what is newsworthy and correspondingly Wikipedians decide what is Wikipedia-worthy. Literature on newsworthiness notes news values such as timeliness, impact, geographical proximity, conflict, and organisational hierarchy as important factors in news coverage and levels of interest, with journalists and editors in conversation acting as gatekeepers, shaping the presentation of events, their context, and newsworthiness (de Semir, 1996; Clayman & Reisner, 1998; Lester, 1980). Increasingly however, online news sources are utilising the wealth of data on readers' news consumption patterns to generate the most impactful headlines, reader engagement, and business revenue (Cherubini & Nielsen, 2016). Use of these editorial analytics is thrusting the behaviour of regular readers into the news gatekeeping process (Vu, 2014; Tandoc, 2014). Regular users at the centre of the gatekeeping process is more akin to the recording and access of news on Wikipedia. A greater understanding of the online behaviours of people in reaction to news events, particularly outside of the traditional realms of news publication websites, will help to greater contextualise the impact of bringing editorial analytics into the newsroom.

2.4.4 Connections to Content Analysis

Studying the news through Wikipedia article data can be considered a form of content analysis, a common technique in communication and journalism studies. Of particular interest is the task of automatic content analysis, whereby topics, trends, agenda, etc. are analysed across large corpora of news stories where manual coding is not feasible. Grimmer and Stewart (2013) and Boumans and Trilling (2016) both identify 3 main forms of this. Firstly, rudimentary dictionary

based methods which involve classification based on prevalence of manually supplied keywords (e.g. Van Dalen, de Vreese, & Albæk, 2017). Secondly, supervised machine learning approaches whereby news stories are partitioned into certain classes according to some predefined, labelled categories (e.g. Scharkow, 2013). Finally, unsupervised machine learning approaches seek to identify underlying structure and classify stories accordingly (e.g. Guo, Vargo, Pan, Ding, & Ishwar, 2016). Nicholls and Bright (2019) offer a novel unsupervised approach in which individual news stories are clustered into 'news story chains' according to textual similarity (particularly with novel words), grouping individual articles and their follow-ups into single entities. This method of grouping of news stories is an innovative feat for an often overlooked task, yet is clearly of substantial importance since how news stories relate to each other and wider context is critical to understanding their reception.

2.4.5 Identifying Events on Wikipedia

Returning the focus to Wikipedia, several papers cover methods for the detection and summarisation of news events using activity on the platform. In "Temporal summarization of event-related updates in Wikipedia", Georgescu et al. (2013) identify events by bursts in edit activity. They categorise and summarise events according to the nature of the edit, clustering updated sentences temporally, semantically, and by position on page. Summarising sentences are selected by taking the top ranked sentences from the top ranked clusters according to the number of updates. "WikiTopics: What is popular on Wikipedia and why?" (Ahn, Van Durme, & Callison-Burch, 2011) focusses on the use of bursts in page views to find topics popular on Wikipedia. They then compare different methods for clustering the articles by both topic modelling and article network structure as well as methods for selecting a sentence to summarise the article, as tested against human judgement. Attempts to test event selection against the curated current events portal of Wikipedia prove unsuccessful, indicating that what are deemed as significant news events are not necessarily characterised by bursts of attention towards one article.

Efforts have also been made to enhance news event detection by combining Wikipedia and

social media analysis. In "Bieber no more: First Story Detection using Twitter and Wikipedia", M. Osborne, Petrovic, McCreadie, Macdonald, and Ounis (2012) used Twitter and Wikipedia in conjunction for first story detection. After clustering tweets close in time according to semantic similarity, they compare these event clusters against Wikipedia page view data, keeping events where related articles on Wikipedia receive a spike in page views greater than 3.5 standard deviations over a 48-hour window. This in effect filters out noise from Twitter to ensure the detected tweet clusters correspond with real world events. Similarly, in "MJ no more: Using Concurrent Wikipedia Edit Spikes with Social Network Plausibility Checks for Breaking News Detection", Steiner, Van Hooland, and Summers (2013) detect concurrent spikes across different language Wikipedias to produce an event shortlist, then manually confirm these by performing a human search on social media.

2.4.6 Characterising Events

Limited work, however, has been done in terms of defining and characterising different news events on Wikipedia. In "Hot Off the Wiki: Structures and Dynamics of Wikipedia's Coverage of Breaking News Events", Keegan, Gergle, and Contractor (2013) consider the evolution of the article and editor network for pages that cover breaking, non-breaking, and historical events, focussing on whether collaboration networks exhibit features of organisational regeneration, having similar collaboration dynamics through time. They find that both breaking and non-breaking news events exhibit similarities in their density, clustering, and distribution of editor and article connectivity in the long run. Typical breaking news events progress via an initial decentralisation of editor activity, followed by a regression to the activity found for non-breaking and historical articles. This work demonstrates the use of these network features, beyond simple page view spikes, in measuring and categorising events, as well observed differences in the measures between different broad event types. In a similar paper "Hot off the Wiki: dynamics, practices, and structures in Wikipedia's coverage of the Tōhoku catastrophes", Keegan, Gergle, and Contractor (2011) use similar metrics to cover one event in depth. The edit and attention dynamics for the network of pages associated with the event were analysed, including the intense

interest focussed on the topic followed by a decline and dispersion of attention among newly created articles.

These works demonstrate the variance in measured online effects between different broad event classes. News events are rapidly covered on Wikipedia, yet events are not limited to transient attention based effects. They can have a lasting impact on article content and network structure, driving information production and revision by editors, incrementally contributing to the collective knowledge base of Wikipedia. Clearer connection to news value theory is required, as well as more general approaches to identifying how events are represented and responded to.

2.5 Dynamics of Collective Attention Online

2.5.1 Classes of Attention

Studying user attention and navigation on Wikipedia and indeed the web at large informs us of what information people care about and how they go about accessing it. Here I focus on the computational social science notion of collective attention, characterising aggregate population-level measures of attention, rather than the issue of allocation of attention at the individual level of cognitive psychology (Anderson, 2005). Work in this area is often related to the ideas of attention economics, where attention is treated as a scarce commodity, which also has applications in advertising and controlling the spread of information, such as with spam emails. Using Wikipedia as a setting for research means that attention towards any concept with some record on Wikipedia can be studied and directly compared against other topics.

The study of attention should not be limited to certain types of event, nor are the dynamics of one universal class. Crane and Sornette (2008) identify 3 classes of collective attention dynamics according to the dynamics of a burst and (power-law) relaxation of views towards YouTube videos, attributing particular features towards endogenous and exogenous effects. Identifying characteristic peak profiles is also the focus of J. Yang and Leskovec (2011), who in developing the K-Spectral Centroid (KSC) clustering algorithm identify 6 distinct peak profiles which are associated with unique patterns of coverage by different types of accounts (blogs, newspapers,

TV, etc.). Similar efforts including Wang, Song, and Barabási (2013); Kwon, Cha, Jung, Chen, and Wang (2013); Matsubara, Sakurai, Prakash, Li, and Faloutsos (2012); Kobayashi, Gildersleve, Uno, and Lambiotte (2021) typically consider attention towards a subject and the resulting peak dynamics as the result of sharing content and spreading mechanisms, with a particular focus on viral trends and a tendency towards Susceptible, Infected Recovered (SIR) style models. An informative coarse-grained time series analysis of Twitter hashtag data is undertaken by Lehmann, Gonçalves, Ramasco, and Cattuto (2012). The authors are able to classify attention patterns towards certain topics based on the fraction of activity that occurs on the day, before, and after a peak, finding 4 classes; anticipation dominated hashtags, fallout dominated hashtags, peak dominated hashtags, and symmetric activity hashtags. Various classes of collective online response evidently emerge due to both exogenous and endogenous influences. It is important to clarify what findings are the result of the nature of the phenomenon observed and/or the platform they are being studied on, as well how they might apply to attention patterns more widely.

2.5.2 Competition for Attention

Other work more generally models the dynamics of attention online and how it is distributed, where factors such as supply, demand, and competition affect what subjects attract interest. In "The production of information in the attention economy", Ciampaglia, Flammini, and Menczer (2015) studied the dynamics of Wikipedia article page views, modelling an attention economy in terms of the supply and demand for information on Wikipedia. They relate bursts in traffic towards pages related to a topic prior to its creation as an article as indicative of a model where demand drives the supply of information. Conversely, traffic bursts following the creation of articles indicate a model where demand follows supply. Traffic patterns for newly created articles are observed to be different from regular pages, and whilst information demand both preceding and following supply is observed, there is a significant shift towards demand preceding supply compared to a baseline of traffic for typical Wikipedia pages, indicating the considerable number of regular users who turn to Wikipedia as an information source in the wake of

events. The issue of competition for attention is central to the work of Lorenz-Spreen, Mønsted, Hövel, and Lehmann (2019). The authors use a Lotka-Volterra ('predator-prey') based model to explain shortening attention spans—as measured by the relative weights contained in peaks of activity—through increased competition across the domains of Twitter hashtags, Google trends, Reddit comment counts, Weekly box-office sales, and yearly n-gram occurrences in the Google books corpus. Notably, there is a negligible 'shortening attention span' effect in monthly scientific citations and daily Wikipedia page views. The authors attribute this to the citation and Wikipedia view dynamics being based on knowledge communication rather than entertainment consumption. Whilst (fortunately for this thesis) the effects on Wikipedia are negligible, this work does still question of how attention dynamics have changed more broadly with technology.

2.5.3 Attention and Technological Affordances

These issues are approached in a pair of papers; "Traffic in social media I: Paths through information networks" (Ratkiewicz, Flammini, & Menczer, 2010) and "Traffic in social media II: Modeling bursty popularity" (Ratkiewicz, Menczer, Fortunato, Flammini, & Vespignani, 2010). In the first of these, the authors use page view and privately collected navigational data to study how users navigate Wikipedia as compared to other websites, finding that it acts as a traffic sink and that usage patterns suggest users primarily use it for browsing and as an encyclopaedia, rather than for search or as a directory, as might be expected. They also investigate the correlation between the bursty page view activity of Wikipedia pages against Google trends search data, indicating that Wikipedia articles receive spikes in attention according to both internal Wikipedia dynamics and predominantly due to external world events. Correlations are also observed between the time series of hits for neighbouring pages, finding that this is often caused by direct traffic between them. Finally, an attempt is made at predicting article categories by their neighbours' categories according to cosine similarity, page view time series correlation and inter page traffic, though this is not so successful, with a mean average precision of around 25% at best. In the second paper, Ratkiewicz et al. measure and model the bursty popularity of pages in terms of page in-degree and page views for both Wikipedia and the Chilean web. They

observe power law relationships for both burst magnitude and time interval between bursts.

Ratkiewicz et al.'s findings contrast with what has previously been observed by Wu and Huberman (2008) with news driven events on Digg. The authors successfully model content popularity using 3 content ordering strategies, including a rank based rich get richer model together with a random ranking shift, designed to mimic when a previously unpopular topic suddenly receives a lot of attention, e.g. due to appearing in the news. Different content ordering strategies are preferable at different modelled "novelty decay rates", with a strategy prioritising existing popularity maximising attention for a slow novelty decay rate and after passing a critical value there is a phase transition towards a novelty prioritising strategy working best for fast novelty decay rates. The interaction between the social dynamics of individuals' information seeking behaviour and the technological affordances of the platform(s) of access is thus an important issue when considering different facets of online attention. These papers separately address navigation through networks and the modelling of individual page attention bursts. Synthesising these ideas as part of this project, analysing bursts of attention, as well as other patterns, through networks of pages on Wikipedia would help to complete the picture.

2.5.4 Navigation

How accessible Wikipedia's knowledge network is determines both the completeness and utility of the information readily available to users. "Evaluating and Improving Navigability of Wikipedia: A Comparative Study of Eight Language Editions" by Lamprecht, Dimitrov, Helic, and Strohmaier (2016) examines the navigability of several Wikipedias when restricted to certain sections of the articles. This is motivated by the fact that the majority of user clicks to other pages come through a relatively small number of links, focussed on sections such as the lead paragraph and infobox. With a 'bow tie' model of the network of articles on Wikipedia, they observe that the strongly connected component, pages mutually reachable, is severely reduced in size to as little as 16-37% of pages when only considering limited views (e.g. the introduction) of articles. For most users browsing, the vast wealth of information on Wikipedia is not evidently available. The authors go on to develop a link recommendation system based on improving the

navigability of the Wikipedia network. Related research (Gildersleve & Yasseri, 2018; Dimitrov, Lemmerich, Flöck, & Strohmaier, 2018) identifies the unique roles that individual articles, article types, and topics play in shaping navigation through the site. Finer grain user tracking in "Why We Read Wikipedia" (Singer et al., 2017) provides a comprehensive overview of navigation on Wikipedia to create a taxonomy of users, their behaviours and their motivations by matching survey responses with data including user clickstreams. A wide range of navigational patterns are observed, including fast-paced random exploration, current events driven navigation, and long sessions of work and research. Whilst research into navigation of the website itself is clearly applicable to the users who do browse the network of content, at the full population level navigation is a second-order effect. Around 30% of traffic to Wikipedia pages is from other Wikipedia articles and 70% from external sources, primarily search engines (Lamprecht et al., 2016).

2.5.5 Attention Towards Events

A foundational study in understanding the evolution of news issues is Watt, Mazza, and Snyder's "Agenda-Setting Effects of Television News Coverage and the Effects Decay Curve" (1993). The authors relate the coverage of particular topics on television news to audience polling data on news issue salience over 5 years with an exponential memory decay process. Characteristic decay timescales for the issues studied varied from 10-200 days, meaning an *accumulation* of coverage from regular events was effectively measured to be agenda-setting to the news audience. However, news media has moved on very quickly in the digital age, with more rapidly accessible, and arguably forgettable, coverage possibly having very different audience effects. Bright and Nicholls (2014), for example, find that individual news story life cycles on front pages online only last for a matter of hours.

Online data has now been well utilised, both as a predictor and as a measurement of collective response for particular types of event. Lerman and Hogg (2010) are able to model popularity of stories on social news site Digg, and predict their total popularity from early popularity levels, taking into account direct influence of social voting patterns on the site. On Wikipedia, "Dynamics and Biases of Online Attention: The Case of Aircraft Crashes" García-Gavilanes,

Tsvetkova, and Yasseri (2016) study the variation in attention given towards aircraft crashes as influenced by different parameters such as number of deaths, airline region, and event locale and date. Interestingly, the eventual decay of the spike in attention following a crash is independent of the number of deaths or initial impact, typically taking 3-10 days. Additionally, it is observed that there are two attention regimes, a low impact regime where the maximum lasting attention is independent of the number of deaths, and a high impact regime, where parameters such as the airline region and the impact of the event significantly influence attention. This is an important study for modelling collective attention towards a particular class of event. Kummer (2014) studies the spillover of attention to neighbouring linked articles in the wake of an article being featured on the Wikipedia front page (this is independent of any current events). Shocks of attention in the article network such as this drive up to a 70% increase in views as well as short edits, however, "deep edits" are not sensitive to these shocks. The results from this study act as a baseline for analysis of news events, exogenous drivers of attention towards articles across a network.

There is further work that goes onto use this online data on attention, such as Wikipedia page view statistics, news media mentions, and Google search trends, in predicting 'real world' information. For example, Yasseri and Bright (2014, 2016) approach the issue of election prediction by considering information seeking behaviour online for individual politicians and parties, particularly for swing voters and new parties, across several elections. Results from Google search and Wikipedia are not always correlated, implying different uses and predictive utility for the two data sources. Additionally, whilst Wikipedia data is of relatively minor importance in predicting absolute vote share, it is an important feature in predicting vote swing. This indicates that this information seeking is more common for those considering changing vote, though this relationship is exaggerated for newer parties. Kobayashi et al. (2021) find associations in the page view time series between more general classes of planned events (movie releases, holidays, elections, sporting competitions, and football matches), with clear circadian identifiers to events more popular in different regions, identifiable patterns based on event outcome (match

win/draw/loss), and predictable collective attention dynamics based on a simple peak model. More of these kind of tasks, relating Wikipedia data to 'real-world' outcomes, are undertaken in research on movie box office performance (Mestyán, Yasseri, & Kertész, 2013), influenza outbreaks (McIver & Brownstein, 2014), as well as stock and cryptocurrency trends (Moat et al., 2013; ElBahrawy, Alessandretti, & Baronchelli, 2019).

Attention is clearly a defining feature of online activity in response to news events; further work is needed to better understand the competition for attention for topics both within events and between different events, the resulting classes of dynamics that may emerge, and crucially how this relates to external news media.

2.6 Conclusion and Research Questions

Wikipedia is well established as a cornerstone of the modern Internet. Its encyclopaedic knowledge base supplements the developing record of current events that proves so popular with users. Prior literature has covered how editors have recorded current events, as well as case studies on the wider reception of particular kinds of event, yet significant gaps in the literature remain. Given the importance of news media in communicating information on current events it is crucial to understand the coverage and access of this information from the perspective of news value theory. In fact Wikipedia's unique status as independent extra-media data grants us the ability to draw conclusions on wider matters in news media. The high concentration of activity and interest around the time of events also motivates the more careful study of peaks in collective attention. This research is often limited to studying peaks generated by spreading processes, but the communication mechanics of news and motivations of users on Wikipedia compared to social media are not necessarily the same. Wikipedia's status as an independent, intentional, low cost information resource again acts to alleviate endogeneity concerns, granting an 'unfiltered' perspective of collective attention towards news events.

Ultimately, work is needed to understand what kinds of event are covered and how they are accessed by Wikipedia's audience. In addition, we must relate Wikipedia coverage and access

patterns to that of news media, in turn contributing to news value theory. Finally, we must give added focus to peaks of collective attention, where much of the online activity around a news event is concentrated. I articulate and approach these issues through three central research questions and their constituent sub-questions:

- **RQ1:** How are current events represented in the knowledge structures and access patterns of Wikipedia and its users?
  - **RQ1a:** How can we identify and sample the groups of Wikipedia articles associated with a given news event?
  - **RQ1b:** How are these groups of articles from different events related and do they form coherent topics?
- **RQ2:** How are traditional conceptions of news values and newsworthiness of events reflected in extra-media data?
  - **RQ2a:** What types of events, according to news values, are recorded on Wikipedia?
  - **RQ2b:** How does the complementarity hypothesis apply to extra-media data?
  - **RQ2c:** How are news values associated with newsworthiness?
- **RQ3:** How can we model and predict peaks of collective attention towards news events?
  - **RQ3a:** What characteristic shapes of peaks of collective attention arise in response to news events?
  - **RQ3b:** How do different dynamics affect the rise and fall of peaks of collective attention?
  - **RQ3c:** How well can peak models predict collective attention in the aftermath of events?

These challenges are addressed in chapters 4, 5, and 6.

Chapter 3

Data

3.1 Wikipedia data

The three primary classes of Wikipedia data used in this thesis are information on news events that occur from the Wikipedia Current Events Portal (*Portal:Current events - Wikipedia*, 2021), data for the article network of Wikipedia, i.e., the article names and what hyperlinks exist between them, and time series data for the daily and hourly page views to each article. The Wikimedia Foundation generously make this data readily available in several forms (through API[1] access, or various database dumps). Due to the scale of the data needed to be gathered, I turned to the data dumps and prioritise local computation over time spent making API calls.

3.1.1 Current Events Portal

The Wikipedia current events portal is a daily archive of news events as recorded by Wikipedia editors. Events are sorted in 10 categories (Armed conflicts and attacks, Law and crime, Arts and culture, Politics and elections, Business and economy, Science and technology, International relations, Sports, Health and medicine, Disasters and accidents) and each event is written as a summary sentence with links to relevant articles. In addition to the daily list of events, there are monthly sections on important recent, ongoing, and upcoming events, sport, elections, deaths, trials, and conflicts. A partial snapshot is displayed in Figure 3.1. This portal is a ready-made

[1] "Application Programming Interface", a technology that allows remote control of a service or access to database information through the construction of specific queries in code, e.g., request page views for a Wikipedia article over a certain time period.

wealth of data for event networks across many categories that may exhibit varying dynamics. The criteria for an event's appearance on the page is simply that the community believe it is a significant story, subject to guiding principles such as the length and depth of existing news coverage, though ultimately, decisions are based on consensus of editors on the individual merits of events (*How the Current events page works*, 2021; *In the news*, 2021). I scrape the page to sample a full year of events from 1st December 2017 to 30th November 2018 from the current events portal. Initial data gathered for each includes date of the event, category of the event, full text description, and the linked Wikipedia articles (henceforth referred to as "core articles") in each description. For example, in Figure 3.1, the final event on 01/04/2017 in the 'Disasters and Accidents' category is described as "Authorities cannot contact the *South Korean* cargo freighter *Stellar Daisy*. It is believed that the ship sunk off the coast of *Uruguay*". I extract the linked pages *South Korea* (displayed text does not have to match article title), *Stellar Daisy*, and *Uruguay* as the core articles for this event.

3.1.2 Clickstream Networks

For the Wikipedia network data, I use dumps from the Wikipedia Clickstream (*Analytics Datasets: Clickstream*, 2021). The Wikipedia Clickstream contains monthly aggregated counts for the number of times links are accessed on Wikipedia, and crucially where from, in (referrer, resource)—equivalently (source, target)—pairs. This is supplied in multiple languages though I only study the English language Wikipedia. Visits from popular sources outside Wikipedia are also recorded, although I only include hyperlinks between Wikipedia articles, excluding links from external to Wikipedia and from Wikipedia's Main Page. In addition, in the raw data only links with > 10 clicks over the course of each month are supplied. Essentially, this represents an edgelist that forms a directed, weighted network of monthly navigation between Wikipedia articles. The clickstream data is used since firstly it offers a fast, reliable snapshot of the past article network structure of Wikipedia (compared to the time taken wrangling HTML from the complete Wikipedia dumps, or from individual article revisions through API calls), and secondly the weighting allows for a cutoff for spurious links. The edge weights do offer interesting

Figure 3.1: A snapshot of the Wikipedia current events portal.

navigational information but these weights are not used in the later analysis since the monthly resolution is not fine enough when studying news events whose typical timescales are on the order of hours or days. The period of study for news events is December 2017–November 2018, so clickstream data for the English Wikipedia was downloaded for November 2017–December 2018, allowing a one month buffer for studying news events at the start and end of the time period.

3.1.3 Page View Time Series

Page view data for all articles is also downloaded from the Wikimedia data dumps (Wikimedia, 2021). The page view data is grouped into monthly datasets with hourly granularity for each

Wikimedia project. I focussed on the page views towards articles in the English Wikipedia, which are reported under the prefixes for desktop Wikipedia (en.z), the mobile Wikipedia (en.m), and Wikipedia Zero (en.zero), a now discontinued project where Wikipedia access was available for free in developing countries through zero-rating (included for completeness though in practice en.zero accounts for relatively few views). The raw data is stored in a highly compressed format and relatively slow to access on demand. I identify the networks of articles linked to the entries in the current events portal for which page view time series are required (more details in Chapter 4), and process the raw compressed time series data to more accessible HDF5 format. This data was downloaded for the period November 2017–December 2018.

3.1.4 Redirects

A somewhat minor, yet often overlooked, facet of research on Wikipedia is the issue of page redirects. A redirect is a page which automatically sends visitors to another page with the 'correct' title, for example searching for 'USA' or clicking a wikilink titled 'USA' automatically redirects the user to the page 'United States'[2]. Redirects are very important since users and editors constantly search for articles, click on links, and edit links with alternative/abbreviated/previous names. In fact, it has been estimated up to 55% of the articles in the main namespace of Wikipedia are redirect articles (Hill & Shaw, 2014). In order to ensure information for identical articles in the data is not being duplicated, redirects need to be resolved. Redirects are already resolved in the clickstream dataset (though pages may change in name over time e.g. 'Meghan Markle' to 'Meghan, Duchess of Sussex') but they are unresolved in the page view dataset, i.e. hits on redirect pages are recorded separately from hits for the true page name[3]. For all articles in the dataset (which includes page views towards the 'correct' article names as well as redirect pages), Wikimedia API calls for redirects (MediaWiki, 2021) were used to create a mapping using 1) what 'correct' title they redirect to (if necessary) 2) all other names that redirect to the article. Page views for individual articles were then calculated by summing those for the groups

[2]More information available here: en.wikipedia.org/wiki/Wikipedia:Redirect
[3]This is further detailed at wikitech.wikimedia.org/wiki/Analytics/Data_Lake/Traffic/Pageviews/Redirects.

of their redirects. When mapping the page view data to the articles in the clickstream data this guarantees correct correspondence. The different forms of data are summarised in Table 3.1.

	Timespan	Resolution	Source	Redirects	Notes
Current Events Portal	12/17-11/18	Daily	Scraped	Unresolved	Categorised by news topic
Clickstream Networks	11/17-12/18	Monthly	Data Dumps	Resolved	10 click threshold
Page View Time Series	11/17-12/18	Hourly	Data Dumps	Unresolved	Separated by wiki (z, m, zero)

Table 3.1: A summary of the Wikipedia data sources used for this thesis.

3.2 News article data

I link the records of current events on Wikipedia to a database of news published in a wide selection of prominent worldwide outlets. The LexisNexis 'Major World Publications' database (*Nexis UK*, 2018) compiles news from 153 outlets (listed in Appendix A). Articles are supplied in docx format, with supporting information such as headline, byline, and media descriptions also stored. An example of the interface and output is shown in Figures 3.2 and 3.3.

Searches for relevant news articles based on the current events portal entries are undertaken. Queries to the LexisNexis database are constructed by taking the core Wikipedia article names from the event description as a comma separated search term, limiting results to ± 1 day from the listed event date, and scraping the first 500 'most relevant' items. Having retrieved up to 500 news articles per event query I run a second level of news article to Wikipedia event record matching, since the relatively loose nature of the search term means many false positives are included in the results. I use Universal Sentence Encoder (Cer et al., 2018) to generate a text similarity score between the body of each news article in each Wikipedia events record's search results and the text description of the event from the Wikipedia current events portal. A threshold similarity is selected by manual labelling of the (typically 500) returned news articles for 50 randomly selected events as either relevant or not relevant to the event. Optimising for mean absolute error (MAE) gives the threshold of a minimum similarity of 0.46 (corresponding MAE $= \pm 13.2$ articles), which is then applied across the full dataset of events. In total, there are $114,721$ news articles selected, written by $\approx 10,000$ unique authors.

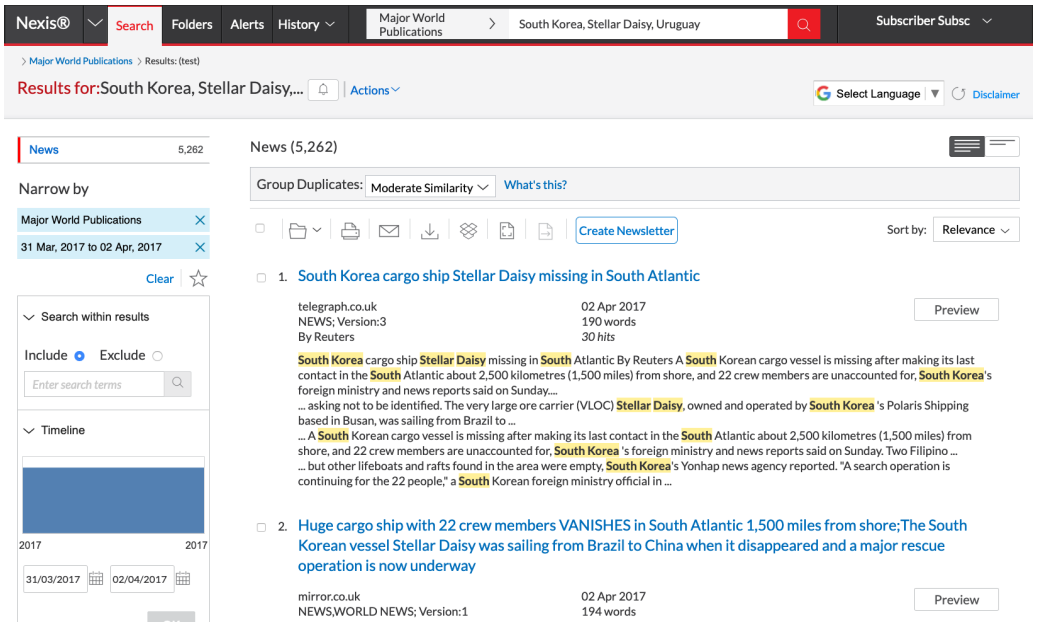

Figure 3.2: The NexisUK interface with an example query based the event "Authorities cannot contact the *South Korean* cargo freighter *Stellar Daisy*. It is believed that the ship sunk off the coast of *Uruguay*" from 01/04/2017 in Figure 3.1.

South Korea cargo ship Stellar Daisy missing in South Atlantic

telegraph.co.uk

April 2, 2017 Sunday 8:31 AM GMT

Section: NEWS; Version:3

Length: 190 words

Byline: By Reuters

Body

A South Korean cargo vessel is missing after making its last contact in the South Atlantic about 2,500 kilometres (1,500 miles) from shore, and 22 crew members are unaccounted for, South Korea's foreign ministry and news reports said on Sunday.

Two Filipino crew members were rescued floating in a life raft on Saturday, but other lifeboats and rafts found in the area were empty, South Korea's Yonhap news agency reported.

"A search operation is continuing for the 22 people," a South Korean foreign ministry official in Seoul said by telephone, adding that eight of the missing are South Korean nationals and 14 are Filipinos.

South Korea has requested Brazil and Uruguay to aid in the search and rescue, the official said, asking not to be identified.

The very large ore carrier (VLOC) Stellar Daisy, owned and operated by South Korea's Polaris Shipping based in Busan, was sailing from Brazil to China carrying iron ore when it sent a distress signal to the ship operator on Friday, Yonhap said.

A message received on Friday by Polaris from a crew member said the ship was taking in water on the port side and was listing rapidly, Yonhap said.

Load-Date: April 2, 2017

End of Document

Figure 3.3: The NexisUK output with one example story.

Chapter 4

Topics of Attention on Wikipedia

4.1 Introduction

In order to study the impact of news on Wikipedia, we must first identify what news events are represented and how they are manifested. Unlike a news website, where there may be a single news article, or occasionally a small series of articles, the information on Wikipedia is stored in a network of Wikipedia articles with no inherent order or bound. Information seeking behaviour relating to news on the website is thus exploratory. Individuals browse groups of articles relating to events in the news, but what are the various groups that are browsed? Are they consistent between events in forming news topics? And how well do access patterns align with the structure of the information in the encyclopaedia? Indeed, only when the representations of reactions on Wikipedia's article network to news events have been formalised can we study any further network wiki news phenomena.

Undertaking this task using Wikipedia provides several perks one may not typically encounter when studying events using social or news media. The assembled knowledge network establishes relations between constituent actors, settings, and further subjects of events. The existence of the various regions of this knowledge network are not dependent on the occurrence of an event either, enabling long term study of activity both before and after an event occurs. As already argued, Wikipedia itself also acts as a relatively independent barometer for collective response to current events, compared to the findings one would find when directly considering the output

of news media or algorithmically served content from social media.

For the most part, prior literature exploring the effects of current events on Wikipedia either selects individual articles from a list of related events (García-Gavilanes et al., 2016, 2017; Aragon et al., 2012) or identifies articles through event that receive bursts of attention in the form of page views or edits (Georgescu et al., 2013; Ahn et al., 2011; M. Osborne et al., 2012; Steiner et al., 2013). Whilst these methods may have been adequate in addressing the specific research questions of their respective papers, they are not sufficient in selecting a range of events and their relation to established knowledge for a full comparative study of their evolution.

In this chapter I present a topic detection model for Wikipedia. I use both the Wikipedia network structure and correlations between time series for page views, combined with a database of events from Wikipedia's current events portal to identify groups of articles that are both well connected and exhibit similar patterns of page views around individual events. These groups of articles are then connected through time to identify the recurrent topics attracting attention on Wikipedia. To thoroughly study this phenomenon, one requires the extraction of objects of study that are rooted in a wide sample of news events, incorporate the network structure of articles (in effect a long-term memory based process) as well as short-term attention dynamics, yet are not selected for solely based on such signals. I detail in this chapter how I achieve this with a temporal network community detection approach towards identifying the prominent themes linking individual news events to chains of stories and wider news topics. The resulting 'Event Reactions' and 'Topics of Attention', which encompass news topics as well as attention towards background topics on Wikipedia (and can be resolved as such), act as the primary objects of study for the later chapters of this thesis. The approach is built around **RQ1**.

- **RQ1:** How are current events represented in the knowledge structures and access patterns of Wikipedia and its users?
  - RQ1a: How can we identify and sample the groups of Wikipedia articles associated with a given news event?

- RQ1b: How are these groups of articles from different events related and do they form coherent topics?

4.2 Sampling Requirements and Prior Approaches

No previous work fulfils the aforementioned requirements of; sensitivity to short and long term effects through explicit relation to news events and usage of the knowledge network, generality across topics, and finally an independence of detection from particular attention dynamics. I elaborate on these, and explain why they are necessary for this thesis.

Explicit relation to external news events

Many works explore the various online dynamics in response to current events. Occasionally these are explicitly tied to particular news events, especially when the relevant articles, hashtags, videos, etc. are manually supplied (e.g. the approach of Keegan et al., 2011). However, in cases where larger datasets are used, it is frequently the case that properties such as page views of a large number of pages are studied independent of any explanatory description, with any detected interesting features such as peaks later being ascribed meaning (likely some external event) by the researcher(s) (e.g. the approach of Lehmann et al., 2012). Whilst an issue of causality behind whether the external event is responsible for the signal(s) observed on Wikipedia remains, there is an important distinction between starting from the point of news events and understanding their dynamics, rather than observing particular dynamics and attempting to relate to news events. Firstly, sampling-wise we may only select for particular dynamics when adopting the latter approach (detailed further below). Secondly, immediately linking to news events assists in later stages of interpreting results—even the most well-read researcher would not be able to interpret results across a wide range of topics and provide explanatory news events without supporting documentation of such events linked.

Articles vs networks

As according to the premise of this thesis, and supported by Kane (2009); Milne and Witten (2008); Ciampaglia, Shiralkar, et al. (2015), information on Wikipedia, including news events,

is partly defined according to its relations to other articles, whether that's as a 'sub-event' as part of a larger ongoing set of events, or an event that heavily links several individuals, places, objects, etc. Previous literature typically concerns itself with events that have dedicated articles (M. Osborne et al., 2012; García-Gavilanes et al., 2016). However, news events can also be documented within one already existing article or across several different articles. For example, the death of a prominent figures is contained within that person's article, and stories about one public figure making newsworthy comments about another may be separately recorded on their respective articles. Analysing the dynamics of news events in the context of their links to related topics thus necessitates a network based approach.

Generality across topics

Event selection as according to a Wikipedia list of related events not only suffers from the same issues as only selecting those events with dedicated articles, but also may be limited in the range of article dynamics observed and how this might apply to other topics. Collecting and comparing events from a common category is informative in answering how event dynamics vary according to category specific parameters, and will certainly form part of this work. However, just as important is the comparison of different event categories. By analysing different event categories, a much wider range of event dynamics can be observed.

Event detection through specific signals

Bursts of attention towards a topic seems like a natural way of selecting news events, and are frequently used in various general event detection algorithms. However, I argue that the level of attention towards a news event is a feature for analysis in itself, rather than simply a defining characteristic of what a news event is. For a start, a burst of attention towards a topic does not mean it is a significant news event, for example, pages for TV shows receive bursts of attention when episodes air and certain news events do not attract bursts of attention (whether that be due to a lack of attention or a consistently high level of attention) (Ahn et al., 2011). Selecting events according to attention thus introduces a sampling bias, studying events that do not receive bursts of attention and why this is the case may prove to be just as important an analysis.

To account for these conditions in my work, several sources of data—on news events, article network structure, and page view dynamics—must be synthesised.

4.3 Sampling News Events

As detailed in Chapter 3, the three primary classes of Wikipedia data used are information on news events that occur from the Current Events Portal (*Portal:Current events - Wikipedia*, 2021), data for the article network of Wikipedia, i.e. the article names and what hyperlinks exist between them, and time series data for the daily page views to each article. Here I detail the exact pipeline by which I generate a network of related articles and associated page view time series related to each news event (the 'Event Networks'), analyse these for communities of articles representing distinct content and dynamic based 'Event Reactions', and finally cluster these communities (the 'Event Reactions') based on overlapping constituent Wikipedia articles to identify 'Topics of Attention'. These concepts are the key levels of analysis in this chapter, and are more clearly defined as follows:

- **Event Network:** The hyperlink network of Wikipedia articles and associated page view time series related to a particular news event.
- **Event Reaction:** A community of articles within a single Event Network that are relatively strongly linked and receive correlated patterns of page views.
- **Topic of Attention:** A cluster of Event Reactions, grouped according to common constituent Wikipedia articles.

4.3.1 Processing Current Events Portal Data

Firstly, news entries from the current events portal must be related to networks of Wikipedia articles and page view data. The process to generate 'Event Networks' runs as follows:

- For each news event:
  - Scrape aforementioned data from current events portal.

  - Resolve redirects of pages linked in news descriptions—henceforth referred to as 'core' articles.
  - Use clickstream data to create network of all articles that link to, and are linked to by, as well as all links between these articles over a window of 61 days centred on the recorded event date. Edge weights are a weighted average of the monthly totals.
  - Keep all edges with estimated weight > 100, remove any isolates. This is done due to computational limitations regarding the sparsity of graphs during community detection.

- Collect all article names in the networks, with all redirects, and the period of time they are in the news and require page view data for.
- Process the page view data, keeping data for all required article names, with redirects, over the required time periods.
- Assign 61-day time series (30 days before/after event date) to each news event for each article in the respective networks.

The Event Networks encapsulate the network structure and page view dynamics in response to current events, however, they are not a wholly satisfactory description when attempting to generate summary statistics for each event. It is not the case, given the breadth of pages included in each network, that all their articles will exhibit the same signals for page views or edits, many simply being unrelated in the context of the news event. As such, simple averaging techniques for network level features will likely wash out any useful information.

4.3.2 From Event Networks to Event Reactions

One might think that the issues with the Event Networks are a result of the sampling strategy. In abstract terms, there may well be one true signal for each news event, yet it is obfuscated by the noise of less related pages picked up in the network, or concurrent news events involving the same pages, and that the solution is simply some filter or averaging process. I argue that

on the contrary it is the very nature of current events that when studying the underlying constituent concepts and their response that there will be longer term effects from historical events, structural effects from related information, as well as associations with other current events. This could lead to a variety of different responses. This may seem trivial, but it often does not explicitly emerge in research where the objects of study in focus are specific individual hashtags, news articles, YouTube videos, etc.

I have already argued that the constituent Wikipedia articles relating to individual events exhibit a variety of different dynamics tied to historical, structural, and concurrent news effects. I propose a method to separate responses across both content (structure) and attention (dynamics), to identify which groups of articles are both well connected *and* exhibit similar page view time series. Simply taking clusters according to network structure ignores short term associations and their strength, on the other hand, simply taking pages with correlated responses ignores context of the related content, and could also introduce spurious associations. This approach takes both factors into account.

The chosen two-stage temporal community detection approach disentangles the different response signals across each Event Network into communities of articles termed 'Event Reactions'. Each news event is partitioned into a handful of 'Event Reactions' across the different subjects represented. These 'Event Reactions' are then clustered with those from other news events according to common constituent articles (via a weighted Jaccard index), to detect broader topics, termed 'Topics of Attention'. A schematic of the full process is shown in Figure 4.1[1].

4.4 Extracting Event Reactions

4.4.1 Temporal Community Detection of Signals on Knowledge Networks

Together with Section 4.3, here I tackle **RQ1a:** *How can we identify and sample the groups of Wikipedia articles associated with a given news event?* For a given news event i, with associated

[1]The reader familiar with topic modelling may find the following analogy useful. Individual Event Networks represent 'documents', that are made up of Wikipedia articles, akin to 'words', and each news event represents a sample of wider news 'topics', that I term 'Topics of Attention'.

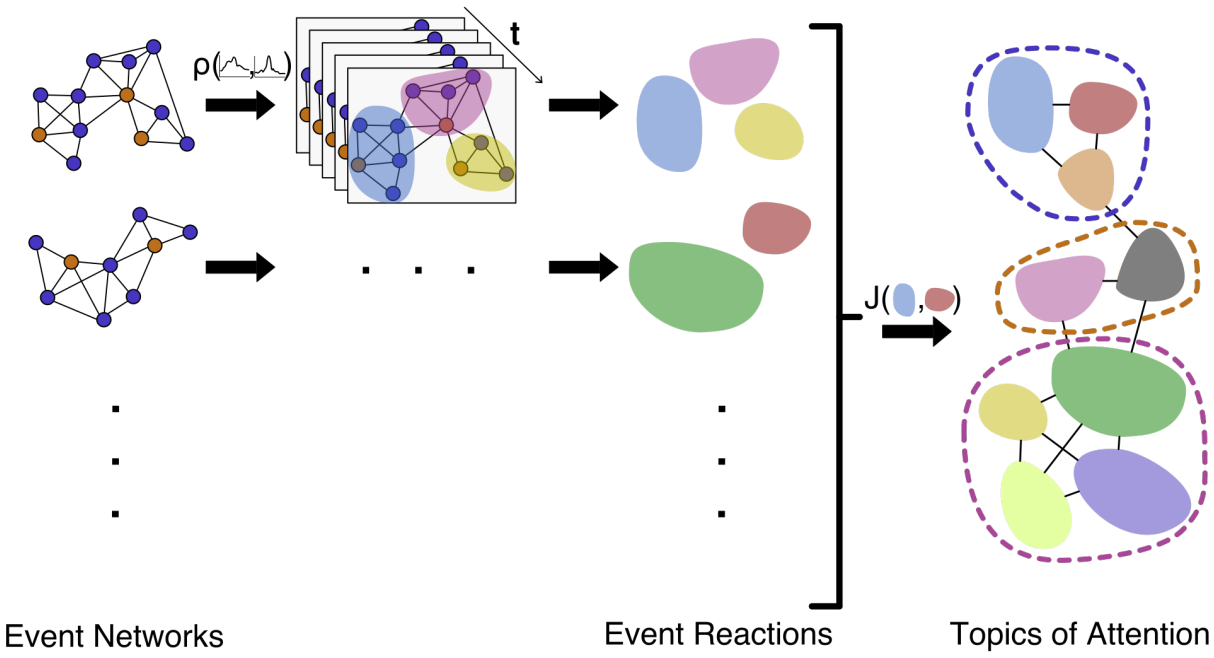

Figure 4.1: A schematic of the processing of Event Networks to Event Reactions to Topics of Attention.

network of articles $G_i(V, E)$ (nodes representing articles, edges representing hyperlinks between them), there is an associated set of time series for page views towards articles $p_n(t) \forall n \in V$ with T timesteps. Correlation matrices (for all combinations of nodes) over a rolling 7 day window, using Pearson coefficients, are calculated, yielding W, which is a $|V| \times |V| \times (T-6)$ dimensional array. In order to restrict the analysis to hyperlinks (i.e. pairs of pages) which experience substantial increased correlation in traffic, I generate two further arrays $W_{\text{edges}} = W \circ A$ and $W_{\text{non-edges}} = W \circ (1 - A)$, where A is the (unweighted) adjacency matrix of G_i[2].

W_{edges} represents an undirected, weighted temporal network $G'_i(t)(V, E(t))$, on which I perform community detection to identify groups of articles that are both well connected by hyperlinks, and exhibit correlated patterns of page views.

The Leiden algorithm (Traag, Waltman, & van Eck, 2019) is selected for temporal community

[2]I explore the effect of thresholding the edge weights in this network in Appendix B.1, but find no significant variation or improvement in performance so proceed with all edges. Note that the imposed graph structure already removes the vast majority of time series correlations between nodes.

detection. This is an extension of the popular Louvain algorithm, but addresses an issue whereby communities may be arbitrarily badly connected, it also runs faster than the Louvain algorithm. Rather than simple modularity, the Constant Potts Model (Traag, Van Dooren, & Nesterov, 2011) is used for the quality function, since it can handle both positive and negative edge weights (which can be observed depending on the edge weight threshold), the readily interpretable resolution parameter, and the independence of communities from the observed graph/subgraph. The temporal method adapts the approach of Mucha, Richardson, Macon, Porter, and Onnela (2010) with the Leiden algorithm, using layers of $G_i'(t)(V, E(t))$, with a constant interlayer coupling of $\tau = 1$ between identical nodes in neighbouring layers, in addition a resolution parameter is required. A search for this parameter with a robustness test on a 50 event sample is carried out in Appendix B.1, with the resolution being set to $r = 0.25$.

For each Event Network, the obtained partition $\mathbf{P}_i$ is comprised of a handful of communities C_{ij}. Any detected communities which contain at least one of the 'core' articles from the descriptive text of the event, and that overlap in time with the day of the event are kept as Event Reactions—R_{ij}. Each of these elements is in effect a building block of wider Topics of Attention. The discrepancy in timescales between fast-pace attention towards news events and the more slowly evolving structure of the Wikipedia article network means these topics are not necessarily reflected in solely the hyperlink structure (or aggregated navigational structure from clickstream logs), or solely through correlated short term page views. In addition, satisfactory temporal community detection on one network for one year over the ≈ 6 million English Wikipedia articles is not computationally feasible.

4.4.2 Baseline Community Detection Comparison

I compare the obtained communities from the combined network structure and page view correlations against those generated from a solely network structure approach. If the communities from the combined approach are no different from those for the network structure baseline, then this indicates attention dynamics in response to any news event have very little effect, and the community is well represented solely by the network. Any 'disturbance' by the news event is

either minimal, or closely aligns with the way information is already represented on Wikipedia. On the other hand, if the communities obtained from each approach are quite different, then the variation in page view dynamics among the articles in the network is important in producing the Event Reactions. Any 'disturbance' by the news event is of sufficient magnitude that the association between concepts related to the event is then not well represented by the relatively static structure of information on Wikipedia.

The approach for each event is based on comparing the communities already obtained from the temporal network of correlated time series to those we obtain from community detection on a single-layer, unweighted, graph representing the structure of the article network (i.e. G_i from Section 4.4.1). For each event i I run community detection on the graph G_i over the same logarithmic range of resolutions $r \in [1.23 \times 10^{-4}, 1]$ from the temporal robustness tests (Appendix B.1), yielding the partitions $\mathbf{P}'_{ir}$. Each Event Reaction from the temporal approach (R_{ij}) receives a 'Structural Similarity' score s_{ij}. This score is defined as the maximum of the similarities between R_{ij} and each community obtained from the non-temporal approach across all resolutions $C'_{ikr} \in \mathbf{P}'_{ir} \ : \ 1.23 \times 10^{-4} \leq r \leq 1$. Thus,

$$s_{ij} = max(J_w(R_{ij}, C'_{ikr}) \ \forall \ C'_{ikr} \in \mathbf{P}'_{ir} \ : \ 1.23 \times 10^{-4} \leq r \leq 1), \tag{4.1}$$

where $J_w(x, y)$ is the Jaccard similarity between communities x and y, weighted by the PageRank scores of nodes in the subgraphs x and y. This takes into account both the content of the community, and the relative importance of articles within the community.

The Structural Similarity score describes how dependent the observed community R_{ij} is on variation in short term correlated attention dynamics. If all page view time series were uniformly correlated, we would expect $s_{ij} \approx 1$. If on the other hand a subset of articles receive strongly correlated page views, uncorrelated with the page views to other articles, we would expect $s_{ij} \approx 0$. The distribution of s across all Event Reactions is shown in Figure 4.2. We observe a range of behaviours; the largest mode has relatively low structural similarity (i.e. page views are important), there is an intermediate mode around $s = 0.6$ (page views have some effect), and finally the mode around $s = 1$ (page views have no effect).

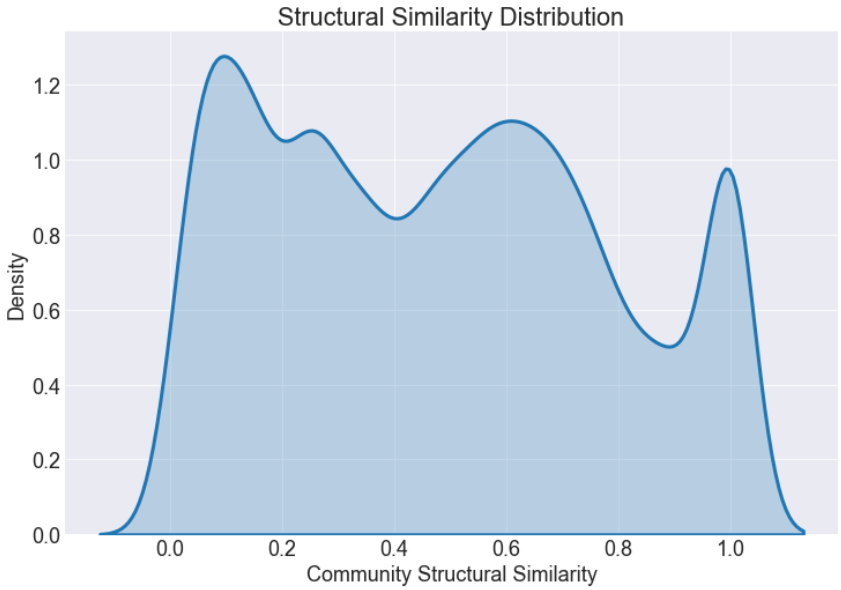

Figure 4.2: Distribution for the structural similarity scores of all Event Reactions.

4.5 Establishing News Topics

4.5.1 Higher-level Topics of Attention

Over all events, we now have a collection of Event Reactions. Many of these will be related through covering different stages of the same continuous event (e.g. different rounds of the FIFA World Cup), or through the re-emergence of events and news topics in time (e.g. updates related to the Mueller report, or new natural disasters). I now turn to **RQ1b:** *How are these groups of articles from different events related and do they form coherent topics?* I seek to identify the recurring groups of Wikipedia articles associated with news events—the topics that are represented. Event Reactions from different events that are made up of broadly the same collection of Wikipedia articles are emblematic of a wider concept receiving repeated news exposure. I look to quantify the similarity between Event Reactions and use this to find the more closely related groups that represent Topics of Attention. I construct a higher-level network $H(V, E)$ of all Event Reactions. Edge weights are set as the weighted Jaccard similarity (Ioffe, 2010) between the sets of articles of each Event Reaction, weighted by their PageRank centrality in their respective networks (Page, Brin, Motwani, & Winograd, 1999), indicating similarity in

content (and weighting more important articles to the concept more highly). This network contains all recorded instances of Event Reactions in the sample, representing their relation to one another over the course of one year. In order to identify the Topics of Attention (groups of related Event Reactions) I run a further stage of community detection over this network $H(V, E)$, using the Leiden algorithm with the Constant Potts Model as before. Whilst the nodes in the network represent snapshots of events centred on different points in time, $H(V, E)$ is not a temporal network. The resolution parameter is set at $r = 0.12$, according to the robustness test set out in Appendix B.2. This process yields a partition of communities that are the Topics of Attention which I go on to label, validate, and explore.

4.5.2 Topic Labelling and Validation

In line with literature on news values and newsworthiness (more in Chapter 5), the Topics of Attention were sorted by several features detailed in Table 4.1, with the top topics across each feature manually labelled. Several of these are based on the constructed time series, W_{ij} for each Event Reaction (R_{ij}). This is a sum of the daily page views to each article in an Event Reaction ($p_k(t)$), weighted by their PageRank centrality (w_k) to the network they form;

$$W_{ij}(t) = \sum_{k \in R_{ij}} w_k p_k(t). \tag{4.2}$$

The time series is then centred in time to the max value occurring ± 1 day from the recorded date of the event. For a Topic of Attention A_α with constituent Event Reactions R_i (i acting as an index for the Event Reactions in A_α and no longer referring to a specific event) and their associated time series W_i, the prominence, magnitude, and deviance of a topic are then accordingly

$$\text{Prominence}_\alpha = \frac{\sum_{W_i \in A_\alpha} \text{median}(W_i(-30, -29, ..., 0))}{|A_\alpha|}, \tag{4.3}$$

$$\text{Magnitude}_\alpha = \frac{\sum_{W_i \in A_\alpha} W_i(0) - \text{median}(W_i(-30, -29, ..., 0))}{|A_\alpha|}, \tag{4.4}$$

Table 4.1: A summary of the features with which I sort and examine the Topics of Attention.

Feature	Description
Number of associated Events	Topics most frequently featured in the news (often background)
Prominence	Topics which on average have the largest level of pre-existing attention, median page views (often background)
Magnitude	Topics which on average receive the largest increase in page views when in the news
Deviance	Topics which on average receive the largest increase in page views, relative to their prominence, when in the news

$$\text{Deviance}_{\alpha} = \frac{\sum_{W_i \in A_{\alpha}} \left(W_i(0) - \text{median}(W_i(-30, -29, ..., 0))\right)/\text{median}(W_i(-30, -29, ..., 0))}{|A_{\alpha}|}. \tag{4.5}$$

Two coders independently manually labelled a subset of 150 of the Topics of Attention by examining the constituent Wikipedia articles for each topic and the news events most associated with them (five events initially, with the option to see more) using the interface in Figure 4.3. This set of 'top topics' were selected by taking the top 50 topics across each feature in Table 4.1. Each coder was then presented with the combined list of labels and independently tasked with identifying where there was 'strong agreement', 'partial agreement', or 'weak/no agreement' between labels. For the top topics, 70% of labels were in unanimous strong agreement, 22% in partial agreement, and 8% in weak/no agreement. The procedure demonstrates validity of the interpretable topics. For the purposes of display in figures and tables, cases where there was not unanimous strong agreement between coders were passed to a second round where coders were invited to resubmit labels, giving greater scrutiny to the available information, and the agreement between labels evaluated again. Ultimately, this resulted in 92% of labels in strong agreement, 5% in partial agreement, and 3% in weak/no agreement. A selection of these are displayed in Table 4.2.

```
Cluster  9

Most Frequently Occurring Core Articles:
European Union         49
European Commission    12
European Parliament     5
Donald Tusk             3
European Council        2
dtype: int64

Most Frequently Occurring Articles in Networks:
European Union                   59
European Commission              22
European Parliament              22
European Council                 20
Council of the European Union    18
dtype: int64

Random Sample of Events:

20180212
Cyprus–Turkey maritime zones dispute
The European Union urges calm and restraint after Turkish Navy warships obstruct a Cypriot offshore
drilling vessel in the Eastern Mediterranean, which was approaching an area to explore for natural
gas. (Reuters)

20181121
Kosovo puts a 100% trade tariff on all goods imported from Serbia and Bosnia and Herzegovina. The
European Union says this is a "clear violation" of the Central European Free Trade Agreement. (BBC)

20180824
European migrant crisis
Representatives from 12 European Union countries do not reach an agreement on Italy's request to take
in the 150 migrants still remaining on the Diciotti-class vessel in Catania, Sicily, after their
rescue by the Italian Coast Guard nine days ago. (Deutsche Welle) (ANSA)

20180904
The European Ombudsman, Emily O'Reilly, mentions four counts of maladministration by the European
Commission in the fast-track nomination of Martin Selmayr as its Secretary-General in February. (BBC)

20181027
ALDE European Parliament group expels Catalan nationalist party PDECat amid corruption scandals that
affect the CDC predecessor party. (El Periódico)


Would you like to see more event descriptions? (Y/N):
n

Describe this cluster:
```

Figure 4.3: The interface for labelling Topics of Attention, showing the most frequently occurring core articles, regular articles, and a sample of related events.

Table 4.2: Top topics by certain measures (min 10 events). Colour indicates quartile of structural similarity score, from red=bottom quartile to green=top quartile.

	# Events	Prominence	Magnitude	Deviance
1	Countries	FIFA World Cup	FIFA World Cup	Iranian Protests
2	US States	FIFA World Cup 2	FIFA World Cup 2	Caracas drone attack
3	Middle East	Countries 2	Kavanaugh Supreme Court Nomination	Peruvian politics
4	Global Cities	Countries	US Senate	California Wildfires
5	US Presidents	East Asia	US gun violence	Mariano Rajoy
6	Israel-Palestine relations	Economy of various countries	Khashoggi assasination	US political sexual scandals
7	North Korean politics	UK Prime Minister	US history	Gaza border protests
8	Africa	US Presidents	Trump & Kavanaugh	Trump & Kavanaugh
9	Trump	Trump Presidency	Syria	Hawaii earthquake & volcano
10	Latin America	European politics	Hurricanes	US gun violence
11	War in Afghanistan	Trump	US Presidents	Spanish Politics
12	North Korea-South Korea relations	US Finance	ISIL in Syria	Inter-Korean summit
13	ISIL in Syria	US Cities	MeToo & Weinstein Effect	Afrin Offensive, Syria
14	Global Finance	Facebook	Mueller Investigation	Kabul
15	Economy of various countries	Trump & Kavanaugh	Hawaii earthquake & volcano	Skripal poisoning
16	Syria	US Military	Italian Politics	Kavanaugh Supreme Court Nomination
17	Putin & Russia	Big Tech	Russian political interference in US	SK
18	US Presidency	Far-right politics in Europe	California Wildfires	US history
19	US Military	US States	Iranian Protests	Venezuelan politics & Maduro
20	US Political Houses	UK	Chemical weapons in Syrian Civil War	Crimean Annexation

4.6 Results and Discussion

Studying the contents of the emergent Topics of Attention in Table 4.2 reveals various interesting details on current events as recorded on (English) Wikipedia. Identified features include; a background concept space, strong geographical effects (including a heavy Anglosphere/US focussed bias), a focus on individuals, and breakout subtopics.

Background concept space

Several of the top Topics of Attention by number of associated events (Countries, Global Cities, US Presidents, etc.) are those of lasting historical context. These topics also typically have high structural similarity—attention towards the topic is correlated with its structural composition on Wikipedia. Whilst the Event Networks are sampled from a current events records, much of the related content is built on and widely considered as part of long established encyclopaedic knowledge. This supports the case of news events contributing to longer term narratives.

Strong geographical effects

The Topics of Attention are strongly characterised by geography. Many of the labelled topics are specified by the region they are relevant to. It is also clear that when incorporating the structure of the knowledge graph and attention that many of the most prominent topics are Anglosphere focussed. This is of course partly a consequence of studying the English Wikipedia. However, the current events portal's nominal aim, and that of English Wikipedia as a whole, is to objectively cover global events and knowledge—something it still falls short on. Topics relating to the US and UK are covered with far higher granularity than those relating to other countries. That is, there are several top topics related to the intricacies of US politics, yet other countries typically have all related news summarised within a single topic. This is not entirely unsurprising, given prior work on Wikipedia biases (Graham et al., 2014; Callahan & Herring, 2011; Hecht & Gergle, 2009) as well as this work's focus on the English language Wikipedia. Nevertheless, this further validates assertions that rather than being the "sum of all human knowledge", Wikipedia (in its

various languages), through its content, structure, and access patterns is highly sensitive to its cultural setting.

Focus on individuals

Several of the top labelled topics are focussed on, or strongly feature a powerful individual (e.g. Trump, Putin & Russia, Mariano Rajoy, Maduro & Venezuela). This points to an audience sensitivity towards people that can be related to or reviled and is reflective of findings on the news values of celebrity/power elite. I further explore this relation to news values in the next chapter.

Breakout subtopics

There are several cases where topics may be strongly related, yet one cluster achieves breakout popularity enough to distinguish itself from the original topic. These could correspond to the well studied phenomena of "media storms" (Boydstun, Hardy, & Walgrave, 2014), whereby there is intense media focus on a single issue. An example of this is the topic for the Brett Kavanaugh Supreme court nomination—representing an overview of related events—and the Trump & Kavanaugh topic—which is the subject of more intense focus by the audience as indicated by the differing struxtural similarity scores. Another example is the two FIFA World Cup topics, where one represents stable knowledge attracting attetnion around the event and one represents new, unusual combinations of articles accessed concurrently. This may be a consequence of the choice of Jaccard similarity for the higher-level graph edge weights, where news events create strong, synchronous deviations from typical page view behaviour across a very small group of articles that are still related to a wider group. Since the number of deviating articles for the individual event is small, the edge weight through Jaccard similarity to other related events with a larger set of articles is also small, leading to it not being included in the Topic of Attention. An alternative similarity metric such as the overlap (Szymkiewicz-Simpson) coefficient could account for this effect, though using this would likely smooth over any breakout clusters, interesting features unto themselves.

Further remarks

The qualitative discussion of results and exploration of content from this chapter is important in contextualising findings in the later chapters, as well as Wikipedia status as an 'independent' data source for news media. Beyond simply being indicators for the notable issues in the news over a year, the detected Topics of Attention and their properties are demonstrative of the previous assertion that there is a disconnect between the 'editor's' Wikipedia and the 'page viewer's' Wikipedia. The central tension of Wikipedia as both a slow moving encyclopaedic knowledge base and fast moving current events record is displayed in the ways the topics are constructed. In the first mode, the audience's access patterns align with the established structure of knowledge on Wikipedia. The truly interesting mode occurs when the audiences attention does not align with the article network and in effect establishes its own communities of related articles. This collective behaviour is what stretches Wikipedia both towards updating its content and remaining a popular information resource, and away from its traditional encyclopaedic grounding.

There are several limitations to the methods proposed in this chapter. Firstly is the issue of this being a single language study. There have been a number of articles on the varying content, coverage, and use of different language Wikipedias based on linguistic, cultural, and national focusses (Aragon et al., 2012; Bao et al., 2012; Hale, 2014; Lemmerich, Sáez-Trumper, West, & Zia, 2019). One could contend that this means a single story from one community of people editing and viewing Wikipedia. A strong Anglosphere bias is indeed observed but I see it as the case the English Wikipedia is not the product of, nor the information tool, for a single, large, homogeneous community. Welser et al. (2011); West, Weber, and Castillo (2012); D. Yang, Halfaker, Kraut, and Hovy (2016) all observe that certain editors occupy particular roles in lending substantive expertise towards particular categories, whether that be due to identity, education, or other personal interest, and the same would be expected of regular users (to some extent also supported by Singer et al., 2017). In addition, the majority of the methods used are language agnostic, and may be swiftly applied to other language Wikipedias, which may be a

fruitful avenue to pursue beyond this thesis.

The current events portal is clearly not an exhaustive source of news stories, many of which would have no discernible effect on Wikipedia. Explicit editing guidelines state that "Stories added to the main portal page should be of international interest" (*How the Current events page works*, 2021). Beyond this restriction, there are a very large number of people who regularly access information related to sports, entertainment, and popular culture, whose news stories are rarely featured on the current events portal. Celebrity deaths, for instance, have their own summary article, rather than residing on the current events portal. The topic map thus does not cover the universe of what might be considered news, and is sensitive to the contents of the news story source, the Wikipedia Current Events Portal, raising the issue of endogeneity. Sampling news events by 'Wikifying' (Milne & Witten, 2008) alternative sources such as news website RSS feeds would indeed yield a different set of events, though unfortunately due to editorial decisions, we of course arrive at a similar obstacle where there is no objective set of events. A more thorough comparison between the topic landscape of several different news outlets would be of interest and an immediate application of the developed methods towards agenda-setting research, yet is outside the scope of this thesis. An argument in favour of choosing the Wikipedia Current Events Portal is that this collaborative recording of news is representative of the collective received importance of events, incorporating what the news recording and accessing communities consider relevant. This, together with time constraints and the simplicity of selecting descriptions already formatted with Wikipedia links, resulted in the decision being made to concentrate on the Current Events Portal.

4.7 Conclusion

The encyclopaedic origins of Wikipedia mean it is not set up as a ready-made data source for the study of news events—events and news topics do not have an established natural representation on Wikipedia. Equally, the broadly consistent, common structured information available to its huge audience, as well as how this audience accesses content on current events, is too appealing

to ignore. In order to take advantage of this, one must establish a framework for event and topic level study using Wikipedia data. To this end, I have developed an approach for topic detection on Wikipedia, with a focus on news topics, that takes into account article network structure, dynamics, and content. The graph supported correlation network approach towards temporal community detection successfully detects stable Event Reactions, relating both short-term dynamics of attention through page views as well as long-term knowledge structures thus addressing **RQ1a**. I have demonstrated its utility in identifying and exploring different Event Reactions, and in their aggregate how they represent Topics of Attention, the objective of **RQ1b**. These objects of study improve upon those used in prior work for their generality across topics, usage of the knowledge network rather than focus on individual articles, explicit relation to news events, incorporation of short and long term effects, and lack of reliance on detection through particular attention dynamics. The Topics of Attention on Wikipedia exhibit a background historical concept space, strong geographical effects, a focus on individuals, and breakout subtopics. The Topics of Attention may be resolved to volatile news topics and stable background topics. More importantly, they represent both facets of Wikipedia as a stable knowledge base and rapidly updating current events record.

Detecting Topics of Attention using Wikipedia has proven to be a non-trivial task. It is important to encapsulate the contrasting timescales of news and existing knowledge, the many to many relationship between news events and topics they feature, together with the corresponding dynamics of attention and memory building. Through this process we gain insight how news topics are represented and accessed on Wikipedia and on which events are considered important enough to make it into the encyclopaedic record. Finally, generating Event Reactions, and wider Topics of Attention enables the detailed event and topic level study that is the focus of the remaining thesis chapters. Now that we have established representations of events and topics on Wikipedia, we can move on to quantitatively studying them and addressing questions on news media theory in ways not previously possible without this kind of massive audience level data. This task is undertaken in Chapter 5.

Chapter 5

What Makes an Event News?

5.1 Introduction

In Chapter 4 I mapped the content and topics covering one year of events on Wikipedia. Yet the mechanism by which information on events is communicated and drives users to both visit and edit the site—the news process—remains unstudied. This begs the question; what makes events news? One might expect that certain measurable properties, perhaps associated with topic, affect the likelihood of a story being published. At a basic level, we tend to think things like more deadly disasters, more important individuals, and more unusual incidents mean an event is more likely become news and attract more coverage. But how can heuristics like these be made consistent and operationalised? In this chapter, I investigate wider properties of news events, centred around the journalism studies theory of news values.

News values are the set of journalistic criteria and properties of events that define how likely it is to be selected for news coverage—its "newsworthiness". Various formulations haven been proposed since Galtung and Ruge's (1965) initial proposal, but the modern standard tends to be based around Harcup and O'Neill (2017) which covers news values of Magnitude, Relevance/Proximity, Prominence, Organisation Agenda, amongst others. Coupled with their proposed news values, Galtung and Ruge also put forward several hypotheses as to how they relate to each other and to newsworthiness. Firstly, the 'complementarity' hypothesis; that events low in a given news value should exhibit high values in other news values—we should "expect

a negative correlation between two news factors for a defined universe of news items" (as later articulated by Sande, 1971). Secondly, the 'additivity' hypothesis; that events which exhibit high values in several news values are more newsworthy, i.e., are more likely to be selected for news coverage. Finally the 'exclusion' hypothesis; that events which exhibit low values in several news values are less newsworthy. Rosengren (1970), however, takes great issue with how these values are typically applied and tested for. Galtung and Ruge's original study, and many more following it, only consider events as reported in news media. As such, there is no way of knowing the effect of news values about unreported or underreported events—a survivorship bias is introduced. Rosengren calls for objective "extra-media" data in order to truly evaluate the nature of news values and newsworthiness. Absolute objective databases are of course difficult to come by; news is practically the means by which information about events is publicly communicated. However, this does not mean attempts at improving the objectivity and independence of event databases should not be made.

Following the early development of news values theory, the emergence of the Internet has brought with it, through the tracking of users' browsing patterns, wide ranging extra-media data on audience perception of events on an unprecedented scale. We can can use this opportunity to identify comprehensive extra-media data sources on events and understand how they reflect and relate to traditional news media. I present Wikipedia as one such data source. Wikipedia editors practice a form of citizen journalism, yet this is typically distant from any original or institutional reporting of the events themselves, as well as from any individual news source. Wikipedia editors are individually influenced by a range of news sources (in addition to historical records and other primary sources), yet must collaboratively generate some representation of events. Furthermore, the guiding principles of Wikipedia as a collaborative encyclopaedia encourages consideration of enduring notability and neutral point of view editing, and discourages any original reporting. The purpose is not to report news, but to record events of relevance to the encyclopaedic record. We then further exploit the disconnect between the browsing behaviour of users interested in current events (who are guided by news value and newsworthiness indicators)

and the encyclopaedic information available. This may be combined with a large database of news reports from varied sources, where newsworthiness can be evaluated on a population level by the number of articles in different outlets. I thus use Wikipedia, in both the structured information crafted by its editors and how it is accessed by its audience, as a baseline for the study of news events.

In this chapter, I use Wikipedia to establish and explore quantitative formalisations of news values, test the key hypotheses and their relation to newsworthiness. To do this, I evaluate what it means for Wikipedia to be an 'independent' current events source, explore how news values can be represented in the Wikipedia current events portal, use this database of events to explore the relation between news values and news topics, test relations between news values (the complementarity hypothesis), and model news coverage based on events' fulfilment of these news values—testing the additivity/exclusion hypotheses. This is framed under **RQ2**.

- **RQ2:** How are traditional conceptions of news values and newsworthiness of events reflected in extra-media data?
  - RQ2a: What types of events, according to news values, are recorded on Wikipedia?
  - RQ2b: How does the complementarity hypothesis apply to extra-media data?
  - RQ2c: How are news values associated with newsworthiness?

5.2 Wikipedia as Extra-Media News Data

As detailed in the Wikipedia editing guidelines, "Wikipedia is not a newspaper"—there are various practices and policies on the site that differ it from news media including its own notability criteria, limitations on the commitments of volunteer editors, and prohibition of primary research and source materials (*What Wikipedia is not*, 2021). It is however referred to as a "news backgrounder" as a complement to the rather less successful Wikinews site, which is designed explicitly for breaking news coverage. The Wikimedia Foundation still proudly praises Wikipedia's news credentials in its relation to Wikinews as regular "professional news organizations have no

encyclopaedia as a sibling project to call upon". I would argue that, whilst there are some specific beneficial integrations with Wikinews, Wikipedia's scope, size, and wide adoption make it the de facto sibling encyclopaedia for any professional news organisation and indeed a large proportion of the information on the wider web. Platforms such as Facebook, Google, YouTube, Twitter, Amazon Alexa, and Apple Siri routinely rely on the labour of Wikipedia editors and Wikimedia staff in producing their own knowledge graphs, informing automated search results and infoboxes, verification of notable persons, and directing their users to authoritative sources on issues of conspiracies and misinformation (Matsakis, 2018; Withers, 2018; Perez, 2020; TwitterInc., 2020; Vincent & Hecht, 2021). This is demonstrative of the fact that in many cases inclusion of a subject on Wikipedia in and of itself constitutes notability. There is broad (if sometimes misguided) consensus that Wikipedia is now *the* authoritative source for important information on the web.

But what are Wikipedia's rules for inclusion of information? Wikipedia's general notability guideline is that the "topic has received *significant coverage* in *reliable sources* that are *independent* of the subject, it is *presumed* to be suitable for a stand-alone article or list" (*Wikipedia: Notability*, 2021). More specifically, for the notability of current events editors are encouraged to consider the "impact, depth, duration, geographical scope, diversity and reliability of the coverage, as well whether the coverage is routine" (*Wikipedia: Notability (Events)*, 2021). There are also warnings to editors on recentism—the tendency for recent events to seem more important than they might do in in a few years time—and a wariness towards deviating too far from its encyclopaedic origins. Keegan (2013) notes how some of these notability criteria are not dissimilar from Galtung and Ruge's news values. We can see already some of these guidelines will tend to exclude certain kinds of stories that appear in regular news media. For example, daily celebrity news, detailed accounts of individual sporting events, events of solely local importance, and tabloid speculation. Having said that, entertainment and sport are two of the most popular topics of interest to regular viewers of Wikipedia (Singer et al., 2017)—mirroring the struggle within journalism between the popularity of soft 'infotainment' news and hard news. Even rel-

atively small events with little to no detailed record on the encyclopaedia drive large volumes of users interested in background information to the website.

Whilst there is a degree of shared values and criteria towards event selection in both Wikipedia and news media, the incentives that drive publication in news outlets are not necessarily the same as those that fuel the generation of content on Wikipedia and its use as a news backgrounder. Wikipedia editors are in principle unpaid and uncredited (at least on the main view of a typical article). Wikipedia editors, through the collaborative editing process and "neutral point of view" policy are encouraged to make relatively small deviations around 'established facts', with a regression towards some middle ground. News articles on the other hand may exhibit more variance, from partisan thinkpieces to original investigative reporting. The Wikipedia editor can afford, and is encouraged, to remain comfortably conservative in the content they produce, whereas the journalist must push the boat out in producing hot takes and providing interesting original insight and information. In terms of the news audience, Wikipedia readers typically navigate to specific areas of the site after finding out about events from elsewhere. This is a rather more targeted information search than browsing news sites to discover the day's events, or being algorithmically served the next news item of interest. Articles on Wikipedia do not have to compete for clicks with each other in the way news headlines must do. Aside from a short 'In the News' section on the Wikipedia homepage, there is little done to push users towards information on current events, nevertheless users still flock to the encyclopaedia's articles in line with the latest trends (Yoshida, Arase, Tsunoda, & Yamamoto, 2015).

We can contextualise this with some basic stats on the data used in this chapter; the Wikipedia current events portal and a LexisNexis news article database from 153 major news publications in Table 5.1 (see Chapter 3 for more details). The figures help characterise the practices of shorter, incremental contributions by a wider range of individuals on Wikipedia, compared to longer pieces from a smaller range of individuals and organisations in the news story data.

Table 5.1: Figures on news and Wikipedia article authorship. Edits are collected for the current events portal across the full 12 months, and edits to the associated articles in the Event Reactions are collected ±1 day from their respective event date. News articles from the LexisNexis database are collected according to search terms from the Wikipedia event records and a second stage of text similarity scoring (further detailed in Section 5.3.4).

	Articles	Revisions	Unique editors/authors	Median contribution (characters)
LexisNexis news database	114,721	-	$\approx 10,000$[1]	2,566
Wikipedia Current Events Portal	12	23,557	2,522	49[2]
Associated Wikipedia articles	33,676	620,565	121,536	33[3]

We see the similarities and differences of news recording on Wikipedia and news media. Wikipedia policies and practices are the way that news values manifest in the audience's remembrance and recording of events to collective history. I set about operationalising and measuring these extra-media news values, and explore how the core hypotheses apply to the news audience.

5.3 News Value and Newsworthiness Features

There are a variety of different overlapping perspectives with different conceptions of news values. I select a subset of 8 news values to be studied mostly based on the Harcup and O'Neill (2017) taxonomy (though they are also reflected in many other proposals); Prominence, Magnitude, Surprise, Uniqueness, Follow-up, Power Elite, Proximity, and Good/Bad news. These news values were also selected for their capacity to be operationalised using the Wikipedia data.

I detail 8 features as proxies to news values measurable from the Wikipedia events data, which are also summarised in Table 5.2. The Wikipedia events data is composed of 3,250 events, each with a network of related Wikipedia articles and associated page view time series. Through the work of Chapter 4, each event is deconstructed into (≈ 5) Event Reactions—unique responses over distinct groups of more closely related articles—for each of which I measure the 8 news values. In Chapter 4 I also detail Topics of Attention—groups of Event Reactions from

[1]Parsing of author names is done from article byline. This is likely an upper bound as many names would remain uncombined e.g. P Gildersleve & Patrick Robert Gildersleve.

[2]This through only considering edits which net contribute content, there are also many removals and reverts of changes. Overall median revision difference = +12 characters

[3]This through only considering edits which net contribute content, there are also many removals and reverts of changes. Overall median revision difference = +2 characters

different events linked by common content—which are also utilised as topic labels in this chapter. News value features are calculated for each Event Reaction (and are later combined for event-level analysis). Several of these features are based on a constructed time series for each Event Reaction. The time series, $W(t)$, is a sum of the hourly page views ($p_i(t)$) of each article in a Event Reaction R weighted by their PageRank centrality (w_i) to the network of articles;

$$W(t) = \sum_{i \in R} w_i p_i(t) \tag{5.1}$$

The time series is then centred in time to the max value occurring ± 24 hours from the day of the event (i.e. total window size of 72 hours).

5.3.1 Attention Based Features

Prominence

Prominence is widely thought of as a measure of the importance of a subject, independent of the particular event taking place. I measure the importance of the constituent Wikipedia articles in a news reaction using the median weighted page views for the 30 days (720 hours) prior to the event;

$$\text{Prominence}_i = \text{median}(W(-720, -719, ..., 0)) \tag{5.2}$$

Magnitude

If the event itself is of significant, wide ranging importance it will attract a large number of Wikipedia viewers above its baseline. This feature measures the amplitude of any increase in page views related to the event

$$\text{Magnitude}_i = W(0) - \text{Prominence}_i \tag{5.3}$$

This describes the excess of attention attributable to the event to the articles in the community. Events of high Magnitude attract a large number of excess views.

Surprise

High Surprise events can be characterised by a relatively large, sudden increase in page views. I scale the weighted time series $W'(t) = \frac{W(t) - \text{median}(W(-720, -719, ..., 0))}{\text{IQR(W(-720, -719, ..., 0))}}$ (according to median and interquartile range (IQR)) and take the maximum gradient to the peak value from the prior timesteps;

$$\text{Surprise}_i = \max\left(\frac{W'(0) - W'(-n)}{n} \forall n \in [1, 48]\right) \tag{5.4}$$

5.3.2 Relational Features

Uniqueness

If a story is the first in time (or only) in a Topic of Attention, it is unique (at the time of release), in that the subjects of the associated Wikipedia articles have not been previously observed in the news. The inverse Uniqueness score for a news reaction (N_i) is calculated by counting the number of news reactions (N_j) in the parent topic (T_{N_i}) that occurred prior to the event in question, divided by the amount of time (in days) passed from the beginning of the data window, with Uniqueness correspondingly defined as

$$\text{Uniqueness}_i = \left(\frac{|\{N_j \in T_{N_i} : date_N < date_{N_i}\}|}{date_{N_i} - date_0}\right)^{-1} \tag{5.5}$$

where $date_0$ is 01/12/2017, the first day in the period of news being studied.

Follow Up

Some news stories are produced as a follow-up on news events fresh in the public consciousness. Our news topics allow us to link stories according to content and timescale. To determine whether a news reaction is a follow-up, I compare it to the other items in the attention topic, counting the number of news reactions that precede it in the previous week. In this instance, news reactions from the same topic but well separated in time are not considered follow-ups. This extra time-sensitivity distinguishes this feature from (inverse) Uniqueness. For a news reaction N_i

$$\text{Follow-up}_i = |\{N_j \in T_{N_i} : 0 < date_{N_i} - date_{N_j} < 168\}| \tag{5.6}$$

5.3.3 Event Content Features

Power Elite

The Power Elite news value concerns events related to famous individuals and powerful organisations. Entity recognition using spaCy Python library (Honnibal, Montani, Van Landeghem, & Boyd, 2020) is performed on the event descriptions, linking entities identified as persons or organisations to their Wikipedia article if it is present in the news reaction. The median of the partial PageRank weighted sum of page views (as used in Equation 5.1) towards articles for persons/organisations is then calculated.

$$\text{Power Elite}_i = \text{median}(\sum_{j \in \text{persons/orgs}} w_j p_j(t)) \tag{5.7}$$

Proximity

Proximity (or relevance) is typically a descriptor of the geographical or, more frequently, cultural proximity of the news events to the audience. Given I am studying the English language Wikipedia, I take a simple measure of considering proximal stories to be those concerning the 'core Anglosphere' (A) – the UK, US, Canada, Australia, New Zealand (Bennett, 2007). Entity recognition is applied to the event descriptive text, and any locations are parsed to the country level using OpenStreetMap to give the country set L for each news reaction. The fraction of the locations in each descriptive text that are part of the 'core Anglosphere' is then stored.

$$\text{Proximity}_i = \frac{|L \cap A|}{|L|} \tag{5.8}$$

Good/Bad News (Sentiment)

How good/bad an event is is frequently thought of as a news value. Sentiment analysis on the event descriptions was applied using VADER NLP (Hutto & Gilbert, 2014). VADER assigns a string of text a score from 0 to 1 on its 'positive', 'negative', and 'neutral' sentiment, as well as an aggregate 'compound' measure from -1 (negative) to +1 (positive). In principle one could use two features for Good News and Bad News separately. However, I choose to take the

Table 5.2: A summary of the news value features.

News Value	Description	Formula
Prominence	How important or popular the subjects of the event are prior to it occurring.	$\text{median}(W(-720, -719, ..., 0))$
Magnitude	How important the event itself is and how wide-ranging its effects are.	$(0) - \text{Prominence}$
Surprise	How unexpected the event is.	$\max(\frac{W'(0)-W'(-n)}{n} \forall n \in [1, 48])$
Uniqueness	How much the event is unrelated to all other events.	$\left(\frac{\lvert\{N_j \in T_{N_i} : date_{N_j} < date_{N_i}\}\rvert}{date_{N_i} - date_0}\right)^{-1}$
Follow-up	How much the event is related to other recently occurring events.	$\lvert\{N_j \in T_{N_i} : 0 < date_{N_i} - date_{N_j} < 168\}\rvert$
Power Elite	How much the event concerns famous or powerful individuals and organisations.	$\text{median}(\sum_{j \in \text{persons/orgs}} w_j p_j(t))$
Proximity	How culturally or geographically relevant the event is to the audience.	$\frac{\lvert L \cap A\rvert}{\lvert L\rvert}$
Sentiment	How good/bad the event is in nature.	$VADER(text_{N_i})_{\text{compound}}$

singular measure of compound score, for better agreement with manual labelling in validation (section 5.3.5), and to reduce later model complexity. The feature is correspondingly labelled as Sentiment.

$$\text{Sentiment}_i = VADER(text_{N_i})_{\text{compound}} \tag{5.9}$$

Whilst relatively basic and not providing a value judgement on what should be good or bad news, the sentiment analysis is capable in identifying positive/negative scores for descriptions including words such as 'win', 'celebrate', 'death', 'crisis' etc.

5.3.4 Newsworthiness

I also consider the level of news coverage of an event, in effect its newsworthiness. Using an online database from LexisNexis of one year of news articles published in major news outlets

worldwide (see Chapter 3 for more details) in the same time period one may assess how the number of news articles published on each event relates to the news values as measured from Wikipedia data. If an event is more newsworthy, it is more likely to be published across the news outlets and more news articles about it will be detected in the dataset. In total, there are $114,721$ news articles selected, written by $\approx 10,000$ unique authors. An event's Newsworthiness (# news articles) is then given by the number of matched news articles from this database.

5.3.5 Feature Preparation and Validation

Validation by two human coders was carried out for the NLP based features (Sentiment[4], Power Elite, Proximity). Coders were informed of what the respective news values correspond to, and asked to identify whether a particular event exhibited them, based on the constituent articles and event description (Power Elite: 0 or 1, Proximity: 0 or 1, Sentiment: -1, 0, or 1 (bad, neutral, good)). This was conducted across a random sample of 100 news reactions. The results were compared against categorical versions (Power Elite: 0 or > 0, Proximity: 0 or > 0, Sentiment: < -0.05, 0, or > 0.05) of the extracted features and against each other, with the results in Table 5.3. Strong precision and recall across all the metrics, as well as inter-coder consistency indicates the features are appropriate to proceed with. A drop in precision and recall is observed for Sentiment, though NLP F1 scores are comparable to inter-coder F1, suggesting that news being good/bad is a somewhat subjective judgement, as might be expected. In addition, in further analysis, long tailed features (Prominence, Magnitude, Surprise, Follow-up, Uniqueness, Power Elite, Newsworthiness) are transformed according to $x' = \log(1 + x)$. All feature distributions and correlations are displayed in Figure 5.1. Prominence, Magnitude, Surprise, Follow-up, Uniqueness, Power Elite, Newsworthiness are all already transformed in these figures. Since there are more than 1000 topics containing more than one reaction, fine grain interpretation of distribution is not possible, but there appear to be some groupings and trends. A notable example is the secondary spike in the Prominence distribution is highly associated with a low

[4] An earlier version of this manuscript had separate bad and good news features according to the negative and positive sentiment, but better agreement with manual labels is achieved with the compound score.

Table 5.3: Precision, recall, and F1 score for the NLP-based features against a manually labelled sample of 100 event descriptions by two independent coders. Inter-coder F1 score also provided.

	Metric	Power Elite	Proximity	Sentiment
Inter-coder	F1	0.91	0.84	0.71
NLP vs Coder 1	Precision	0.83	0.76	0.64
	Recall	0.89	0.81	0.66
	F1	0.86	0.78	0.61
NLP vs Coder 2	Precision	0.78	0.79	0.69
	Recall	0.94	0.72	0.70
	F1	0.85	0.75	0.69

Uniqueness group, almost exclusively formed of the Countries 'background' topic.

5.4 News Reactions, Topics, and Clusters

In this section I address **RQ2a:** *What types of events, according to news values, are recorded on Wikipedia?* The plots in Figure 5.1 suggest there could be some association between news values and news topic, which I investigate further in this section. Standardising the news value distributions to mean = 0, SD = 1, and comparing the average standard deviation for each news value by topic leads to the results in Table 5.4. The reduced within-topic variance, indicates that news values are associated with news topic. Verifying this effect across all news values, one can consider the Euclidean distance between all Event Reactions. The average intra- and inter- topic distance can then be compared. Average intra-topic distance = 2.34. and average inter-topic distance = 4.03. The effect is consistent across all reactions and across all topics. 99.4% of reactions are on average closer in news values space to members of the same topic, than those of different topics. All news topics also have a smaller average distance amongst their own members, than to members of other topics. This is a strong indicator that any modelling of news values and newsworthiness, within and beyond this work must control for news topic.

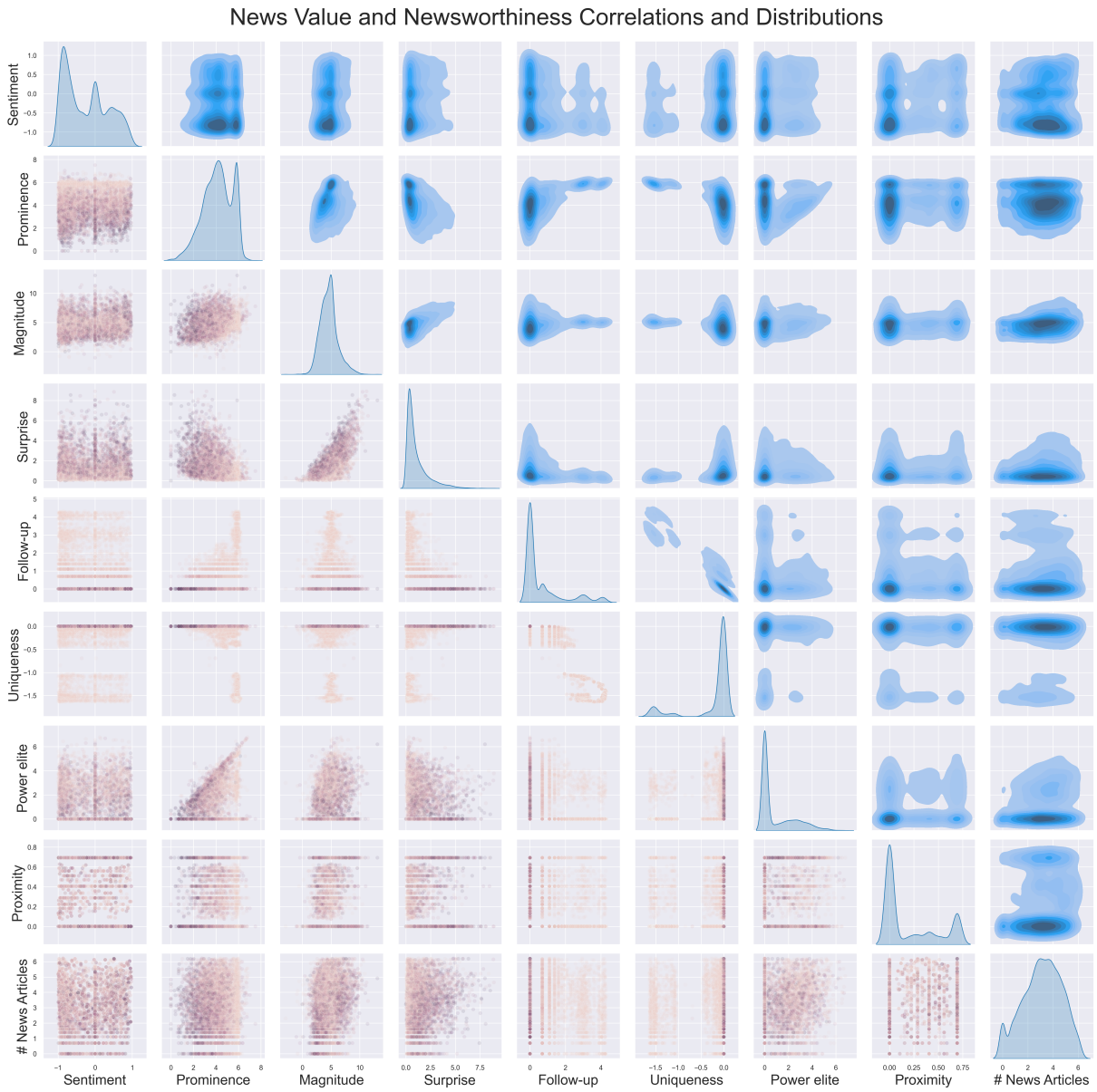

Figure 5.1: Kernel density estimate (KDE) plots of the news value feature distributions are provided along the diagonal. Scatter plots for feature correlations in the lower triangle (colour coded by the reaction topics). The corresponding 2D KDE plots for the pairs of features are provided in the upper triangle.

Table 5.4: Mean standardised standard deviation ($\bar{\sigma_{\mathrm{T}}}$) across topics for each news value. Population standardised standard deviation = 1, by definition.

σ	$\bar{\sigma_{\mathrm{T}}}$
Prominence	0.287
Magnitude	0.648
Surprise	0.623
Uniqueness	0.017
Follow-up	0.180
Power Elite	0.384
Proximity	0.450
Sentiment	0.656

5.4.1 Clustering

On evaluation of the findings of the previous section, one might expect this difference in intra-/inter- cluster distance to result in some clustering effect in the news value feature space. I perform hierarchical clustering to evaluate this on the individual news reactions, using the Euclidean distance metric with Ward linkage and calculate the full dendrogram (presented with the distance matrix in Figure 5.2).

Figure 5.2 seems to indicate some clustering within the data. I use the silhouette coefficient (Rousseeuw, 1987) to quantitatively evaluate cluster quality at different n clusters (Figure 5.4). Whilst there is no clear global maximum (aside from at initial $N = 2$ clusters), the first local maximum occurs at $N = 6$ clusters. I also consider how well the hierarchical clustering obtained matches the topic labels using adjusted mutual information (AMI) (Vinh, Epps, & Bailey, 2010). A peak value of $AMI = 0.37$ is also obtained with 6 clusters (also Figure 5.3). Whilst cluster alignment with the news topic labels is not strictly the aim, the additional support for meaningful separation in a 6 cluster model from the silhouette coefficient means this model is worth exploring further. I thus investigate the properties of these 6 clusters, with cluster feature means plotted in Figure 5.4. An important caveat is that the lack of global maximum in silhouette coefficient means there is no definitive clustering scale, pointing towards relatively continuous variation over all news reactions. Clear clusters are likely recoverable if considering

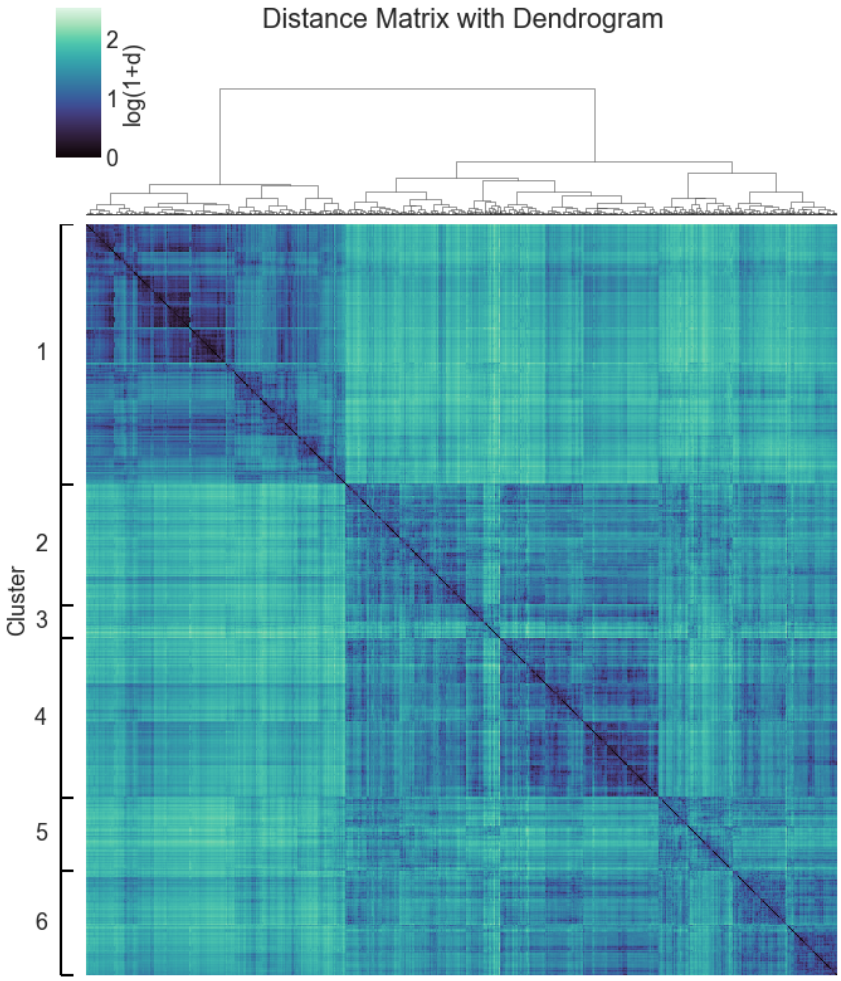

Figure 5.2: Clustering dendrogram with sorted distance matrix indicating the Euclidean distances between Event Reactions. Cluster memberships are also indicated. (Log scale colour bar for distance for visualisation purposes)

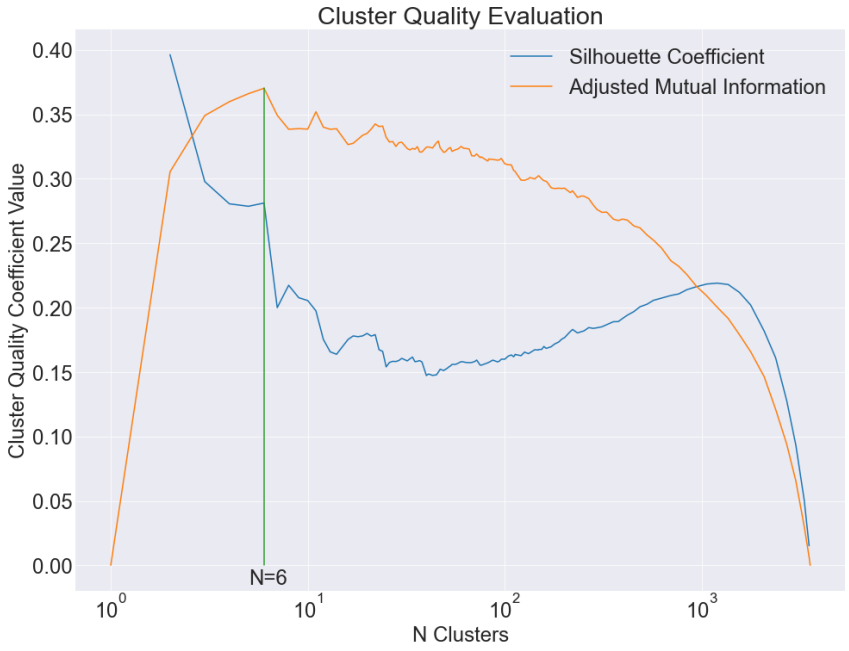

Figure 5.3: Hierarchical cluster quality, as evaluated by silhouette coefficient, and Adjusted Mutual Information to the topic labels.

points from some small subset of topics, but the large number of varied events across a range of topics washes out the full population level effect. Nevertheless, there is still merit in exploring the 6 cluster model as a distilled description of all Event Reactions.

The six clusters can be described by:

1. Prominent, continuous, background topics.
2. Power Elite focused topics.
3. High Surprise and Magnitude topics.
4. Weak response across almost all values.
5. Power Elite, proximal, prominent, positive Sentiment topics.
6. Proximity focused topics.

Examples of members of each of the Event Reaction clusters are provided in Table 5.5.

Figure 5.4: Mean news values for each cluster (Population mean=0 SD=1).

Table 5.5: Example Event Reactions closest to each cluster centre. Each event will also likely have other Event Reactions over a different set of Wikipedia articles centred on a different key article.

Cluster	Event Title	Event description	Reaction key article(s)	Notes
1	Oil bribery case in Italy	(26/11/2018) A court in Milan, Italy, examines evidence submitted by the campaign group Global Witness, that alleges bribery around the attribution of the OPL 245 oil prospecting license in 2011 led to a loss for the Nigerian state estimated at US$6 billion	Italy	Prominent article, but no additional relevance to story beyond location. Little additional online response to key article and its neighbours.
2	French protest against Macron's economic reforms	(22/03/2018) Protests against Emmanuel Macron. People in 150 places across France take to the streets peacefully in a general strike to protest President Macron's economic reforms. Railways, airways, schools and power generation are affected.	Emmanuel Macron	Reaction focussed on prominent individual, though not Anglosphere focussed or shock response to event.
3	FBI director resigns	(29/01/2018) Presidency of Donald Trump. Andrew McCabe resigns as Deputy Director of the Federal Bureau of Investigation amid a dispute with President Donald Trump.	Andrew McCabe, Deputy Director of the Federal Bureau of Investigation	High Magnitude and Surprise reaction. Individual in question however is not particularly prominent prior to this story.
4	Referendum proposed on anniversary of Iranian Revolution	(11/02/2018) Anniversary of the Iranian Revolution, 2017–18 Iranian protests. Iranian president Hassan Rouhani proposes a referendum to heal country's divisions, according to the Article 59 of constitution.	Iranian Revolution	Low level of news value fulfilment and online response around key article and neighbours.
5	Resumption of peace talks between USA and North Korea	(25/05/2018) 2018 North Korea–United States summit. U.S. President Donald Trump tweets that "very productive talks" are being held with North Korean leader Kim Jong-un on reinstating the June 12 Singapore summit, which he had cancelled Thursday.	Kim Jong-un, North Korea	Reaction centred on prominent individual, event is relevant to Anglosphere and is of a positive leaning.
6	Calgary residents vote on Winter Olympics bid	(13/11/2018) 2026 Winter Olympics. Residents of the city of Calgary head to the polls to vote on a non-binding plebiscite to determine if the city should make a bid to host the 2026 Winter Olympic Games. The unofficial results suggest the No vote is in the lead, and only a 40% voter turnout. Official results will be released Friday, November 16.	Calgary	Event of Anglosphere relevance, but relatively low Prominence and response.

Clusters 1 and 4, whilst descriptive of Topics of Attention on Wikipedia, do not seem to map well towards would normally be considered news topics. Cluster 1 has high Prominence, but relatively low Magnitude (despite the average absolute value), similarly cluster 4 has average Prominence, relatively low Magnitude, and weak response across other news values. The key difference between them is on the axes of Follow-up and Uniqueness. Cluster 1 reactions are regularly part of the associated information with news records on Wikipedia, whereas cluster 4 topics are less frequently mentioned in the news cycle.

The typical 'shock' breaking news story reactions are captured by cluster 3, with high Surprise and Magnitude responses to the event, interestingly with slightly lower Proximity. Clusters 2, 5, and 6 are characterised by their strong association with Power Elite (2), Proximity (6), or both (5). Cluster 5, concerning high Power Elite and Proximity reactions, is also of relatively higher Prominence and Magnitude than clusters 2 and 6. Cluster 5 also has the largest absolute deviation for Sentiment, with cluster 5 events typically more positive than those in other clusters.

5.5 The Complementarity Hypothesis

In Galtung and Ruge's seminal news value paper, they hypothesise that existence of one news value for an event tends to exclude the existence of others. There is a degree of ambiguity as to the extent to which this covers news values in general, specific combinations of news values, and whether it applies for all events or topics. As we have already seen, news values are associated with news topic, so one must investigate the effect of controlling for news topic. Furthermore, I can make use of the news article data to weight the correlations to better represent news media vs the Wikipedia recording of events. This motivates **RQ2b:** *How does the complementarity hypothesis apply to extra-media data?* I propose three variations of the complementarity hypothesis to test against. In this analysis all values are standardised to mean = 0, SD = 1.

- **The Weak Complementarity Hypothesis:** Minimum news value is negatively correlated with maximum news value of each Event Reaction.

- **The Intermediate Complementarity Hypothesis:** Each news value is negatively correlated with maximum of the other news values.
- **The Strong Complementarity Hypothesis:** Each news value is negatively correlated with each of the other news values.

These three variants, along with not controlling for / controlling for topic, and not weighting / weighting by number of news articles, give 12 total analyses. Weighted Pearson correlations are determined by weighting points in the mean and covariance functions by the transformed number of news articles (Newsworthiness), and calculating p-values with bootstrapped standard errors. To control across topics, relevant correlations are calculated for each topic (containing minimum 10 different reactions), and the mean correlation across topics is then calculated. In this case, p-values are evaluated according to a t-test on the distribution of correlations. Detailed results for the correlations are given in Tables 5.6 (weak), 5.7 (intermediate), and 5.8-5.11 (strong).

To summarise the tables: the weak complementarity hypothesis is confirmed for Wikipedia recorded events and their relative coverage in news media, when not controlling for news topic ($\rho = -0.23^{***}$ and $\rho = -0.17^{***}$ respectively). However, this negative correlation fades to a weak positive correlation ($\bar{\rho} = 0.11^{***}$ and $\bar{\rho} = 0.10^{***}$) when one does control for the topics. Minimum and maximum news values are anticorrelated between topics, but correlated within topics—in effect an instance of Simpson's Paradox (Lerman, 2018). Considering the intermediate hypothesis, the only news value that is consistently negatively correlated with the maximum news value is Uniqueness. Non-unique events must stand out in some other way—though this is effect is weakened when controlling for topics, so it is perhaps more appropriate to say non-

Table 5.6: Weak complementarity hypothesis Pearson correlations. Correlations are taken between the minimum and maximum news value for each news reaction. $^{*}: p < 0.05$, $^{**}: p < 0.01$, $^{***}: p < 0.001$.

	Wikipedia	News weighted
Population	-0.23***	-0.17***
Topic controlled	0.11***	0.10***

Table 5.7: Intermediate complementarity hypothesis Pearson correlations. Correlations are taken between the specified news value and the maximum of the remaining news values. * : $p < 0.05$, ** : $p < 0.01$, *** : $p < 0.001$.

	Wikipedia		News weighted	
	Population	Topic controlled	Population	Topic controlled
Sentiment	0.11***	0.03	0.14***	0.04
Prominence	0.26***	0.06	0.22***	0.07*
Magnitude	0.55***	0.30***	0.56***	0.29***
Surprise	0.16***	0.23***	0.21***	0.23***
Follow-up	0.04***	0.03	0.00	0.03
Uniqueness	-0.39***	-0.06*	-0.32***	-0.04
Power Elite	0.06***	0.03	0.07***	0.04
Proximity	0.11***	0.04	0.12***	0.02

Table 5.8: Strong complementarity hypothesis Pearson correlations for the Wikipedia record of events across all Event Reactions. * : $p < 0.05$, ** : $p < 0.01$, *** : $p < 0.001$.

	Sentiment	Prominence	Magnitude	Surprise	Follow-up	Uniqueness	Power Elite	Proximity
Sentiment	-	-0.007	0.137***	0.116***	-0.086***	0.063***	0.149***	0.157***
Prominence	-0.007	-	0.267***	-0.417***	0.562***	-0.561***	0.084***	0.059***
Magnitude	0.137***	0.267***	-	0.635***	0.138***	-0.118***	0.185***	0.111***
Surprise	0.116***	-0.417***	0.635***	-	-0.256***	0.247***	0.082***	0.05***
Follow-up	-0.086***	0.562***	0.138***	-0.256***	-	-0.924***	-0.079***	-0.053***
Uniqueness	0.063***	-0.561***	-0.118***	0.247***	-0.924***	-	0.107***	0.063***
Power Elite	0.149***	0.084***	0.185***	0.082***	-0.079***	0.107***	-	0.074***
Proximity	0.157***	0.059***	0.111***	0.05***	-0.053***	0.063***	0.074***	-

Table 5.9: Strong complementarity hypothesis Pearson correlations, weighted by news coverage, across all Event Reactions. * : $p < 0.05$, ** : $p < 0.01$, *** : $p < 0.001$.

	Sentiment	Prominence	Magnitude	Surprise	Follow-up	Uniqueness	Power Elite	Proximity
Sentiment	-	0.014	0.162***	0.124***	-0.087***	0.057***	0.144***	0.177***
Prominence	0.014	-	0.219***	-0.420***	0.544***	-0.546***	0.105***	0.056***
Magnitude	0.162***	0.219***	-	0.667***	0.088***	-0.067***	0.182***	0.119***
Surprise	0.124***	-0.420***	0.667***	-	-0.262***	0.252***	0.074***	0.064***
Follow-up	-0.087***	0.544***	0.088***	-0.262***	-	-0.912***	-0.075***	-0.056***
Uniqueness	0.057***	-0.546***	-0.067***	0.252***	-0.912***	-	0.105***	0.060***
Power Elite	0.144***	0.105***	0.182***	0.074***	-0.075***	0.105***	-	0.077***
Proximity	0.177***	0.056***	0.119***	0.064***	-0.056***	0.060***	0.077***	-

Table 5.10: Strong complementarity hypothesis Pearson correlations for the Wikipedia record of events and controlling for topic. * : $p < 0.05$, ** : $p < 0.01$, *** : $p < 0.001$.

	Sentiment	Prominence	Magnitude	Surprise	Follow-up	Uniqueness	Power Elite	Proximity
Sentiment	-	0.006	0.048	0.040	-0.041	-0.017	0.030	0.083*
Prominence	0.006	-	0.148***	-0.150***	0.112***	-0.127***	0.051	-0.006
Magnitude	0.048	0.148***	-	0.704***	0.096**	-0.012	0.075*	0.005
Surprise	0.040	-0.150***	0.704***	-	-0.007	-0.025	0.044	0.008
Follow-up	-0.041	0.112***	0.096**	-0.007	-	-0.313***	0.028	-0.036
Uniqueness	-0.017	-0.127***	-0.012	-0.025	-0.313***	-	0.028	-0.036
Power Elite	0.030	0.051	0.075*	0.044	0.028	0.028	-	-0.054*
Proximity	0.083*	-0.006	0.005	0.008	-0.036	-0.036	-0.054*	-

Table 5.11: Strong complementarity hypothesis Pearson correlations, weighted by news coverage, and controlling for topic. * : $p < 0.05$, ** : $p < 0.01$, *** : $p < 0.001$.

	Sentiment	Prominence	Magnitude	Surprise	Follow-up	Uniqueness	Power Elite	Proximity
Sentiment	-	0.012	0.063*	0.044	-0.036	-0.015	0.047	0.066*
Prominence	0.012	-	0.131***	-0.163***	0.097***	-0.135***	0.066*	0.003
Magnitude	0.063*	0.131***	-	0.700***	0.095**	0.000	0.086**	-0.005
Surprise	0.044	-0.163***	0.700***	-	-0.000	-0.014	0.052	-0.014
Follow-up	-0.036	0.097***	0.095**	-0.000	-	-0.325***	0.020	-0.037
Uniqueness	-0.015	-0.135***	0.000	-0.014	-0.325***	-	0.040	-0.042
Power Elite	0.047	0.066*	0.086**	0.052	0.020	0.040	-	-0.053*
Proximity	0.066*	0.003	-0.005	-0.014	-0.037	-0.042	-0.053*	-

unique topics are associated with higher maximum news value. Correlations when weighting for news coverage remain relatively unchanged. Finally, For the strong complementarity hypothesis, we observe consistent negative correlations for the news value combinations of [Surprise & Prominence], [Uniqueness & Prominence], and [Uniqueness & Follow-up]. However, the overall picture here is more of uncorrelated news values and a handful of significant positive correlations. With credit to Galtung and Ruge, this granularity of the intermediate and strong hypotheses is well beyond the realms of their initial proposal. Though it does allow us to see in which specific news value combinations the hypothesis still applies. I find that whilst the complementarity hypothesis holds in the most general sense, it does not hold when controlling for topics, or at finer news value granularity. In addition, there is little qualitative difference in results between the record of events on Wikipedia, and the events weighted by their news coverage.

5.6 The Additivity Hypothesis

5.6.1 From Wikipedia Articles to News Articles

I have thoroughly explored how news values vary and relate to each other over a population of one year of news events. But I am yet to evaluate their core purpose—describing how likely a story will be published about a particular event. This addresses **RQ2c:** *How are news values associated with newsworthiness?*

Here, I must be clear on the direction of causality. Clearly, it is not the popularity or content of Wikipedia that determines whether a journalist will publish a news story. Likewise however, it is not the content of the news outlets from the LexisNexis data that exclusively determine what news is recorded on Wikipedia, as well as how it is recorded. Each event record on the current events portal must cite at least one news source, but just 7.6% of them cite one or more of the 153 outlets from the LexisNexis data and 81% of the LexisNexis news sources are not cited at all. That is not to say that Wikipedia is without journalistic influence—no new information would be recorded without it—but that given the diversity of editors, their own diverse influences, as well as competing encyclopaedic editing principles, that make it about as good an independent baseline as one could expect. Furthermore, many of the coded news values are based on measurements from a time or timescale that occurs before or faster than the bulk of the news coverage, or would be expected to be resilient to any particular media manipulation. Prominence, Power Elite, Uniqueness, and Follow-up are primarily based on Wikipedia article popularity and relations to wider knowledge structures before an event occurs. Magnitude and Surprise, whilst clearly dependent on communication about the event itself, typically unfold over a matter of hours which is faster than the timescale that many of the news articles in the LexisNexis dataset are published and eventually read (see Chapter 6 for greater discussion of attention peak timescales). Proximity and Sentiment are based on the content of the event (where it was, to whom it may concern, what happened) which is clearly difficult for the entire journalism industry to get wrong or manipulate (though certain outlets might still try). I take the news values as measured from Wikipedia as a proxy for those that also inform the process

of news publishing.

5.6.2 Modelling Newsworthiness

Having disentangled separate Event Reactions for each event in Chapter 4, I must recombine them to obtain a single event-wide set of news values to model their relationship with the number of news articles published for a given event. The model takes into account that different topics may attract different baseline levels of news coverage and that, since news values are associated with topics, the respective news values *relative* to those typical for the topic. For the set of news values $\{v_1, v_2, ...\}$ from an Event Reaction r in topic t, the standardised news values $\{v'_1, v'_2, ...\}$ are given by

$$v'_\alpha = \frac{v_\alpha - \mu_t(v_\alpha)}{\sigma_t(v_\alpha)}, \tag{5.10}$$

where $\mu_t(v_\alpha)$, $\sigma_t(v_\alpha)$ are the mean and standard deviation of the news value v_α in topic t. Then a single set of news values is calculated for each event by averaging over the constituent Event Reactions from different topics. For an event E with constituent set of reactions R, its news values $\{V_1, V_2, ...\}$ are then expressed by

$$V_\alpha = \frac{\sum_{i \in R} v'_{\alpha i}}{|R|}. \tag{5.11}$$

The 100 topics are coded as binary variables, with membership = 1, non-membership = 0. We then have 100 topic variables and 8 event-level news value variables, so a total of 108 independent variables to be used in a linear regression model to predict the (log-transformed) number of news articles about each event (its Newsworthiness). Results are given in Table 5.12 (full regression coefficients in Appendix C), with model $R^2 = 0.40$, $R^2_{\text{adj}} = 0.35$. 38 of the 100 baseline topic variables and 4 of the 8 news values are significant to at least the 5% level. Significant positive coefficients for Magnitude and Power Elite ($B = 0.15^{***}$, $B = 0.13^{***}$) show that events which attract a large of excess of attention on Wikipedia and those involving prominent persons and organisations are also published about more in the news. In standardised log-transformed terms, a unit increase in an event's Magnitude or Power Elite news values results in a 0.14 or 0.13 respective increase in Newsworthiness. Re-exponentiating, this corresponds to

a multiplicative increase of a factor of $e^{0.1479} = 1.16$, $e^{0.1345} = 1.14$ in the number of news articles written about an event, equivalent to an increase above the geometric mean of 26.8 news articles per event of 4.3 or 3.9 more articles respectively. A significant negative coefficient for Sentiment ($B = -0.07^{**}$) indicates bad news is more newsworthy than good news, and that a unit increase in Sentiment news value results in a decrease in news articles written according to a multiplicative factor of $e^{-0.0746} = 0.93$. This is equivalent to a decrease of 1.9 fewer news articles below the geometric mean of 26.8 news articles per event.

A significant negative coefficient for Prominence is somewhat unexpected. This can likely be traced to how the Event Reactions are sampled in Chapter 4. Event Reactions are extracted using both the Wikipedia article network structure and correlated page view time series for individual articles. Highly newsworthy news events are more likely to generate a high Magnitude impulse of page views on some of these articles. This increases the average distance between the key articles with strong signals and other wider context articles with weaker to no signal (resulting in a lower structural similarity score). The community detection process then includes fewer of these context articles in the Event Reaction. Many of these excluded context articles are high traffic and do not contribute to the weighted page view sum (equation 5.1), reducing the Prominence. This 'within-topic' Prominence score then in effect captures how (un)focused within a topic an event is. Rather than taking the individual Prominence score of an Event Reaction, one could instead consider the average Prominence of the topic the Event Reaction belongs to as a better indication of the Prominence of the event's subject matter. Whilst this specific analysis was not conducted, this effect is essentially already captured by the 100 topic baseline variables. 40 of these are significant indicating that there is association between pre-existing importance of event subject matter and Newsworthiness.

I find no significant association between Newsworthiness and Surprise, Follow-up, Uniqueness, or Proximity. It is possible that the data collection strategy for news articles may disproportionately affect the predictiveness of these measures. News articles are only collected ± 1 day from the event, but particularly surprising or unique events may generate a greater share of

the news articles about them > 1 day from the event compared to typical news events. There may be similar issues surrounding the uncertainty in time of Follow-up events. Unfortunately, extending the time window increases the false positive rate of the automated collection to an unacceptable level. It is also possible that response to Surprise is nonlinear. One might initially intuit that surprising events are more likely to attract news coverage, however, events that are anticipated or more slowly evolve have greater opportunity for news coverage to be published before or as they occur—something not possible for complete surprise events. This subtlety is not captured by the model. For Proximity, this coarser grain Anglosphere measure is not significant for the worldwide news database, but a measure more tailored to a more local news database, set of events, and audience could prove more meaningful.

Table 5.12: Regression summary statistics and standardised news value coefficients (full coefficients in Appendix C).

Dep. Variable:	Newsworthiness	**R-squared:**	0.400
Model:	OLS	**Adj. R-squared:**	0.351
No. Observations:	1374	**F-statistic:**	8.146
Df Residuals:	1269	**Prob (F-statistic):**	3.0×10^{-84}
Df Model:	104	**Log-Likelihood:**	-1550.0
AIC:	3310	**BIC:**	3859

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Sentiment	-0.0746	0.025	-3.022	0.003	-0.123	-0.026
Prominence	-0.1469	0.033	-4.467	0.000	-0.211	-0.082
Magnitude	0.1479	0.029	5.090	0.000	0.091	0.205
Surprise	0.0065	0.026	0.251	0.802	-0.044	0.057
Follow-up	0.0553	0.035	1.579	0.114	-0.013	0.124
Uniqueness	0.0046	0.036	0.130	0.897	-0.065	0.075
Power-elite	0.1345	0.024	5.669	0.000	0.088	0.181
Proximity	0.0346	0.030	1.168	0.243	-0.024	0.093

5.6.3 The Most Newsworthy Events

For a final stage of analysis, I study the news values of the top decile of most newsworthy events against the rest of the population. Considering the averaged standardised news values as used in the linear regression model, the mean news values for each set of events are given in Table 5.13. We see significant positive difference in means across 5 of the 8 news values, furthering

the case for the additivity hypothesis when applied to the most newsworthy events.

Table 5.13: Mean standardised news values for the top 10% most newsworthy events vs the remaining 90%.

	Top 10% Newsworthy	Remaining 90%	Δ
Sentiment	0.18	0.05	0.14
Prominence	-0.57	-0.41	-0.15
Magnitude	0.70	-0.13	0.83***
Surprise	1.09	0.20	0.89***
Follow-up	-0.11	-0.47	0.36***
Uniqueness	0.04	0.36	-0.32***
Power Elite	1.08	0.26	0.82***
Proximity	0.35	0.07	0.28***

On the whole, this analysis supports the additivity hypothesis; the more an event fulfils particular news values, the more likely it is to be published about[5], though not across all values as measured in the Wikipedia data. The explained variance is not outstanding, but is satisfactory given the independent and dependent variables are generated from disconnected data sources (most news value studies take news value measures directly from news media source material or ask participants for their own views on the source material). Crucially, these news values come as measured from an audience-centric extra-media source and have controlled for the various news topics expressed by a given event. Beyond the suitability of the features as proxies for news values and random effects, there are other important factors not approached by the model that may be responsible for remaining variance. Competition amongst events and news services in generating the most attractive news articles on a given day is one factor that may affect the number of articles about a given event (consider the phenomenon of 'slow news days'). The explicit relation between the Wikipedia audience and what news they consume is also not captured. These factors may improve the model, but one would not expect a change in the underlying findings on news value additivity.

[5]Note that given some of data transformations, the relationship in the model is not strictly mathematically additive.

5.7 Discussion

We have observed how quantitative conceptions of news values can be extracted from Wikipedia—an extra-media data source. The observed associations between topics and news values are an important contribution to the literature. Many events are not independent from each other and the constituent elements that contribute towards particular news values and newsworthiness carry over from one event to the next. In the establishment of quantitatively defined topics from Chapter 4 I have systematically taken this into account where many prior studies have not (except for broader news categories). Intra-topic and inter-topic studies of news values must be clear over what set of events the observed relations hold true.

I have also delivered much needed theoretical clarification on the complementarity hypothesis. One would not expect such extreme results as the strong complementarity hypothesis always applying, but it has its use in discerning which news values are negatively correlated. These single score correlations act as a first step, but the relationship may not hold across the full range of a news value. Consider the weak complementarity hypothesis (minimum news value is negatively correlated with maximum news value). It might be the case that only for small minimum new values does the maximum news value have to be large and that above a certain minimum news value level the correlation fades. There is more detail to come from this line of inquiry. With regard to the additivity hypothesis, in many cases much of the predictive capacity is largely driven the topic constant in the linear regression. The news values—in this case the within-topic news value scores—have significant effects, but a large degree of the variance is captured by the average newsworthiness of each topic. Again, breaking down events by topic is a key step in modelling the relationship between news values and newsworthiness.

More generally, in some instances we are left to wonder whether the rejection (or a lack of confirmation) of hypotheses is a result of them being objectively false in any/all extra-media data, or that this particular data source and/or experimental design for whatever reason does not reflect news in that specific way. On its own, is this result a claim about Wikipedia or about news? An individual study is rarely enough to reform or overturn established theory.

However, we clearly do not observe the default result of having all traditional hypotheses apply in this extra-media data. Clearly this work should motivate further research into testing the foundational news value hypotheses in further large scale extra-media data from Wikipedia and beyond.

This study is not without limitations. In focusing on "material" news values as intrinsic properties of events I find several established news values not measured or even measurable from the Wikipedia data. In particular, news values described by (Caple & Bednarek, 2013; Bednarek & Caple, 2017) as "cognitive"—based on individual news workers' beliefs on newsworthiness—or "discursive"—based on how news production actively constructs newsworthiness—are excluded. Editors and viewers are also clearly influenced by, even depend on, news media to drive their Wikipedia collaboration and viewing habits. However, given the wide variety of sources they may receive their news from, we do not expect that any results are biased by the decisions of a particular platform or news outlet.

In future work, one could run experiments on or surveys of editors dedicated towards answering how they depend on news media, similar to the efforts of Lemmerich et al. (2019); Singer et al. (2017) in tracing browsing behaviour and linking to survey responses to address more general issues in Wikipedia information production. There is also room for more in depth study of how news values are represented in the LexisNexis data, for example through detailed manual coding or natural language processing. However, there is little apparent information on attention based news values or any link to established knowledge structures in this data. This kind of task on (a typically narrower range of) exclusively news media is also well covered in a number of studies, as previously detailed.

5.8 Conclusion

I have explored the landscape of new values without strong dependence on a particular choice of news media, analysed how they relate to news topics, and tested Galtung and Ruge's hypotheses on complementarity and additivity (/ exclusion). In doing so, I answer Rosengren's appeal and

present the case for Wikipedia as a comprehensive, if imperfect, extra-media site of study for news events. Wikipedia can be considered to be more representative of an audience-centric perspective on events than direct news media, or other previously used data limited by, for example, survey size or news topic.

Addressing **RQ2a** I find association between news values and news content in the form of topics, with clusters of the news value data aligning with topic labels obtained through separate network content analysis in Chapter 4. In testing correlations between news values (**RQ2b**), I find the complementarity hypothesis—that news values are negatively correlated—is valid in its most general form, but does not hold consistently when controlling for news topic or on closer inspection of individual news values. I also take to the task of modelling the relationship between news values and news media coverage for events listed on Wikipedia for **RQ2c**, finding some support, if not universal, for the additivity hypothesis. There is sizeable variation in how relations between news values and news coverage apply to different news topics—which has important ramifications when considering many other studies only study a single type, or limited range of events.

I have provided an extensive analysis of news values theory based on large datasets from both news media and extra-media Wikipedia data. I advance understanding of how news is represented on Wikipedia, yet in linking this to established news media I go beyond the platform and test key hypotheses in the news values literature and how they apply to audience-centric perceptions of news values and newsworthiness. Wikipedia is not a news website, but it is a vital record of current events from which much can be learned about the effects and core theories of news media. We have seen the importance of the attention based news values in characterising events. This, together with how concentrated activity is in the build-up to and fallout from current events, motivates closer study of the rise and fall of peaks of collective attention. I go on to address this in Chapter 6.

Chapter 6

News and Collective Attention: Profiling Peaks and Predicting their Aftermath

6.1 Introduction

Through both Chapter 4 and Chapter 5 I have shown the importance of collective attention in discovering what kinds of current events topics attract interest and how it relates to the newsworthiness of events. Peaks in collective attention in particular, given the level of focussed activity at the moment of event occurrence, are emblematic of online responses to news events. A news event will act as an exogenous force to spur collective attention and information seeking behaviour towards relevant information online, including Wikipedia. This often happens rapidly, owing to the quick spread of shock news or quickly increasing anticipation as some scheduled event is about to occur. As the event fades from relevance and people's informational need is satisfied one would expect collective attention, as measured by online traces, to similarly fall away. In the previous chapter, we saw how the most newsworthy events, attracting the most news coverage, are most strongly associated with the "magnitude" news value (as measured by excess Wikipedia page view counts; a peak). Rather than slow moving changes to base levels of attention, news events—and responses to them on Wikipedia—happen in individual moments over a timescale of hours and days. The redistribution of attention and reorganisation of knowledge (both temporary and accumulative) then is highly dependent on the dynamics

of how concentrated activity emerges and evolves. Clearly peaks are highly important, even characteristic, of how individuals respond to news events on Wikipedia and beyond. But what form(s) might they take, and how might we profile, or even predict, these peaks in collective attention?

Various attempts have been made to model and predict the pattern of attention towards events or spikes in popularity in social media. Understanding how attention is attracted to and leaves a subject, to the point of being able to forecast it, is clearly highly desirable in an age where newsrooms are increasingly turning towards editorial analytics tools. As detailed in the literature review (Chapter 2), previous projects have identified and modelled different classes of peak behaviour through analysis of Twitter hashtags, YouTube videos, petitions, as well as specific classes of events and their record on Wikipedia (Crane & Sornette, 2008; Matsubara et al., 2012; Wang et al., 2013; Kwon et al., 2013; Kobayashi et al., 2021).

Different technical affordances and driving dynamics can lead to different peak profiles, within and between settings, though there is frequently a focus on sharing of content and spreading dynamics. It is not clear that these results on the form of peaks of collective attention apply to a platform (Wikipedia) and phenomenon (news events) that are largely not dependent on such iterative democratised forces. Moreover, in studying attention many online platforms there is the underlying concern that one is simply reverse engineering whatever content delivery algorithms (or even human editorial decisions) are used to deliver information to users. In contrast, the vast majority of users arrive at Wikipedia from external sources (predominantly search) (Dimitrov et al., 2018), and there is little in the way of on-site article promotion or sharing features. Wikipedia page views are a measure of intentional, low-cost attention towards a subject. A user typically consciously navigates to a page rather than having it served to them by a friend / feed algorithm / editor, and does not incur any cost in producing the content or conferring on it some public social signal (e.g. a like or retweet). This is a very much stripped down collective attention measure, in stark contrast to what information on online engagement is typically available to researchers. Of other popular web-based collective attention measures, Google

search trends data comes closest, however the information is normalised to a 0–100 range which restricts its utility. In addition, when studying attention on Wikipedia, we must also consider the basal level of page views towards articles, which may differ before and after a peak. This is both a technical and substantive issue. The limited array of articles on Wikipedia consolidates users to a specific location (and resulting attention signal) where free form social media can leave disconnected, dissipating signals (e.g. different hashtags on Twitter representing different events that involve the same people). Wikipedia can bring a permanence to the effects of events, which may be only transiently felt elsewhere. Many measures of attention on online platforms do not incorporate this base level of attention—or changes to it—towards a subject.

I build on prior work by studying peaks towards news events across several categories, aggregating activity across a variety of event-related pages, as well as considering both the build-up and decay of page views in response to the collective attention dynamics of anticipation and forgetting. I first develop a time series peak clustering algorithm, to identify the typical kinds of peaks present in the Wikipedia news reaction dataset. I review their behaviour and develop a general model for the build-up and decay of attention in response to some driving news event, testing the form and dynamics within the model that best capture the observed peak shapes. Finally, I turn towards the task of predicting attention in the aftermath of news events, utilising results from both the clustering and decay model approaches in forecasting, as well as comparing to a neural network approach and other baselines. This is encapsulated by **RQ3**.

- **RQ3:** How can we model and predict peaks of collective attention towards news events?
  - RQ3a: What characteristic shapes of peaks of collective attention arise in response to news events?
  - RQ3b: How do different dynamics affect the rise and fall of peaks of collective attention?
  - RQ3c: How well can peak models predict collective attention in the aftermath of events?

6.2 Data Processing

Each Event Reaction obtained from the community detection process in Chapter 4 consists of a community of event related Wikipedia articles with an associated time series for page views. The aggregate page view times series for a given reaction is obtained from a PageRank weighted sum of the time series of its constituent articles (see Section 5.3). Not all of these time series are necessarily a peak, so some time series processing is required. In addition, there is typically some seasonal weekly component to the data, which I remove from the time series before fitting any functions / applying any prediction algorithms. In this chapter, I use hourly rather than daily time series resolution, owing to the explicit focus on temporal patterns of collective attention.

For the purpose of seasonal decomposition I take a multiplicative approach; i.e. a raw time series $Y(t) = T(t)S(t)E(t)$, where $T(t)$ is the trend, $S(t)$ is the seasonal, and $E(t)$ is the error component. $S(t)$ is estimated by taking the periodic median of $Y(t)$ (with period = 168 hours = 1 week), $\overset{168}{\tilde{Y}}(i) = \tilde{Y}(i + 168 \times n, i + 168 \times (n + 1), \cdots)$, and dividing by the full median so $S(t) = \overset{168}{\tilde{Y}}(t)/\tilde{Y}(t)$, giving an average shape for the time series over 1 week. Finally the time series is normalised by this weekly component, yielding $Y'(t) = Y(t)/S(t)$.

I then work to identify one peak per time series associated with each event. I consider a peak in the time series ± 36 hours from midday UTC on the listed event date to be associated with the event (to allow for events occurring at different times of the day in different timezones). I scale the time series according to a 2 week rolling median and interquartile range (IQR), with any local maxima $> 2 \times IQR$ labelled as peaks (with minimum peak separation of 24 hours). The time series is then centred on the largest peak within the 72 hour window at time t_p (if one is present). Then, provided there are no larger peaks in the time series within ± 84 hours of t_p (likely corresponding to other events), a one week time series is returned. A schematic is shown in Figure 6.1. In total this yields 4684 peaks for analysis.

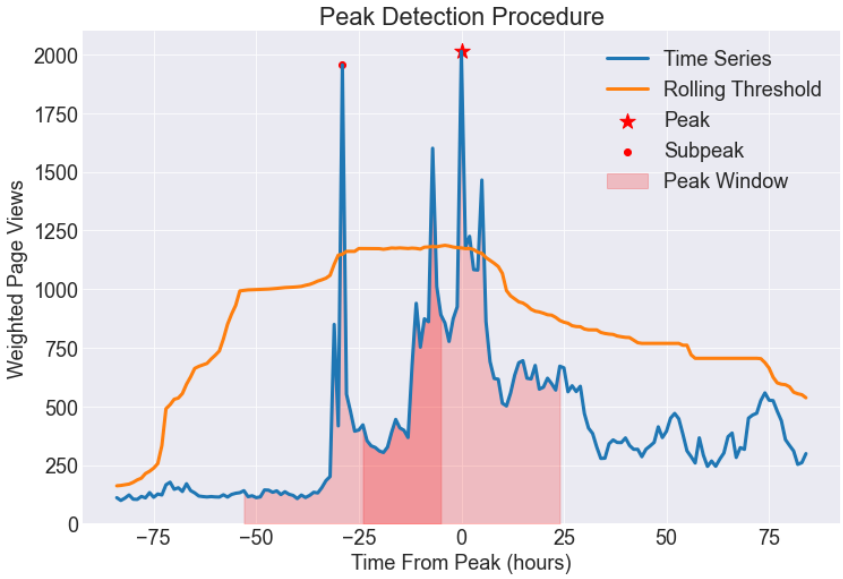

Figure 6.1: The peak detection procedure. Peaks are indicated as the largest local maxima in a 48 hour 'peak window' that exceed the rolling threshold.

6.3 Characteristic Shapes of Attention Peaks

Before developing a collective attention dynamics based model for the growth and decay of peaks, it is instructive to be able to characterise what kind of peak shapes are present in the data in a non-prescriptive, unsupervised way. This clearly aligns with the aims of **RQ3a:** *What characteristic shapes of peaks of collective attention arise in response to news events?*. To accomplish this, I introduce a variant of the K-Spectral Centroid (KSC) clustering algorithm (J. Yang & Leskovec, 2011), termed Weighted K-Spectral Centroid Clustering (WKSC). The regular KSC algorithm takes a distance measure between time series that is invariant to scaling and shifting. However, the distance between values at each time step is equally weighted when calculating the distance between two time series. In cases where there may be secondary peaks, or noise comparable in magnitude to the main peak, the clustering algorithm may in theory group together time series with similar secondary peaks or noise outside of the central peak. I choose then to introduce a window function to weight the distance metric, so as to assign more importance to timesteps closer to the central peak, meaning time series are more likely to be

clustered based on central peak shape.

In the regular KSC algorithm, for time series x and y the distance between $\hat{d}(x, y)$ is given by

$$\hat{d}(x, y) = \min_{\alpha, q} \frac{||x - \alpha y_{(q)}||}{||x||} \tag{6.1}$$

where $y_{(q)}$ is the result of shifting y by q time units and $||\cdot||$ is the l^2 norm. The measure finds the optimal translation q and scaling coefficient α. q is found heuristically by aligning the peak values of the time series and searching a small region either side. Optimal α is found by minimising the convex distance function $\hat{d}(\alpha)$, which is achieved for $\alpha = \frac{x^T y_{(q)}}{||y_{(q)}||^2}$. See Appendix D.1 for more detail and derivation.

In the WKSC variant I set distance between time series as

$$\hat{d}(x, y) = \min_{\alpha, q} \frac{||x - \alpha y_{(q)}||_w}{||x||_w} \tag{6.2}$$

where $||\cdot||_w$ is the norm where elements are weighted by a window w. For a time series expressed by column vector z and window matrix specified by $W = diag(w)$

$$||z||_w = (z^T W z)^{0.5}. \tag{6.3}$$

Optimal q is identified heuristically as previous, whereas optimal α is similarly obtained for minimum $\hat{d}$ with $\alpha = \frac{x^T W y_{(q)}}{||y_{(q)}||_w^2}$. With a flat window (i.e. $W = I$), this clearly reduces to the results for the standard KSC algorithm distance. WKSC then proceeds as KSC; finding optimal cluster centroids μ_k with cluster assignments C_k so as to minimise the function $F = \sum_k \sum_{x_i \in C_k} \hat{d}(x_i, \mu_k)^2$. Full derivations, testing, and model selection in Appendix D.1.

6.3.1 WKSC Peak Clusters

Model selection (Appendix D.1) with the Hamming window WKSC clustering algorithm points towards meaningful separation in a 5 cluster model. The adjusted mutal information (AMI) with the equivalent 5 cluster model for regular KSC is 0.65, indicating some expected similarity, but nevertheless a different clustering tendency. The centroids for the WKSC clusters are plotted in Figure 6.2.

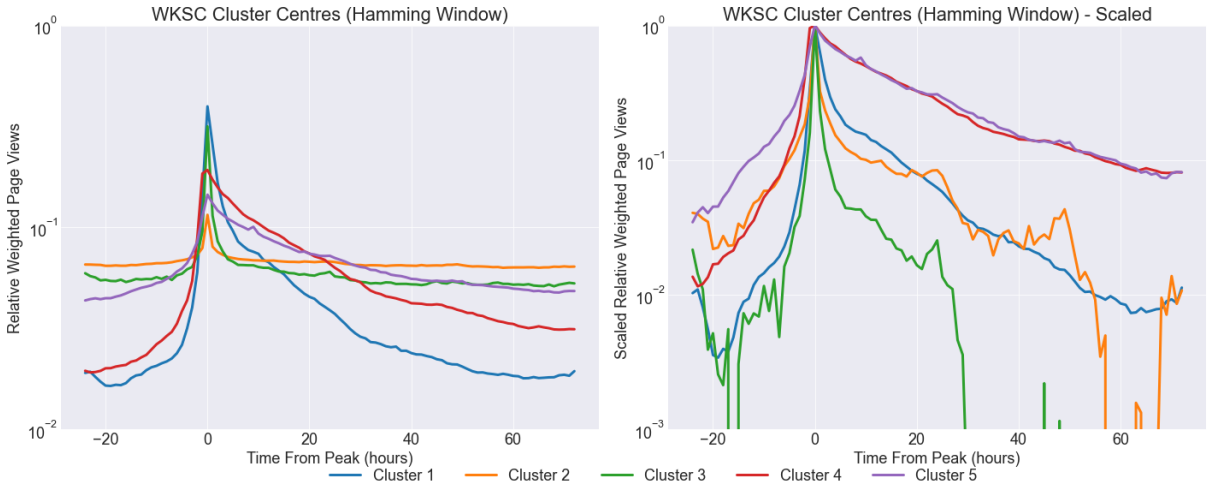

Figure 6.2: WKSC cluster centres obtained with a Hamming window at $n = 5$. The scaled variants are centred to the median of the time series and rescaled to 1 at $t = 0$.

The clusters may be described as:

- Cluster 1, "Shockwaves": Very high relative magnitude peak, sudden rise, medium-term decay.
- Cluster 2, "Blips": Low relative peak magnitude, medium-term rise and decay.
- Cluster 3, "Spikes": Medium relative peak magnitude, sudden rise and decay.
- Cluster 4, "Waves": High relative peak magnitude, medium-term rise, slow decay
- Cluster 5, "Swells": Medium relative peak magnitude, slow rise and decay

There also is close alignment between the different speed rise and decay patterns in the different clusters in the scaled time series (Figure 6.2).

One can consider why these different peak shapes emerge. Different types of events are drivers of different forces in shaping audience response. For example, the degree of foreknowledge and anticipation of the event, event timescale itself (e.g. a 2 hour marathon race vs a full day legal trial), the overall importance and popularity of the subject matter, and the lasting impact and ramifications for the event. An additional specific consideration for this thesis is how suited

Wikipedia is as an information resource for the particular kind of event. A football match might drive short-term communication and entertainment forces that result in intense interest in the moment but quickly fade as the audience moves its focus onto the next match. This is very different from the critical discussion and follow-up responses (from officials and journalists) that might come in the wake of a political scandal. In some ways, explaining the characteristic peak shapes finds close similarity with the exploration of news values in Chapter 5, most clearly with the Prominence, Surprise, and Magnitude news values. However, beyond simply explaining the total newsworthiness of an event, here we more carefully consider how an event might draw attention over time.

Given the different kinds of forces driving peak shapes, we might expect that the peak shapes are associated with news category. Events on the Wikipedia current events portal are recorded under 10 specific news categories (Armed conflicts and attacks, Arts and culture, Business and Economy, Disasters and Accidents, Health and Medicine, International Relations, Law and crime, Politics and Elections, Science and Technology, Sports). The confusion matrix is shown in Table 6.1 together with percentage over/underrepresentation in Table 6.2, and overall AMI = 0.04, $\chi^2 = 715$, $p < 0.001$. On the whole though, peaks of attention towards news events rapidly rise towards sharp peaks (over < 24 hours), and are followed by a wider range of decay timescales. There are some strong individual associations between peak shape and news category (e.g. Sports & Cluster 1), but no consistent peak expression by category. Different kinds of events within a category can produce different peak dynamics.

6.4 Fitting Peaks of Collective Attention

In this section I present a model for the dynamics of peaks of collective attention. Here I am interested in answering **RQ3b:** *How do different dynamics affect the rise and fall of peaks of collective attention?* Variations in setting, experimental design, and driving phenomena have resulted in a range of models being developed in previous studies (Watt et al., 1993; Crane & Sornette, 2008; Matsubara et al., 2012; Kwon et al., 2013; Wang et al., 2013; Kobayashi

Table 6.1: Observed incidence of news category and peak class combinations.

	Shockwaves (1)	Blips (2)	Spikes (3)	Waves (4)	Swells (5)
Armed Conflicts And Attacks	17	400	32	62	116
Arts And Culture	39	85	14	46	56
Business And Economy	9	143	15	53	67
Disasters And Accidents	29	363	39	67	99
Health And Environment	1	20	5	1	10
International Relations	21	429	57	48	137
Law And Crime	45	397	47	149	160
Politics And Elections	116	313	41	151	209
Science And Technology	11	115	18	34	60
Sports	128	82	39	48	41

Table 6.2: Percentage over/under-representation of news category and peak class combination. Defined as $(Observed - Expected)/Expected$, where $Expected = N_{\text{category}}N_{\text{cluster}}/N_{\text{total}}$.

	Shockwaves (1)	Blips (2)	Spikes (3)	Waves (4)	Swells (5)
Armed Conflicts And Attacks	-69.5	27.3	-22.1	-29.7	-9.3
Arts And Culture	83.0	-29.3	-11.0	36.2	14.4
Business And Economy	-64.7	-0.6	-20.3	31.3	14.5
Disasters And Accidents	-45.3	21.3	-0.3	-20.2	-18.7
Health And Environment	-69.6	7.9	106.2	-80.8	32.6
International Relations	-65.8	23.7	25.7	-50.7	-2.9
Law And Crime	-36.5	-0.7	-10.1	32.7	-1.7
Politics And Elections	57.4	-24.7	-24.6	29.3	23.5
Science And Technology	-48.0	-3.6	15.4	1.5	23.6
Sports	326.4	-51.6	76.0	0.9	-40.5

Table 6.3: Simple model fit Mean Squared Error (MSE).

	Pre-peak	Post-peak	Total
Exponential	0.0041	0.0046	0.0044
Power law	0.0051	0.0061	0.0056
SIR	-	-	0.0065

et al., 2021). The majority of these consider some variation on an exponential or power law rise and/or decay, through some epidemiologically inspired Susceptible-Infected-Recovered (SIR) style model (Kermack & McKendrick, 1927), or Hawkes process self excitation model (Hawkes, 1971). These lead to exponential or power law based curves for the rise and fall of attention.

As a first step, I fit simple exponential, power law, and SIR curves to the peaks in the Wikipedia dataset. The results in Table 6.3 indicate the attention curves in the Wikipedia dataset most closely follow exponential functions for their rise and fall, which I pursue with a more general model in the following section. Deviations from these simple fit models are typically the result of overly sharp peaks (very large values very close to t_p), as noted in the exponential approach of Kobayashi et al. (2021).

6.4.1 The Impulse Decay Chain Model

The model is first presented in its most general variant, and takes into account up to three different stages of attention: short-term attention (A_1), medium-term attention (A_2), and long-term attention (A_3), governed by the following differential equations.

$$\frac{dA_1}{dt} = \Lambda_1^{\text{in}}(t)A_1 - \lambda_1^{\text{out}}A_1 - \lambda_1^{\text{T}}A_1 \tag{6.4}$$

$$\frac{dA_2}{dt} = \Lambda_2^{\text{in}}(t)A_2 + \lambda_1^{\text{T}}A_1 - \lambda_2^{\text{out}}A_2 - \lambda_2^{\text{T}}A_2 \tag{6.5}$$

$$\frac{dA_3}{dt} = \lambda_2^{\text{T}}A_2 \tag{6.6}$$

Short and medium term attention are governed by the growth, dissipation, and transfer of attention (the decay *chain* element). Built into the model is the assumption that any rises or

falls long term attention occur over a much longer time span than the period of study for any individual news shock (i.e. $\lambda_3^{\text{in/out}} \approx 0$). The presence of medium term attention in the model allows not only for different growth and decay timescales around any short term spikes, but also for temporary elevated attention levels, for example if there is an ongoing multi-day event (such as a sports tournament) that finishes with a final peak.

The form of $\Lambda_i^{\text{in}}(t)$ represents how attention is attracted towards the news event. In a standard SIR-type model, this takes the form $\lambda_i^{\text{in}} S/N$, where $\lambda_i^{\text{in}}, S, N$ are the intrinsic attractiveness ("infection rate"), the susceptible number of 'agents' (free attention), and total agents (total attention), respectively. Instead, in the Impulse Decay Chain model, we exist in the regime where $S \approx N \gg \sum A_i$, and $\lambda_i^{\text{in}} = \lambda_i^{\text{in}}(t)$ i.e. it is not social saturation that limits the growth of attention towards a new story, but a change in the intrinsic attractiveness of an event that limits the growth in attention towards the subject. As a default, we take

$$\Lambda_i^{\text{in}}(t) A_i = \lambda_i^{\text{in}} H(-t) A_i(t) \tag{6.7}$$

Where $H(-t)$ is the Heaviside step function, which is 1 for negative t and 0 for positive t.

Analytic solutions for A_1, A_2, A_3 are then obtainable through Laplace transform of the differential equations and are provided in Appendix D.5. Through the integration constants we may also specify the variables A_1^0, A_2^0, A_3^0, the values of the attention stages at the peak at $t = 0$. To enforce physically appropriate solutions ($A_i \geq 0 \ \forall \ t$, with each parameter ≥ 0) in the curve fitting process, we take an alternative representation of the constants based on the parameters A_1^0, $A_2'^0 = A_2^0 - a/b$, and $A_3^- = A_3(-84)$.

Taking different combinations of parameters to vary and fixing others at zero leads to an array of sub-models. At the simplest level we obtain a simple exponential rise and fall with constant (using parameters $\{\lambda_1^{\text{in}}, \lambda_1^{\text{out}}, A_1^0, A_3^-\}$. Others lead to more complex behaviour, particularly when incorporating the attention transfer dynamics. I fit the peaks in the Wikipedia dataset to evaluate which of the sub-models best (and most parsimoniously) captures the observed peak shapes, and which dynamics are most important. There are a total of 14 valid sub-models, where attention is able to both enter and leave the short and medium term attention classes,

and at least one of λ_1^{out}, $\lambda_2^{\text{out}} > 0$ (i.e. any rise in total attention towards a subject may also fall).

6.4.2 Curve Fitting

The curves are fit from simple, to progressively more complex models using the 'Constrained Trust Region' procedure in the Python scipy.optimize library (Virtanen et al., 2020), with the objective of minimising the mean squared error (MSE). Initial parameter guesses for models 3-14 are set using the parameters from the previous best performing sub-model that uses a subset of the target model parameters. I also impose 3 additional linear constraints: That growth of short term attention is faster than its decay (without which we would not observe a peak)

$$\lambda_1^{\text{in}} > \lambda_1^{\text{out}} + \lambda_1^{\text{T}} \tag{6.8}$$

and that the total inherent growth and decay of short term attention is faster than that of medium term attention;

$$\lambda_1^{\text{in}} > \lambda_2^{\text{in}}, \tag{6.9}$$

$$\lambda_1^{\text{out}} + \lambda_1^{\text{T}} > \lambda_2^{\text{out}} + \lambda_2^{\text{T}}. \tag{6.10}$$

Normalising the time series to peak value = 1, and re-evaluating the fit MSEs, to allow comparison between time series of different scale, gives the aggregated results in Table 6.4. Progressively more complex models are indeed generally more accurate, and remain so when accounting for the additional parameters using R^2_{adj}.

6.4.3 Attention Distribution

Considering the fits for the full IDC model (sub-model 14). I compare the total fraction of short, medium, and long term attention over the course of each event, and at the peak in Figures 6.3 and 6.4 according to $F_i = \frac{A_i(t)}{\sum_j A_j(t)}$. Comparing these figures; typically, the total attention at the peak value ($t = 0$) is focussed towards short term attention (mean values $\bar{F}_1(0) = 0.46$, $\bar{F}_2(0) = 0.24$, $\bar{F}_3(0) = 0.30$), yet over the 1 week window, the total fraction of attention in each

Table 6.4: IDC sub-model fit performance. All models use the parameters $\{\lambda_1^{\text{in}}, A_1^0, A_3^-\}$ in addition to those listed in the listed in the table. Best performance is achieved for the most general sub-model (14), and models without λ_1^{out} do not perform well.

Submodel	Additional parameters	Total # Parameters	MSE	R^2	R^2_{adj}
1	$\{\lambda_1^{\text{out}}\}$	4	0.0052	0.549	0.541
2	$\{\lambda_1^{\text{T}}, \lambda_2^{\text{out}}, A_2'^0\}$	6	0.0664	-7.271	-7.525
3	$\{\lambda_1^{\text{out}}, \lambda_2^{\text{in}}, \lambda_2^{\text{out}}, A_2'^0\}$	7	0.0046	0.584	0.568
4	$\{\lambda_1^{\text{out}}, \lambda_2^{\text{in}}, \lambda_2^{\text{T}}, A_2'^0\}$	7	0.0047	0.579	0.563
5	$\{\lambda_1^{\text{T}}, \lambda_2^{\text{out}}, \lambda_2^{\text{T}}, A_2'^0\}$	7	0.0300	-1.920	-2.028
6	$\{\lambda_1^{\text{T}}, \lambda_2^{\text{in}}, \lambda_2^{\text{out}}, A_2'^0\}$	7	0.0409	-3.736	-3.911
7	$\{\lambda_1^{\text{out}}, \lambda_1^{\text{T}}, \lambda_2^{\text{out}}, A_2'^0\}$	7	0.0047	0.578	0.562
8	$\{\lambda_1^{\text{out}}, \lambda_1^{\text{T}}, \lambda_2^{\text{T}}, A_2'^0\}$	7	0.0048	0.577	0.562
9	$\{\lambda_1^{\text{out}}, \lambda_2^{\text{in}}, \lambda_2^{\text{out}}, \lambda_2^{\text{T}}, A_2'^0\}$	8	0.0044	0.603	0.586
10	$\{\lambda_1^{\text{T}}, \lambda_2^{\text{in}}, \lambda_2^{\text{out}}, \lambda_2^{\text{T}}, A_2'^0\}$	8	0.0119	-0.205	-0.257
11	$\{\lambda_1^{\text{out}}, \lambda_1^{\text{T}}, \lambda_2^{\text{out}}, \lambda_2^{\text{T}}, A_2'^0\}$	8	0.0043	0.608	0.591
12	$\{\lambda_1^{\text{out}}, \lambda_1^{\text{T}}, \lambda_2^{\text{in}}, \lambda_2^{\text{out}}, A_2'^0\}$	8	0.0043	0.608	0.591
13	$\{\lambda_1^{\text{out}}, \lambda_1^{\text{T}}, \lambda_2^{\text{in}}, \lambda_2^{\text{T}}, A_2'^0\}$	8	0.0045	0.595	0.578
14	$\{\lambda_1^{\text{out}}, \lambda_1^{\text{T}}, \lambda_2^{\text{in}}, \lambda_2^{\text{out}}, \lambda_2^{\text{T}}, A_2'^0\}$	9	0.0039	0.636	0.618

time series is far more weighted towards medium and long term attention (mean values 0.08, 0.28, 0.64).

6.4.4 Attention Conservation

More closely examining the dynamics of attention transfer informs us on how shorter term stimuli can have longer lasting effects on the collective attention towards a topic. At each decay stage, a certain fraction of attention dissipates away from the event and the remaining fraction is conserved in passing to the next attention stage. $f_1 = \int \lambda_1^{\text{T}} A_1 dt / \int \lambda_1^{\text{in}} A_1 dt$ describes the fraction of agents attracted whose short-term interest is then captured for paying medium term attention to the event (determined by the relative strength of λ_1^{T}, λ_1^{out}). Similarly, $f_2 = \int \lambda_2^{\text{T}} A_2 dt / (\int \lambda_2^{\text{in}} A_2 + \lambda_1^{\text{T}} A_1 dt)$ describes the fraction of agents attracted whose medium-term interest is then captured for long term attention to the event. $f_T = \int \lambda_2^{\text{T}} A_2 dt / (\int \lambda_1^{\text{in}} A_1 + \lambda_2^{\text{in}} A_2 dt)$ describes the total volume of short/medium-term attention that translates to long term attention. Finally, $f_R = \int \lambda_2^{\text{T}} A_2 dt / A_3^-$ describes the fractional increase in long term attention

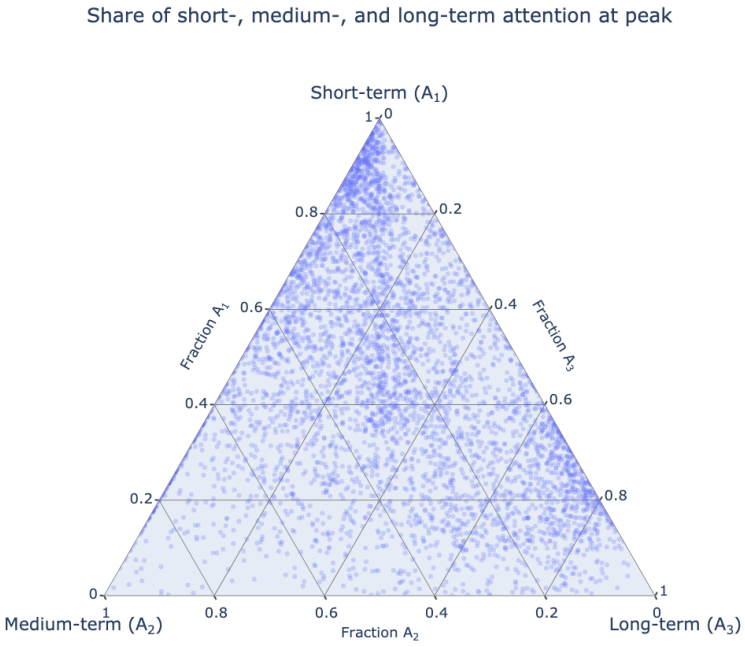

Figure 6.3: Distribution of attention at peak. Many events are short term attention heavy (A_1).

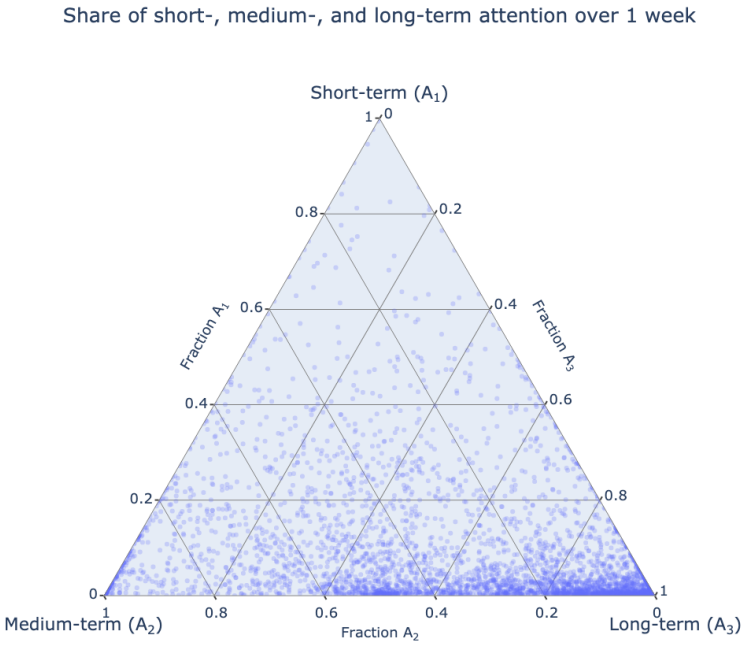

Figure 6.4: Cumulative distribution of attention over 1 week. In contrast to Figure 6.3, events are much more long-term attention dominated (A_3).

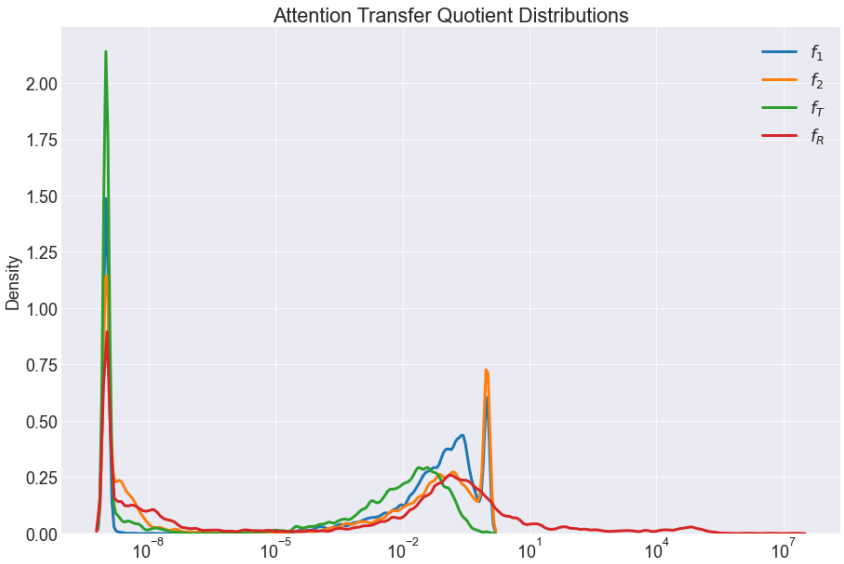

Figure 6.5: Distributions of attention transfer quotients. f_1, f_2, f_T are by definition capped at 1 and artificially clipped at 10^{-9} for plotting purposes. f_R on the other hand is able to range larger than 1, corresponding to a doubling of the long term attention towards a subject.

towards a subject. Distributions for these attention transfer quotients are given in Figure 6.5.

Typically, the capture and transfer of attention is a small fraction of the total attention given towards an event, with median values $\tilde{f_1} = 0.042$, $\tilde{f_2} = 0.011$, and $\tilde{f_T} = 0.0011$. Though this effect appears more substantial when observed from the perspective of the increase in long term attention, with $\tilde{f_R} = 0.028$. Large peaks can generate a substantial increase in attention to previously unpopular subjects, even if the vast majority of the attention within the peaks is not conserved. On average, 0.1% of the short and medium term collective attention directed towards an event is conserved, resulting in a 2.8% increase in the long term attention towards the subject. This mechanism of attention transfer and capture is the process by which individual events accumulate long term attention towards more stable established content on Wikipedia.

6.4.5 Modelling the Characteristic Attention Peaks

I fit the WKSC cluster centres obtained in Section 6.3.1 using the IDC model, with the parameters shown in Table 6.5 and curves displayed in Figure 6.6. There is close to perfect alignment between the cluster centre time series and their respective fits. A slight exception is the pre-peak

Table 6.5: Cluster centre IDC fit values.

	λ_1^{in}	λ_1^{out}	λ_1^{T}	λ_2^{in}	λ_2^{out}	λ_2^{T}	A_1^0	$A_2^{\prime 0}$	A_3^-
Shockwaves (1)	1.9087	0.6247	0.1033	0.1587	0.0606	0.0006	0.3366	0.0168	0.0146
Blips (2)	3.1394	1.5320	0.0242	0.1181	0.0430	0.0010	0.0433	0.0082	0.0628
Spikes (3)	3.8440	1.7461	0.0450	0.1906	0.0811	8×10^{-7}	0.2444	0.0135	0.0524
Waves (4)	1.2356	0.7640	0.0474	0.4572	0.0511	0.0036	0.0500	0.0300	0.0179
Swells (5)	1.2521	0.3450	0.3036	0.1250	0.0446	0.0035	0.0492	0.0293	0.0381

fit in cluster 4, where the rounded peak is not handled as well by the model. Also shown in Figure 6.6 is the proportion of short, medium, and long term attention along the time series. Cluster 1 is short and medium term attention dominated, with the largest short term volume of any cluster. The 'blips' of cluster 2 are long term attention dominated, similar to the larger spikes of Cluster 3. Cluster 4 and 5 are medium term attention dominated (Cluster 5 exhibiting a slower build-up), with a substantial increase in long term attention in both clusters. Several of the cluster centres according to the IDC model parameters apparently differ from their descriptions in Section 6.3.1 based on Figure 6.2, due to the different time series rescaling methods applied. This is most apparent when considering the 'Disasters and Accidents' category in Tables 6.1 and 6.2—events which by their nature are likely to be unexpected—that predominantly fall in Cluster 2 (Blips). The "medium term rise" description of the scaled cluster centre from Figure 6.2 is actually a rapid short term exponential growth when studying the IDC model parameters ($\lambda_1^{\text{in}} = 3.14$, $\lambda_1^{\text{in}} - \lambda_1^{\text{out}} - \lambda_1^{\text{T}} = 1.64$ from Table 6.5). Each cluster's attention transfer quotients are also shown in Table 6.6. As previously noted, there is substantial transfer to long term attention (f_T, f_R) in cluster 4 and 5. Clusters 1 and 5 also show sizeable transfer between short and medium term attention (f_1), where the short term spike feeds into a longer decay. Different timescale attention dynamics across the other centres are relatively independent. The WKSC peaks are then representative of different IDC dynamic regimes.

6.4.6 News Category Dependence

There is also substantial variation in peak dynamics by news category (taken from the record on the Wikipedia Current Events Portal), as shown in Table 6.7, with their graphical representations

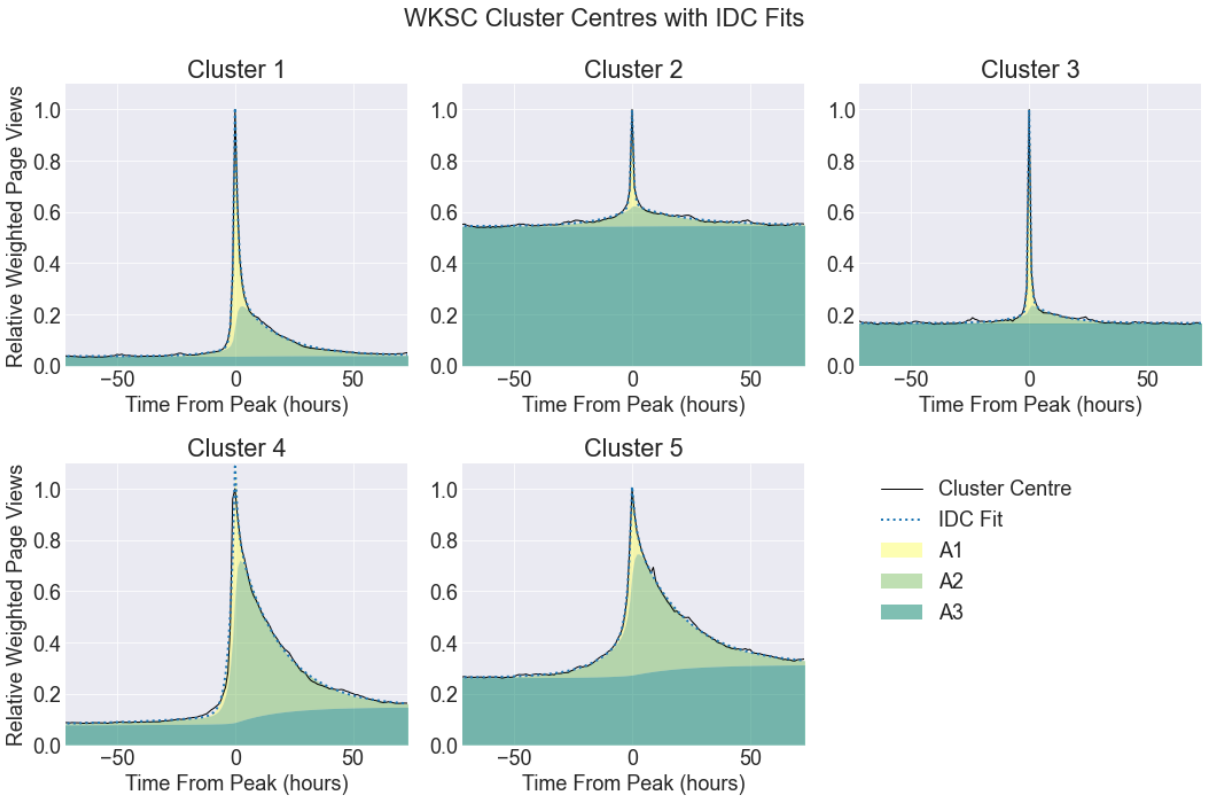

Figure 6.6: WKSC cluster centres as fit by the IDC model.

Table 6.6: Attention transfer quotients for the WKSC cluster centres.

	f_1	f_2	f_T	f_R
Shockwaves (1)	0.142	0.010	0.002	0.077
Blips (2)	0.016	0.022	0.003	0.005
Spikes (3)	0.025	9×10^{-6}	7×10^{-7}	6×10^{-6}
Waves (4)	0.058	0.066	0.035	0.590
Swells (5)	0.468	0.072	0.047	0.191

Table 6.7: Median parameter values by category

Category	$\tilde{\lambda}_1^{\text{in}}$	$\tilde{\lambda}_1^{\text{out}}$	$\tilde{\lambda}_1^{\text{T}}$	$\tilde{\lambda}_2^{\text{in}}$	$\tilde{\lambda}_2^{\text{out}}$	$\tilde{\lambda}_2^{\text{T}}$	$\tilde{A}_1^0$	$\tilde{A}_2'^0$	$\tilde{A_3^-}$
Armed Conflicts And Attacks	6.897	1.876	0.071	0.200	0.033	7.57×10^{-4}	58.7	19.1	44.6
Arts And Culture	3.610	1.049	0.088	0.174	0.041	9.58×10^{-4}	203.7	34.6	59.1
Business And Economy	4.744	0.886	0.039	0.171	0.035	1.71×10^{-3}	92.7	19.3	42.0
Disasters And Accidents	5.858	1.365	0.049	0.141	0.029	1.21×10^{-3}	99.5	22.4	41.8
Health And Environment	6.015	1.500	0.030	0.173	0.041	4.50×10^{-3}	80.6	20.7	53.3
International Relations	6.420	1.359	0.047	0.155	0.033	1.79×10^{-3}	79.7	24.3	44.9
Law And Crime	4.554	1.089	0.046	0.151	0.032	6.66×10^{-4}	104.1	18.6	39.9
Politics And Elections	3.799	0.937	0.061	0.093	0.039	3.77×10^{-4}	123.9	25.8	34.1
Science And Technology	4.411	1.405	0.048	0.105	0.031	5.70×10^{-4}	54.9	14.8	25.9
Sports	2.548	1.060	0.042	0.105	0.050	1.39×10^{-5}	442.3	51.4	44.1
Overall	5.111	1.178	0.050	0.134	0.036	7.41×10^{-4}	99.2	22.8	41.2

in Figure 6.7. The Sports & Arts and Culture categories show the highest peaks of attention, indicating the popularity of entertainment related content on Wikipedia (Singer et al., 2017), yet their average activity level is not necessarily the highest.

The relative proportion of short, medium, and long term attention within each peak also varies by category, as indicated in Figure 6.8. From the long-term attention dominated (at peak and over total time series) Health and Environment news reactions, to Sports events, where short term attention (particularly at peak) is more prevalent.

Attention transfer quotients are shown in Table 6.8, with several orders of magnitude variation by category in some cases. For example, 1.48% of the short and medium term attention towards Health and Environment stories is transferred to a 5.25% increase in long term attention, whereas just 0.0002% of the short and medium term attention towards Sports stories is transferred to a 0.03% increase in long term attention—practically non-existent.

6.5 Predicting Attention Towards News Events

One is able to profile the dynamics for peaks of collective attention towards news events after the fact using the IDC model, and the results from the WKSC algorithm suggest there are characteristic paths of evolution for peaks. But a question remains; **RQ3c:** *How well can peak models predict collective attention in the aftermath of events?* Note that the objective here is

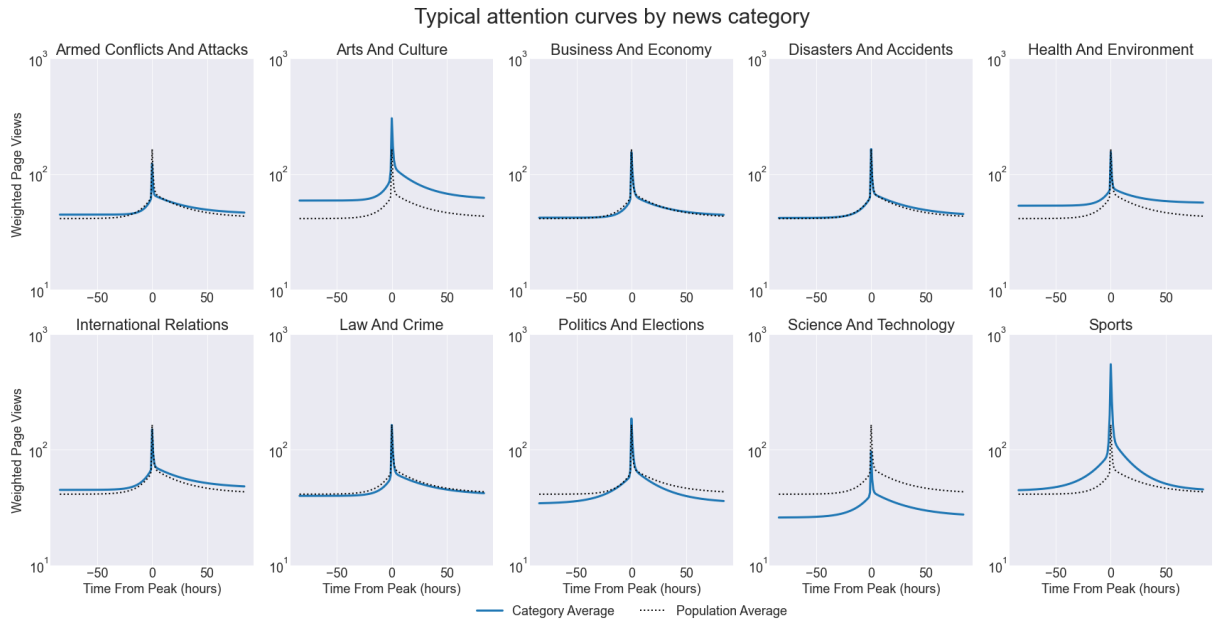

Figure 6.7: Attention curves by category, taking the median value of each fit parameter. Each category curve is also compared to the curve obtained from the median fit parameters for the whole population.

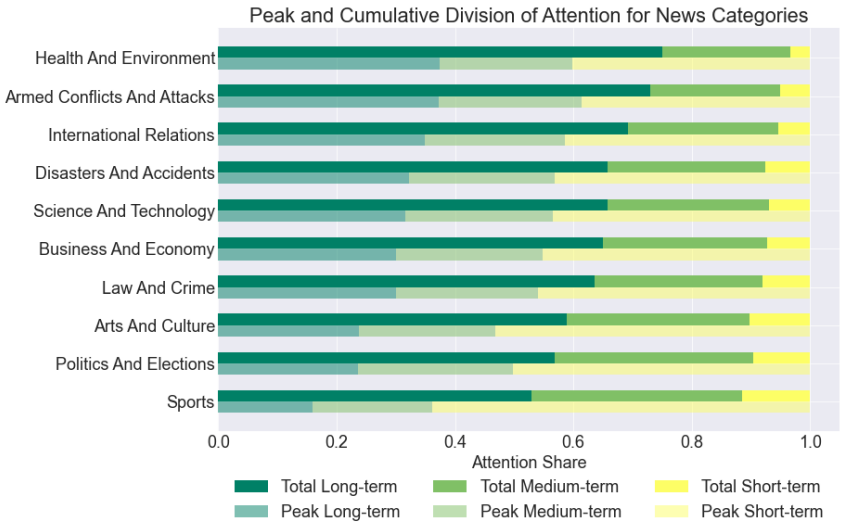

Figure 6.8: Average division of short-, medium-, and long-term attention over the total observed week and at the peak value, organised by news category.

Table 6.8: Median attention transfer quotients by event category.

Category	$\tilde{f}_1$	$\tilde{f}_2$	$\tilde{f}_T$	$\tilde{f}_R$
Armed Conflicts And Attacks	0.0358	0.0090	0.0008	0.0152
Arts And Culture	0.0771	0.0113	0.0018	0.0339
Business And Economy	0.0659	0.0354	0.0029	0.0450
Disasters And Accidents	0.0333	0.0283	0.0018	0.0387
Health And Environment	0.0279	0.0458	0.0148	0.0525
International Relations	0.0352	0.0225	0.0016	0.0353
Law And Crime	0.0356	0.0152	0.0014	0.0310
Politics And Elections	0.0629	0.0081	0.0010	0.0370
Science And Technology	0.0365	0.0056	0.0004	0.0062
Sports	0.0337	0.0001	2×10^{-6}	0.0003
Overall	0.0418	0.0119	0.0011	0.0279

not to predict whether, or when, an event will occur, but to predict the evolution of collective attention towards a subject after a given peak at time $t_p = 0$. I adapt both the IDC model and WKSC algorithm towards this task. The IDC model is fit on values for each time series from $t_- = -84$ up to $t_{\text{obs}} \geq t_p$ to predict page views for t_{obs} to $t_+ = 84$. For the WKSC approach, I take the cluster (from those already obtained) closest in distance (equation 6.2) to the target time series (based on observation window $t_- = -84$ up to $t_{\text{obs}} \geq t_p$), scaled according to the peak value of the target time series. Implementations of the WKSC and IDC fit models do not take into account news category information and we are primarily interested in how existing page views are predictive of future attention levels. As such, category information is not used in this prediction task, but may prove a fruitful avenue for research beyond this work. The proposed forecasting approaches are compared to several baselines as well as a machine learning based model.

6.5.1 Long Short Term Memory Networks

Neural networks, specifically Long Short-Term Memory (LSTM) networks are frequently used in time series forecasting tasks (e.g. Zhao, Chen, Wu, Chen, & Liu, 2017, Siami-Namini, Tavakoli, & Namin, 2018). LSTMs are a variant on recurrent neural networks (RNNs) which use feedback along sequences of data (unlike regular feed-forward neural networks) such as time series or

sentences of words to make predictions. LSTMs use an inbuilt 'memory' in each layer which enables them to learn long term dependencies in the input data (5–10 timesteps) (Zaccone, Karim, & Menshawy, 2017), unlike regular RNNs.

In the LSTM implementation the time series ($Y(t)$) are rescaled from their raw values, so that $P(t_p) = 1$. The models take observation data from $t_- = -84$ up to a time $t_{\text{obs}} \in \{0, 3, 6, 12, 24\}$[1] and predict page views up to a time $t_+ = 84$. Training/test data was created according to a 75%/25% split. Mean squared error (MSE) was selected for loss function, with the models trained via mini-batch gradient descent. For each model, hyperparameters for layer size, number of hidden layers, number of epochs, and batch size were selected with a grid search across hyperparameter combinations, with the best combination selected according to performance in 3-fold cross-validation on the training data.

6.5.2 Baselines

Additional baselines are considered to judge the IDC model prediction against for a time series $P(t)$, listed as follows and summarised in Table 6.9.

Median

Predicted series $Q(t) = \text{median}(P(-84, t_{\text{obs}}))$.

Mirror

Predicted series $Q(t) = P(-|t|)$ (effectively mirroring the pre-peak activity).

Scaled Median Time Series

Predicted series $Q(t)$ is the median of all time series $\tilde{M}(t)$, scaled according to the median and peak values of $P(t)$.

$$Q(t) = \frac{(P(0) - \text{median}(P(-84, t_{\text{obs}})))(\tilde{M}(t) - \text{median}(\tilde{M}(t)))}{\tilde{M}(0) - \text{median}(\tilde{M}(t))} + \text{median}(P(-84, t_{\text{obs}})).$$

[1] Only 5 time windows selected due to computing limitations.

Table 6.9: A summary of the baseline predictive page view models.

Baseline	$Q(t)$
Median	$\text{med}(P(-84, t_{\text{obs}}))$
Mirror	$P(-\lvert t\rvert)$
Scaled Median Time Series	$\frac{(P(0)-\text{med}(P(-84,t_{\text{obs}})))(\tilde{M}(t)-\text{med}(\tilde{M}(t)))}{\tilde{M}(0)-(\tilde{M}(t))} + \text{med}(P(-84, t_{\text{obs}}))$
Scaled Mean Times Series	$\frac{(P(0)-\text{med}(P(-84,t_{\text{obs}})))(\bar{M}(t)-\text{med}(\bar{M}(t)))}{\bar{M}(0)-\text{med}(\bar{M}(t))} + \text{med}(P(-84, t_{\text{obs}}))$

Scaled Mean Times Series

Predicted series $Q(t)$ is the mean of all time series $\bar{M}(t)$, scaled according to the median and peak values of $P(t)$.

$$Q(t) = \frac{(P(0)-\text{median}(P(-84,t_{\text{obs}})))(\bar{M}(t)-\text{median}(\bar{M}(t)))}{\bar{M}(0)-\text{median}(\bar{M}(t))} + \text{median}(P(-84, t_{\text{obs}})).$$

6.5.3 Performance

Example predictions are shown in Figure 6.9 and mean model performance is given in Figure 6.10, with the IDC model predictions comparing favourably to the baselines and performing close to the LSTM model (which has many times more parameters). The WKSC model predictions, whilst not as strong as those from IDC or LSTM, still maintain a sizeable advantage over the baselines. From starting with no post-peak information at $t_{\text{obs}} = 0$ the IDC and LSTM models quickly improve in performance up to $t_{\text{obs}} \approx 3$, before more gradual improvement up to $t_{\text{obs}} = 24$. The predictive power of the IDC model further justifies its use in modelling the full peak time series.

6.6 Discussion

The WKSC algorithm effectively clusters the observed peaks of collective attention, invariant to scaling or shifting of time series, under the assertion that different timesteps are more important in evaluating time series similarity. In practice, we focus more on timesteps close to a central peak, as compared the flat weighting of the original KSC paper. In effect, when taking into account the limited length of the time series, the original KSC paper takes a hat function

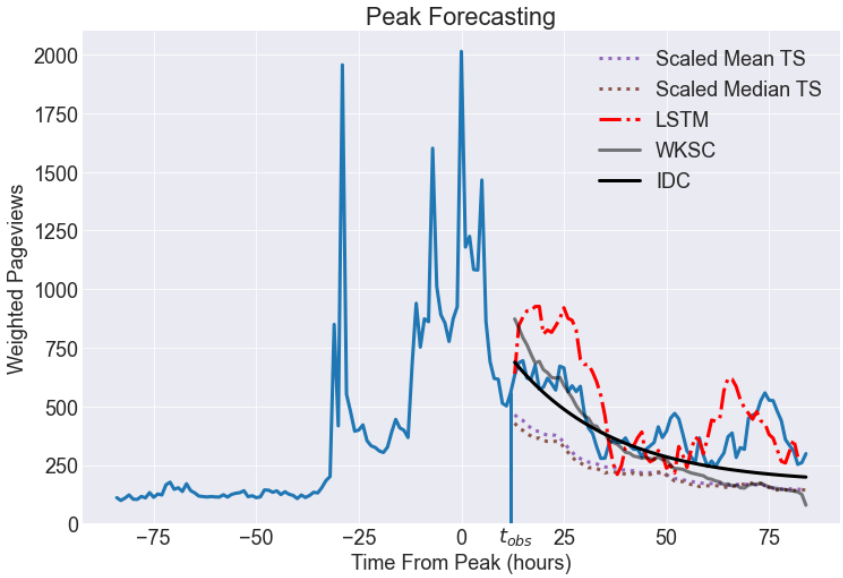

Figure 6.9: Example forecasts for selected models with $t_{\text{obs}} = 12$. The IDC and LSTM models best take into account the increased baseline page views post-peak.

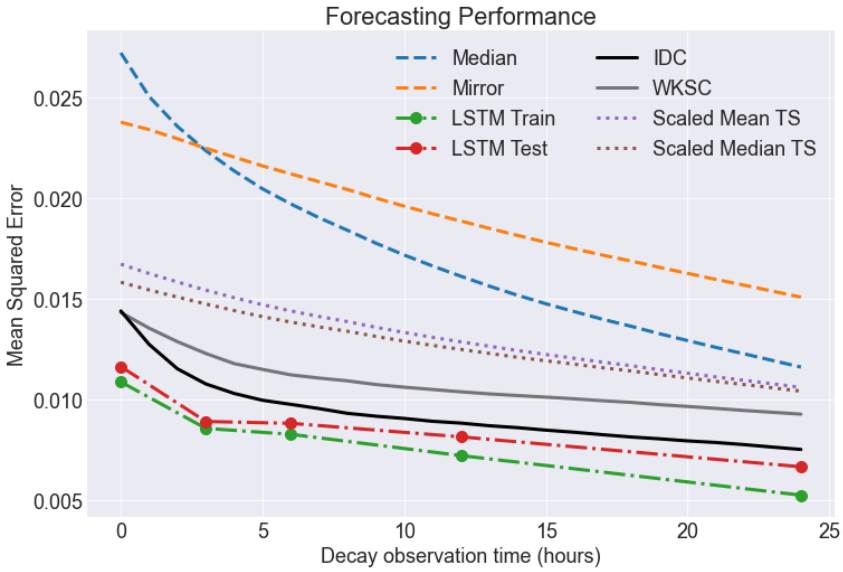

Figure 6.10: Model forecasting performance with increasing observation window. In the IDC model, at $t_{\text{obs}} = 0$ the decay parameters $\lambda_1^{\text{out}}, \lambda_1^{\text{T}}, \lambda_2^{\text{out}}, \lambda_2^{\text{T}}$ are fixed at the population median (since there is no post-peak data to fit them on).

window; equally weighting timesteps inside the window, and discounting (i.e. zero weighting) timesteps outside. Careful theoretically informed choices can inform the window size, but the large discontinuity at the boundaries is not very satisfying. The choice to then progressively weight timesteps closer to the peak more highly is desirable. Of course, having opened the box to allow any window form, this prompts the question of what window performs best? It is not clear what performance measure would be appropriate here, perhaps comparing clustering performance with different windows to some ground truth (not applicable in the case of this work), or exploring how window choice affects cluster separation. There may be yet more subject specific performance measures with which to judge window form, or perhaps automation of window shape is possible with a variant of one of the approaches in various weighted K-means clustering works (Kerdprasop, Kerdprasop, & Sattayatham, 2005; De Amorim & Mirkin, 2012). Going even further, in principle one might be able to allow for nonzero off diagonal elements in the matrix W to explore timestep codependencies—though a reformulation of the algorithm would be required.

The IDC model performs well in fitting the range of observed peaks with the dynamics of growth, dissipation and transfer of attention. Key to the IDC model is the time dependence of the intrinsic attractiveness of an event, rather than the popularity of an event being limited by spreading dynamics and the carrying capacity of a population. In some ways this is not surprising. Newsworthiness, and the attention an event attracts, is time and context sensitive, with several authors going so far as considering timeliness a 'news value' (Schultz, 2007) (see Chapter 5 for greater exploration on newsworthiness and news values). Viral online phenomena, which form much of the research base on collective attention, are less constrained by this. Of course, some news items are about viral phenomena, or may be somewhat viral themselves (and are perhaps captured in the slight sigmoid nature of WKSC cluster 4). However, the vast majority of news reactions are dependent on time sensitive few-to-many broadcast (be it through TV, radio, online) from media sources, rather than spreading dynamics. This observation perhaps has relevance for fake news and misinformation scholars. One expects that fake news items,

given the greater detachment from some objectively true event, are less time sensitive and rely more on occasional viral success through sharing mechanisms. We might then anticipate that the growth in collective attention patterns towards fake news events to be more likely to be sigmoid in form.

Collective attention is dependent the output of time sensitive broadcast media and its audience. The Wikipedia page viewers are not all subject to the exact same synchronous news broadcasts, yet the rapid rise and fall close to the peak (followed by a longer timescale decay) shows the disproportionate level of attention a 'breaking' news story attracts in the moment. Old news, even by a matter of hours, is no longer as interesting, and the audience is acutely aware of this. The observed decay of medium term attention does extend the lifetime of excess attention towards an event—though not to the scale that might be observed by a longer tail power law decay.

The peak shapes are well captured by the single IDC model, but the expressed dynamics are not universal. We observe a variety of peak shapes obtained from the WKSC algorithm which correspond to differently expressed combinations of λ_i parameters in the IDC model. Similarly, we have seen wide variation by news event category. Events in different categories exhibit differing collective attention dynamics during a peak (though not as stark a difference as the directly clustered WKSC time series), but also can have different lasting effect on the attention towards their subject. These changes in base levels of attention towards a subject are frequently not captured when modelling the peaks in collective attention towards often ethereal phenomena on the web.

Finally, journalism is increasingly turning towards automated editorial analytics, and knowing the current and future levels of attention towards an event, in order to best attract readers, is clearly very useful to this practice. This being done in a setting independent from any particular news outlet is particularly appealing, and has strong parallels with 'altmetrics', acting as a complement to more direct measures a news organisation may use such as views or read time on their own website. The WKSC method, in simply forecasting one of the 5 cluster centres,

performs above baseline measures. More accurate, fine tuned predictions are achieved with the IDC fits on the observed time series, which have comparable errors to a much more complex LSTM approach, with the added benefit of being theoretically informed and interpretable, as demonstrated by the exploration of news category dependencies and the dynamics of attention transfer.

6.7 Conclusion

The highly concentrated nature of attention towards current events means it is vital to focus research on the rise and fall of peaks of collective attention that emerge. Understanding how the various types of peaks develop is a valuable avenue for research and content delivery services alike, especially when considering Wikipedia is an independent data source largely free of the editorial and algorithmic newsfeed forces that shape traffic on news and social media sites.

I have presented the Weighted K-Spectral Centroid clustering algorithm as an extension of KSC clustering. WKSC, with appropriate window selection, can identify the characteristic time series shapes and avoid spurious clustering around features at timesteps deemed less relevant. I identify 5 clusters in the Wikipedia events peaks data corresponding to different regimes of collective attention evolution, thus addressing **RQ3a**. I move onto profiling the evolution of these peaks using the Impulse Decay Chain model to answer **RQ3b**. The model takes into account short, medium, and long term stages of attention, and considers the growth, dissipation, and transfer of attention in response to external stimuli. Of all the sub-models, the full parameter set is most effective at fitting the peaks in the data. Using the IDC model we also gain insight from the distribution of attention and rate of attention transfer. The IDC fits for the WKSC centres correspond to qualitatively different IDC model specifications (e.g. fast/slow decay, high/low attention transfer). Moreover, there is substantial variation in IDC peak dynamics by news event category. Finally, I take to the task of predicting the evolution of peaks for **RQ3c**. Forecasts from the nearest neighbour WKSC centres perform above baselines, and more fine tuned least squared fits from the IDC model achieve performance close to a much more complex

LSTM approach, further justifying its application in modelling the full time series.

There remains room to further advance the ideas introduced here. With respect to the WKSC peak clustering, at present the window matrix is chosen somewhat arbitrarily, though future work could perhaps automate timestep weight assignment, in a similar way to some weighted K-means clustering variants. For the IDC model, we choose not to follow the SIR-style population limited growth, instead having the intrinsic attractiveness vary in time. However, we only take a single form of this with the Heaviside step function. Different shape impulses will surely exist, which the step function appears to approximate well enough, yet we have made no attempt to model or derive what form these impulses may take. This is surely a crucial step in further work, given this first stage of detaching the peak dynamics from any population carrying capacity. Predicting pre-peak dynamics remains out of the scope of this work, and given the typically rapid rise in attention for each peak, is likely more a task of predicting whether an event will happen at all. This must surely rely on particular event domain expertise and features. Nevertheless, this work makes material contributions towards literature on time series clustering and the modelling and forecasting of the collective attention towards news events.

Chapter 7

Conclusion

Through the chapters of this thesis I have explored the representation of content in topics, studied the way Wikipedia can develop understandings of news values and newsworthiness in wider news media, and modelled and predicted wider user patterns of attention. This has been driven by three overarching research questions:

- **RQ1:** How are current events represented in the knowledge structures and access patterns of Wikipedia and its users?
- **RQ2:** How are traditional conceptions of news values and newsworthiness of events reflected in extra-media data?
- **RQ3:** How can we model and predict peaks of collective attention towards news events?

The studies in this thesis begin with an attempt at sampling collective responses to events and detecting the topics that link them, working towards **RQ1**. This is an important foundation to the further work in the thesis but there is also much to be learnt simply from the content and basic properties of the reactions and topics, as well as the process by which they are generated. The community detection process employed takes into account both long term knowledge structures as well as short term attention dynamics. As such, it is sensitive to the duality of Wikipedia; its existence as an encyclopaedia for representing long settled information, compared to its frequent use as a news backgrounder. Communities reflecting both of these facets of

Wikipedia are thus identified, and incorporate generality across topics, usage of the knowledge network rather than focus on individual articles, explicit relation to news events, short and long term effects, and lack of reliance on detection through particular attention dynamics.

The Topics of Attention that are consequently revealed by the process are wide ranging in nature, from stable background topics akin to encyclopaedic schema to more volatile news topics. This is captured by the structural similarity score. Wikipedia may never be truly objective or representative, but it can act as an important 'standard candle' for other knowledge representations which are likely formed by smaller communities of production. Future projects could, for example, map topics from different news media sources and compare these agenda against that which is represented by the collective efforts of editors and users on Wikipedia. Developing these separate topic modelling approaches—and crucially working out how to map them against the topics obtained from Wikipedia—is outside the scope of the thesis but poses intriguing questions on how Wikipedia represents various news agenda.

Even without this comparison to other classification systems or news agenda we already note interesting features and biases of the mapped out Topics of Attention. Organisation and access of knowledge according to geography is prevalent, which makes it all the more striking (if perhaps unsurprising) that detailed coverage of Western, particularly US-based, topics dominate the events recorded and article structures that represent them. Another characteristic of interest is the prevalence of powerful individuals as topics and within topics. This ties in with the news values of 'celebrity' and 'power elite', implemented in greater depth in Chapter 5. Audiences take great interest in stories when there is a human to relate to or revile at the centre of it. Despite the biases towards particular countries and individuals, some cross-cultural topics remain represented and of interest to the audience on Wikipedia, e.g., disaster related topics such as hurricanes, earthquakes, and aircraft crashes.

The tension between Wikipedia the traditional encyclopaedia and Wikipedia the real-time current events record is most evident when we observe the existence of 'breakout' topics. In these instances, there are two or more labelled topics representing the same broad concept.

However, one of these has a high structural similarity score whereas others have low structural similarity. This means that that there is variation in how much different news events evoke novel information and connections between concepts in the audience. The Event Reactions in the lower structural similarity topic are more dependent on dynamics of collective attention and the topic does not align as well with how the 'slow' knowledge is structured on Wikipedia.

In Chapter 5, addressing **RQ2**, I paid more explicit attention to the characteristics of news events reported on Wikipedia. Yet it is still reliant on Wikipedia being more than just a news site. News value theory is well developed, but the reliance of much of the literature on news media itself is of concern. In addition, projects have frequently been focussed and narrow in scope, leaving blindspots to variation by topic. Wikipedia has proved well placed to address these issues through this project as an extra-media data source. It, along with other alternative news representations, should be sought out as a form of 'altmetrics' for news media.

I make contributions to technical formulations of the complementarity hypothesis—that news values are negatively correlated—making clear what combinations of news values should be considered for correlation in either the weak, intermediate, or strong complementarity hypotheses, and how events across different topics should be treated. The most general form of the complementarity hypothesis holds true on the Wikipedia data, but more strict versions are only partially supported, and we encounter a Simpson's paradox type effect when one controls for news topic. In future work, formal definitions for the complementarity hypothesis being employed and how it might be fulfilled should be clearly articulated, and any generalisation beyond the topic(s) being studied should be carefully considered. Comparing the Wikipedia driven news values to actual news media data; I find support for the additivity hypothesis, that news values are contribute towards the newsworthiness of an event and the chance of it being selected for news coverage. Again however, there is variation in its application to different topics.

This study is focussed on 'material' news values as intrinsic properties of events. Due to the nature of the extra-media data used, I am not able to comment on news values that emerge in the journalistic editing or production processes. It is not clear what independent extra-media data

could be used for study of these kinds of news values. Work on agenda setting and comparison between both the content and values that instruct the reporting in different news outlets is perhaps more fitting here.

In the final substantive chapter I more closely consider at the dynamics of attention around current events and on what drives peaks in collective attention, the focus of **RQ3**. I develop the Weighted K-Spectral Centroid clustering algorithm and identify 5 distinct peak shapes of collective attention in response to current events. These peaks shapes are then well captured by the proposed Impulse Decay Chain model and both methods are used towards predicting the decay of attention.

With regard to methodological developments, the WKSC algorithm is a forward step for time series clustering taking into account the most relevant timesteps close to the peak. However, the form that the window function takes is somewhat arbitrarily chosen. A question remains for future work on how best to select or automate this window. Any developments on this would require a process sensitive to the data in question, with some target function to optimise. A simple brute force approach could take a clustering performance measure (such as silhouette score) as the objective, and subject the window to a particular form that is tunable by a specific parameter (or set of parameters). The clustering can be rerun over a range of window parameter values with the final window (and subsequent cluster assignments) selected for based on the clustering performance measure. This approach is fairly resource intensive, and in truth it only slightly sidesteps the question about window form since it is still somewhat prescriptive to an overall shape.

The performance of the impulse decay chain model poses important questions for research on attention towards news events. Chapter 6's setup tests for several submodels, yet the full parameter set performs best, even when adjusting for the number of parameters. One might have hoped for the sake of simplicity and interpretability that a submodel with fewer parameters emerged as the leading candidate. Regardless, the dynamics in this system of a time sensitive attractiveness function and exponential decay in popularity have not frequently been encountered

in other settings. I argue that rather than discovering some new mode of attention special to humans doing follow up research on news events, that this is a result of the phenomenon (news) and online setting (Wikipedia) being studied. News events, in contrast to more frequently studied viral/popularity based phenomena, are time sensitive in relevance to their audience and Wikipedia on the whole is not a platform that pulls users towards particular content.

To finish Chapter 6, I took to the task of attention prediction, adapting both the WKSC and IDC approaches. The predictions from the IDC fits are close in performance to a greatly more complex LSTM neural network approach, which would indicate the IDC fits extract nearly all of the predictive information from the observation window. Whilst perhaps not as desirable as attention prediction pre-event or pre-peak (which in this case looks to be a difficult task with peaks typically emerging only over a handful of hours), predicting the legacy of events is still an important task, especially given the wide range of behaviours present (as characterised by the WKSC centres). Knowing this kind of information can allow platforms (including but not limited to Wikipedia), newsrooms, and even the individuals involved to best adapt their outputs and actions in response to the unfolding events.

There are several common themes that run through the studies in this work. Firstly, there is an explicit intention to consider the online reaction towards events over collections of, rather than individual, Wikipedia articles. Selecting an individual digital item as an indicator for attention towards a wider issue is not necessarily a representative or resilient approach. This motivates the community detection process in Chapter 4 and the further use of the Event Reactions in the thesis. I benefit from Wikipedia being a closed, linked system, rather than through detecting various relevant open-ended user generated content on social media.

Whilst Chapter 5 is explicitly focussed on news value theory, chapters 4 and 6 also incorporate it in the design of the topic sorting features and some news values also emerge from the specification of the IDC model (e.g. magnitude and surprise represented by the peak size and timescale parameters). It is critical that these findings are connected to the theoretical core of the thesis, and are part of the essential bridge building in interdisciplinary computational social

science work. The news value concepts can be formalised many different ways, but measures like page views, watch time, shares are relatively meaningless beyond single settings without appeal to a common theoretical grounding.

There is a rich tradition of Wikipedia research fostered by its openly accessible data. Yet I believe greater credence should be given to the site as a social barometer the way much of the rest of social media is treated. The website beats out Twitter, Reddit, Instagram, and Sina Weibo in worldwide Alexa rankings (*Alexa Top Sites*, 2021)[1], and the users are concentrated over a much smaller range of content—everyone sees the same Wikipedia (by language). The idea that Wikipedia is socially produced, accessed, and tacitly endorsed is often forgotten due to the lack of visible individual user profiles and voices.

Wikipedia does have its flaws. My work does not set out to discover or characterise inequalities and biases present but they still rear their head in the emergent news topics of Chapter 5. Plenty of the work uncovering the inadequacies and biases of the platform is only possible because of the wealth of accessible Wikipedia data, something for-profit corporations are far less likely to freely provide if it casts them in a bad light. Indeed some defences of Wikipedia's flaws revolve around it simply being a reflection of the world it represents, which despite not telling the whole story, strengthens my argument for its relevance as a wider social indicator. In instances where this is not the case, there is a large body of work that can fully contextualise how exactly it doesn't measure up, both in in a descriptive and normative sense, which is much more difficult with other platforms.

A final pillar of this work is on relying on Wikipedia as an 'independent' data source. This underpins the notion of Wikipedia as extra-media data in Chapter 5, but also informs the approaches of chapters 4 and 6. Much of this is principled on the fact that the majority of users do not encounter Wikipedia content pushed to them by some common editorial or algorithmic decision on the site, but by direct information seeking behaviour from web search results. Of course Wikipedia is not totally immune from these effects. A small portion of traffic to articles

[1]Mobile app use is not counted here, but the monthly unique user visits are comparable.

on Wikipedia does come from users who encounter featured articles on the front page, and Wikipedia articles are occasionally featured elsewhere on the web—a notable example is as fact checking content with conspiracy related YouTube videos (Solon, 2018). At an even wider scope, the news stories and topics I study are of course subject to the newsfeed and search ranking algorithms which may shape audience interest. Though on the whole one would not expect a consistent strong effect in any one direction from these diverse sources.

Human behaviour on the Web is a tangled sociotechnical system, and analysing what qualities are the result of the technical designs of platforms and what is 'innate' collective behaviour (if it is even appropriate to separate them as such) is a difficult technical and philosophical task. Despite this, it is often convenient or appealing (but also frequently of research interest in its own right) to simply study a platform on which the very content of study is being selectively and opaquely distributed to users, and hope the findings and claims apply beyond the platform at hand, rather than simply reproducing the algorithm. This issue may be alleviated by greater and more open research collaboration between social media giants and academic researchers, or international level policy changes to force such openness. However, this author is pessimistic on how forthcoming details on companies' business secrets will ever be. In using Wikipedia to study a phenomenon that primarily does not take place on Wikipedia, I make efforts to avoid these issues, and call for further work towards this end.

To close, I invite the reader to consider the following extract from Pliny the Elder (translation Doody, 2009), one of the first Roman encyclopaedians.

> *Now all the subjects that the Greeks call enkuklios paideia ought to be dealt with, but they are unknown or made confusing by over-complications, while others are so often discussed that they become tedious. It is a difficult thing to give novelty to the familiar, authority to the brand new, shine to the out-of-date, clarity to the obscure, charm to the dull, authority to the implausible, its nature to everything and all its own to nature. And this is why even if I have not succeeded, it is a brilliant and beautiful enterprise.*

The very same challenges are being addressed today on Wikipedia; inclusionism vs exclusionism, tendencies for recentism, referencing issues for unfolding events, underrepresentation of marginalised groups, and perhaps most prominently, establishing credibility to the encycopaedia that anyone can edit. One may think this represents a lack of progress, yet the encyclopaedia is more information rich, accessible, and popular than ever. Wikipedia has far eclipsed any reasonable expectations for early encyclopaedians of how the format might develop. Its distilled representation of the world is measurable through both its evolving content and, perhaps more importantly, its usage patterns. Wikipedia in this role is humanity's definitive reference work.

Appendices

Appendix A

Data Details

A.1 LexisNexis Major World Publications

The NexisUK major world publications database "contains selected prestigious newspapers, newswires, and magazines from around the world. This group includes major news publications in English, Danish, Dutch, French, German, Italian, and Spanish. These news sources are held in especially high regard by their readers for their reliability and accuracy and for the integrity and objectivity of their reporting". In practice only the English language stories are used, but these make up the overwhelming majority of outlets and stories. The full list of publications is as follows.

- ADWEEK
- Accountancy Age (UK)
- Accounting Today
- Advertising Age
- Africa News
- Airline Business
- Al Jazeera - English
- Australian Financial Review
- Automotive News
- BBC Monitoring: International Reports
- Baltic News Service
- Belfast News Letter
- Belfast Telegraph
- Belfast Telegraph Online
- Billboard
- Birmingham Evening Mail

- Birmingham Post
- Brand Strategy
- Brisbane News
- Builder
- Business & Finance Magazine
- Business Day (South Africa)
- Business Monitor News
- BusinessWorld
- CFO
- CMP Information
- Canberra Times (Australia)
- Chemical Week
- Chicago Tribune
- City A.M.
- Computer Weekly
- Computing
- Contract Journal
- Creative Review
- Daily News (New York)
- Daily Record and Sunday Mail
- Daily Variety
- Design Week
- Detroit Free Press (Michigan)
- EXE
- Electronics Weekly
- Employee Benefits
- Euromoney
- Farmers Weekly
- Financial Adviser
- Financial Director
- Financial Mail (South Africa)
- Financial Times (London, England)
- Flight International
- Forbes
- Herald Sun/Sunday Herald Sun (Melbourne, Australia)
- Hobart Mercury/Sunday Tasmanian (Australia)
- ITAR-TASS
- Industry Week
- Insurance Age
- International Money Marketing
- International New York Times
- Korea Herald
- Korea Times

- Lawyers Weekly
- Legal Week
- Los Angeles Times
- MTI Econews
- MWP Advanced Manufacturing
- Maghreb Confidential
- Management Today
- Marketing Week
- Mergers and Acquisitions, The Dealmaker's Journal
- Middle East Newsfile (Moneyclips)
- Mining Magazine
- Money Marketing
- Moscow News
- Music Week
- National Post (f/k/a The Financial Post)(Canada)
- New Media Age
- New Musical Express
- New Scientist
- New Straits Times (Malaysia)
- Newsday
- Newsweek
- Nikkei Asian Review
- Northern Territory News (Australia)
- Off Licence News
- Ottawa Citizen
- Pharma Marketletter
- Plastics News (tm)
- Platts Energy Business & Technology
- Platts Megawatt Daily
- Polish News Bulletin
- Post Magazine
- Precision Marketing
- Process Engineering
- Professional Broking
- Reinsurance Magazine
- Retail Week
- Revolution
- Rubber & Plastics News
- Satellite Week
- South China Morning Post
- Sydney Morning Herald (Australia)
- Tampa Bay Times
- TechNews

- The Advertiser/Sunday Mail (Adelaide, South Australia)
- The Age (Melbourne, Australia)
- The Australian
- The Banker
- The Boston Globe
- The Business
- The Business Times Singapore
- The Christian Science Monitor
- The Courier Mail/The Sunday Mail (Australia)
- The Daily Mail and Mail on Sunday (London)
- The Daily Telegraph (Australia)
- The Daily Telegraph (London)
- The Dallas Morning News
- The Deal Pipeline
- The Dominion (Wellington)
- The Dominion Post (Wellington, New Zealand)
- The Economist
- The Edge Malaysia
- The Edge Singapore
- The Electricity Journal
- The Engineer
- The Evening Post (Wellington)
- The Evening Standard (London)
- The Express
- The Gazette (Montreal)
- The Globe and Mail (Canada)
- The Grocer
- The Guardian(London)
- The Herald (Glasgow)
- The Independent (United Kingdom)
- The Investors Chronicle
- The Irish Times
- The Japan News
- The Japan Times
- The Jerusalem Post
- The Jerusalem Report
- The Lancet
- The Lawyer
- The Miami Herald
- The Mirror (The Daily Mirror and The Sunday Mirror)

- The Moscow News (RIA Novosti)
- The Moscow Times
- The Nation (Thailand)
- The New York Times
- The New Yorker
- The New Zealand Herald
- The News of the World
- The Observer(London)
- The People
- The Philadelphia Inquirer
- The Press (Christchurch, New Zealand)
- The Straits Times (Singapore)
- The Sun (England)
- The Sunday Herald (Glasgow)
- The Sunday Times (London)
- The Times (London)
- The Toronto Star
- The Washington Times
- The Weekly Times
- The West Australian (Perth)
- USA Today
- Utility Week
- Wall Street Journal Abstracts
- Waste News
- What's New in Industry
- mirror.co.uk
- standard.co.uk
- telegraph.co.uk
- thetimes.co.uk

Appendix B

Robustness Tests

Robustness tests for parameters in the initial stage of temporal community detection on the News Event Article Networks G_i, as well as for community detection on the higher-order network H are carried out.

B.1 Temporal Community Detection

I perform a robustness test to identify the resolution parameter value which gives stable, meaningful partitions when performing community detection. I test a random sample of 50 events, repeating the community detection process over a (logarithmic) range of resolution parameters, and compare the similarity in the obtained partitions. Similarity between partition n and $n-1$ is calculated according to adjusted mutual information and CluSim element-centric similarity (Gates & Ahn, 2019) and shown in Figure B.1. In addition, I explore the variation in clustering similarity when imposing a threshold for edge weight. From -1 (i.e. no threshold), 50th percentile, and 90th percentile of the values from $W_{\text{non-edges}}$, giving W'_{edges} with

$$W'_{\text{edges } ijk} = \begin{cases} W_{\text{edges } ijk}, & \text{if } W_{\text{edges } ijk} \geq c \\ 0, & \text{otherwise} \end{cases}. \tag{B.1}$$

Based on this test, the resolution parameter for further analysis is set according to the peak similarity around $r = 0.25$. In addition, with little consistent variation between partitions with different thresholds, I proceed without the threshold on edge weight. Note that the imposed network structure already removes the vast majority of edges obtained from correlations between

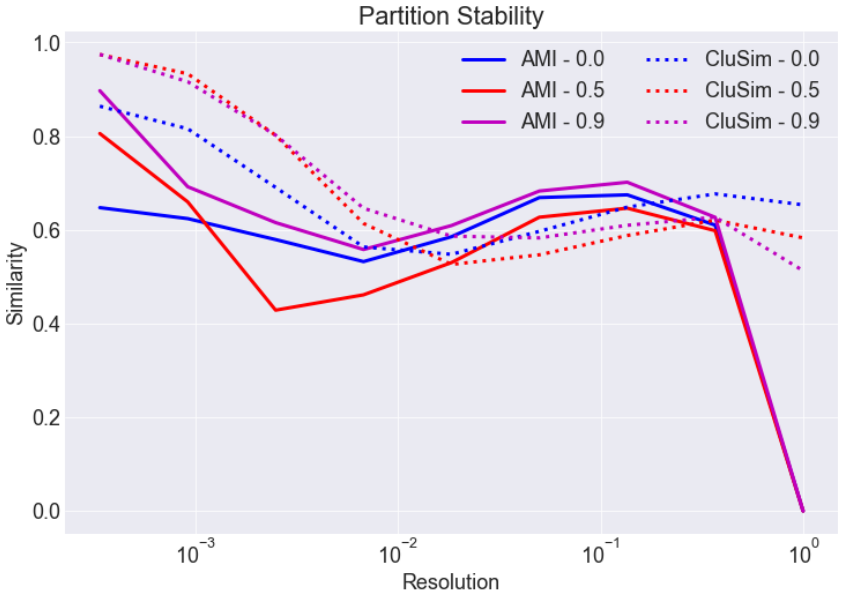

Figure B.1: Partition stability with varying resolution at the 0, 0.5, and 0.9 quantile edge weight thresholds. Partitions at smaller resolutions tend towards a single community and partitions at larger resolutions tend towards a unique community for each node. There are maxima for each similarity measure around $r = 0.25$, as well as little consistent variation between partitions with different thresholds.

nodes, so this task is not necessarily suited to the typical correlation network imposed thresholds.

B.2 Higher-Level Network Community Detection

The approach in B.1 is repeated for community detection on the higher-level network H. In this case, the partition from the maxima around $r = 0.12$ is chosen for further analysis.

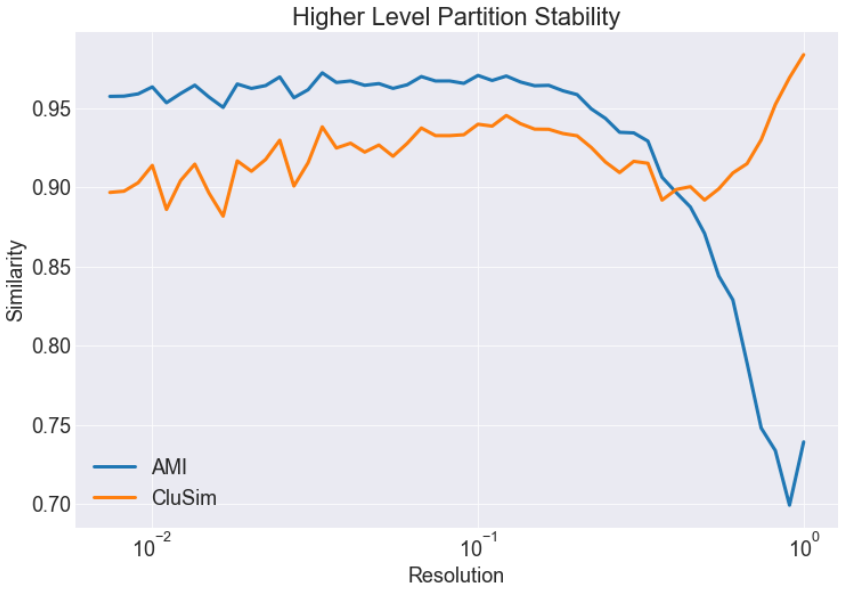

Figure B.2: Higher-level network partition stability with varying resolution. Partitions at smaller resolutions than those shown in the figure tend towards a single community and partitions at larger resolutions tend towards a unique community for each node. There are maxima for each similarity measure around $r = 0.12$.

Appendix C

Full Regression Coefficients

Summary statistics for all coefficients in the additivity hypothesis regression in Chapter 5 are provided here in Table C.1.

Table C.1: Linear regression coefficients.

	coef	std err	t	P> |t|	[0.025	0.975]
const	-0.0672	0.037	-1.812	0.070	-0.140	0.006
Topic 0	2.615e-17	6.13e-16	0.043	0.966	-1.18e-15	1.23e-15
Topic 1	0.0954	0.103	0.924	0.355	-0.107	0.298
Topic 2	0.3384	0.099	3.419	0.001	0.144	0.533
Topic 3	-0.4137	0.121	-3.426	0.001	-0.651	-0.177
Topic 4	1.0605	0.129	8.214	0.000	0.807	1.314
Topic 5	0.4331	0.153	2.822	0.005	0.132	0.734
Topic 6	-0.6754	0.112	-6.025	0.000	-0.895	-0.455
Topic 7	0.0804	0.196	0.409	0.682	-0.305	0.466
Topic 8	0.2823	0.165	1.707	0.088	-0.042	0.607
Topic 9	0.7529	0.185	4.062	0.000	0.389	1.117
Topic 10	0.3078	0.140	2.193	0.028	0.032	0.583
Topic 11	-0.2525	0.173	-1.457	0.145	-0.593	0.088
Topic 12	-0.0765	0.153	-0.499	0.618	-0.377	0.224
Topic 13	-0.2038	0.167	-1.218	0.224	-0.532	0.125
Topic 14	0.3751	0.171	2.190	0.029	0.039	0.711
Topic 15	-0.8166	0.150	-5.453	0.000	-1.110	-0.523
Topic 16	0.0545	0.278	0.196	0.845	-0.490	0.599
Topic 17	-0.4667	0.151	-3.098	0.002	-0.762	-0.171
Topic 18	0.7269	0.347	2.095	0.036	0.046	1.408
Topic 19	0.2400	0.237	1.014	0.311	-0.224	0.704
Topic 20	0.6297	0.255	2.473	0.014	0.130	1.129
Topic 21	0.0436	0.220	0.198	0.843	-0.389	0.476
Topic 22	0.1110	0.196	0.565	0.572	-0.274	0.496
Topic 23	-0.1271	0.200	-0.636	0.525	-0.519	0.265
Topic 24	-0.2132	0.184	-1.158	0.247	-0.574	0.148
Topic 25	0.2273	0.179	1.267	0.205	-0.125	0.579
Topic 26	0.0718	0.277	0.260	0.795	-0.471	0.614
Topic 27	1.1193	0.260	4.298	0.000	0.608	1.630
Topic 28	-0.5821	0.259	-2.250	0.025	-1.090	-0.075
Topic 29	0.0455	0.217	0.210	0.834	-0.380	0.471
Topic 30	-0.3233	0.349	-0.928	0.354	-1.007	0.360
Topic 31	0.3611	0.212	1.701	0.089	-0.055	0.778
Topic 32	-0.6333	0.187	-3.390	0.001	-1.000	-0.267
Topic 33	-0.2989	0.210	-1.425	0.154	-0.710	0.113
Topic 34	-0.8387	0.276	-3.044	0.002	-1.379	-0.298
Topic 35	0.8969	0.207	4.340	0.000	0.492	1.302
Topic 36	-0.1974	0.197	-1.001	0.317	-0.584	0.189
Topic 37	-0.4274	0.183	-2.331	0.020	-0.787	-0.068
Topic 38	-0.3067	0.198	-1.549	0.122	-0.695	0.082
Topic 39	-0.5449	0.187	-2.922	0.004	-0.911	-0.179
Topic 40	-1.2335	0.191	-6.454	0.000	-1.608	-0.859
Topic 41	0.1752	0.348	0.504	0.614	-0.507	0.857
Topic 42	1.0109	0.278	3.635	0.000	0.465	1.557
Topic 43	0.0745	0.227	0.329	0.742	-0.370	0.519
Topic 44	0.9178	0.224	4.094	0.000	0.478	1.358
Topic 45	-0.0547	0.239	-0.229	0.819	-0.523	0.413
Topic 46	-0.1550	0.347	-0.447	0.655	-0.835	0.525
Topic 47	-0.0909	0.235	-0.387	0.699	-0.552	0.370
Topic 48	-0.4050	0.197	-2.051	0.040	-0.792	-0.018
Topic 49	0.5430	0.349	1.554	0.120	-0.142	1.228
Topic 50	0.4725	0.228	2.069	0.039	0.025	0.921
Topic 51	0.7478	0.261	2.869	0.004	0.236	1.259
Topic 52	-0.9008	0.212	-4.245	0.000	-1.317	-0.485
Topic 53	0.6484	0.551	1.176	0.240	-0.433	1.730
Topic 54	0.0498	0.325	0.153	0.878	-0.588	0.687
Topic 55	0.2706	0.318	0.852	0.394	-0.352	0.894
Topic 56	0.6920	0.263	2.634	0.009	0.177	1.207
Topic 57	0.3030	0.391	0.774	0.439	-0.465	1.071
Topic 58	0.1631	0.319	0.511	0.610	-0.463	0.789
Topic 59	-0.3490	0.260	-1.342	0.180	-0.859	0.161
Topic 60	-0.4133	0.236	-1.748	0.081	-0.877	0.051
Topic 61	-0.4028	0.235	-1.715	0.087	-0.864	0.058
Topic 62	-0.3928	0.549	-0.716	0.474	-1.469	0.684
Topic 63	1.049e-16	2.79e-16	0.376	0.707	-4.42e-16	6.52e-16
Topic 64	-0.5234	0.245	-2.132	0.033	-1.005	-0.042
Topic 65	0.8515	0.259	3.283	0.001	0.343	1.360
Topic 66	-0.4159	0.244	-1.701	0.089	-0.895	0.064
Topic 67	0.0352	0.235	0.150	0.881	-0.426	0.496
Topic 68	-0.3332	0.250	-1.330	0.184	-0.825	0.158
Topic 69	0.4450	0.251	1.774	0.076	-0.047	0.937
Topic 70	-1.1071	0.240	-4.619	0.000	-1.577	-0.637
Topic 71	0.2724	0.236	1.155	0.248	-0.190	0.735
Topic 72	0.2479	0.450	0.551	0.581	-0.634	1.130
Topic 73	-0.2745	0.447	-0.615	0.539	-1.151	0.602
Topic 74	0.4541	0.295	1.538	0.124	-0.125	1.033
Topic 75	0.8554	0.279	3.070	0.002	0.309	1.402
Topic 76	-2.222e-16	1.96e-16	-1.132	0.258	-6.07e-16	1.63e-16
Topic 77	-0.0539	0.546	-0.099	0.921	-1.126	1.018
Topic 78	0.2983	0.449	0.665	0.506	-0.582	1.179
Topic 79	-0.1786	0.250	-0.714	0.475	-0.669	0.312
Topic 80	-0.1987	0.451	-0.441	0.659	-1.083	0.686
Topic 81	-0.2557	0.261	-0.978	0.328	-0.769	0.257
Topic 82	-0.2762	0.263	-1.049	0.294	-0.793	0.240
Topic 83	0.4086	0.350	1.169	0.243	-0.277	1.094
Topic 84	0.7224	0.352	2.052	0.040	0.032	1.413
Topic 85	0.1624	0.323	0.503	0.615	-0.472	0.796
Topic 86	-0.4652	0.293	-1.585	0.113	-1.041	0.110
Topic 87	-0.0733	0.237	-0.310	0.757	-0.537	0.391
Topic 88	0.0281	0.260	0.108	0.914	-0.482	0.538
Topic 89	-0.9146	0.548	-1.668	0.096	-1.990	0.161
Topic 90	-0.5061	0.275	-1.843	0.066	-1.045	0.033
Topic 91	-0.8334	0.262	-3.179	0.002	-1.348	-0.319
Topic 92	-0.6804	0.263	-2.592	0.010	-1.196	-0.165
Topic 93	-0.0912	0.276	-0.330	0.741	-0.633	0.451
Topic 94	0.0542	0.277	0.196	0.845	-0.489	0.597
Topic 96	-0.4778	0.262	-1.822	0.069	-0.992	0.037
Topic 97	-0.6514	0.248	-2.628	0.009	-1.138	-0.165
Topic 98	-0.1317	0.278	-0.475	0.635	-0.676	0.413
Topic 99	0.7910	0.388	2.037	0.042	0.029	1.553
Topic 100	0.2223	0.447	0.498	0.619	-0.654	1.098
Sentiment	-0.0746	0.025	-3.022	0.003	-0.123	-0.026
Prominence	-0.1469	0.033	-4.467	0.000	-0.211	-0.082
Magnitude	0.1479	0.029	5.090	0.000	0.091	0.205
Surprise	0.0065	0.026	0.251	0.802	-0.044	0.057
Follow-up	0.0553	0.035	1.579	0.114	-0.013	0.124
Uniqueness	0.0046	0.036	0.130	0.897	-0.065	0.075
Power-elite	0.1345	0.024	5.669	0.000	0.088	0.181
Proximity	0.0346	0.030	1.168	0.243	-0.024	0.093

Appendix D

Peak Models

D.1 WKSC

The squared distance measure is evaluated as

$$\begin{aligned} \hat{d}^2 &= \frac{(x-\alpha y)^T W (x-\alpha y)}{||x||_w^2} \\ &= x^T W x + \alpha^2 y^T W y - \alpha (y^T W x + x^T W y) \\ &= x^T W x + \alpha^2 y^T W y - 2\alpha y^T W x \end{aligned} \tag{D.1}$$

(since $x^T W y = y^T W x$ as W is diagonal)

Minimising this measure by differentiating with respect to α, and taking the value of alpha at $\frac{d\hat{d}^2}{d\alpha} = 0$;

$$\begin{aligned} \frac{d\hat{d}^2}{d\alpha} &= 2\alpha y^T W y - 2x^T W x \\ &= 2\alpha ||y||_w^2 - 2x^T W y \end{aligned} \tag{D.2}$$

Giving

$$\alpha = \frac{x^T W y}{||y||_w^2} \tag{D.3}$$

D.2 Cluster Centres

The centre of each cluster μ_k^* is given by the minimiser of the sum of $\hat{d}(x_i, \mu)^2$ over all $x_i \in C_k$:

$$
\begin{aligned}
\mu_k^* &= \arg\min_{\mu} \sum_{x_i \in C_k} \hat{d}(x_i, \mu)^2 \\
&= \arg\min_{\mu} \sum_{x_i \in C_k} \frac{||\alpha_i x_{iq_i)} - \mu||_w^2}{||\mu||_w^2} \\
&= \arg\min_{\mu} \frac{1}{||\mu||_w^2} \sum_{x_i \in C_k} ||\frac{x_i^T W \mu}{||x_i||_w^2} x_{i(q_i)} - \mu||_w^2 \\
&= \arg\min_{\mu} \frac{1}{||\mu||_w^2} \sum_{x_i \in C_k} ||\frac{x_i x_i^T W \mu}{||x_i||_w^2} - \mu||_w^2 \\
&= \arg\min_{\mu} \frac{1}{||\mu||_w^2} \sum_{x_i \in C_k} ||(\frac{x_i x_i^T W}{||x_i||_w^2} - I)\mu||_w^2 \\
&= \arg\min_{\mu} \frac{1}{||\mu||_w^2} \mu^T \sum_{x_i \in C_k} (\frac{x_i x_i^T W}{||x_i||_w^2} - I)^T W (\frac{x_i x_i^T W}{||x_i||_w^2} - I) \\
&= \arg\min_{\mu} \frac{1}{||\mu||_w^2} \mu^T \sum_{x_i \in C_k} (\frac{W^T x_i x_i^T}{||x_i||_w^2} - I) W (\frac{x_i x_i^T W}{||x_i||_w^2} - I)\mu \\
&= \arg\min_{\mu} \frac{1}{||\mu||_w^2} \mu^T \sum_{x_i \in C_k} (W + \frac{W x_i x_i^T W x_i x_i^T W}{||x_i||_w^4} - 2\frac{W x_i x_i^T W}{||x_i||_w^2})\mu \\
&= \arg\min_{\mu} \frac{1}{||\mu||_w^2} \mu^T \sum_{x_i \in C_k} (W - \frac{W x_i x_i^T W}{||x_i||_w^2})\mu \\
&= \arg\min_{\mu} \frac{1}{||\mu||_w^2} \mu^T W^{\frac{1}{2}} \sum_{x_i \in C_k} (I - \frac{W^{\frac{1}{2}} x_i x_i^T W^{\frac{1}{2}}}{||x_i||_w^2}) W^{\frac{1}{2}} \mu
\end{aligned}
\tag{D.4}
$$

Letting $\nu = W^{\frac{1}{2}}\mu$ and substituting $\sum_{x_i \in C_k}(I - \frac{W^{\frac{1}{2}} x_i x_i^T W^{\frac{1}{2}}}{||x_i||_w^2})$ with M leads to the following

$$
\mu_k^* = \arg\min_{\mu} \frac{\nu^T M \nu}{||\nu||^2} \tag{D.5}
$$

Then, in line with the original work, the solution is the eigenvector v_m corresponding to the smallest eigenvalue λ_m of matrix M. Rewriting in terms of μ, we find the the spectral centroid $\mu_k^* = W^{-\frac{1}{2}} v_m$.

D.3 Testing WKSC

Consider the toy example in Figure D.1. In this case there are 200 time series, 100 of which are based on a central symmetric exponential peak signal (signal group 1), and 100 which are based on a central Hamming peak signal (signal group 2), all with added Gaussian noise. 20 of the 100 time series in each group are perturbed with an additional exponential peak signal added at timestep -120. Our objective is to cluster the 200 time series based on the peak shape. In this case we are more interested in the central peak shape, rather than any secondary signals, so we would like to recover signal group 1 and signal group 2 with a 2 cluster model. With a flat window (i.e. the regular KSC algorithm) in calculating distance between time series, equal weight is given to timesteps far from the peak and close to the peak. This means that the perturbed time series in both groups may be close in distance because of the common perturbation, despite the dissimilar central overall peak shape.

We see in the Figure that the flat window cluster 1 recovers the pure exponential signal shape, but that cluster 2 recovers the Hamming signals, the perturbed Hamming signals, as well as the perturbed exponential signals in one group—with a cluster centre in between a Hamming and exponential shape. It has clustered signals based on the perturbation, rather than the central shape. In contrast, applying weights to the distance measure such as through a Hamming or exponential window (unrelated to the toy signals) with the WKSC algorithm ensures greater focus is placed on the central timesteps and the central peak shape. This means signal groups 1 and 2 are recovered better, with each group including the 20 perturbed signals. There is a slight departure from perfection for an exponential window, which in fact a result of overfitting, where too much weight is given to the noise at the most central timesteps.

I apply a similar perturbation to a random 20% of the time series in the Wikipedia data and evaluate models for 2-10 clusters with flat, Hamming, and exponential windows, comparing them to the partitions from the unperturbed dataset using adjusted mutual information (AMI). The results in Figure D.2 show a consistent higher similarity between the clusters obtained in the WKSC approach, compared to the flat KSC approach.

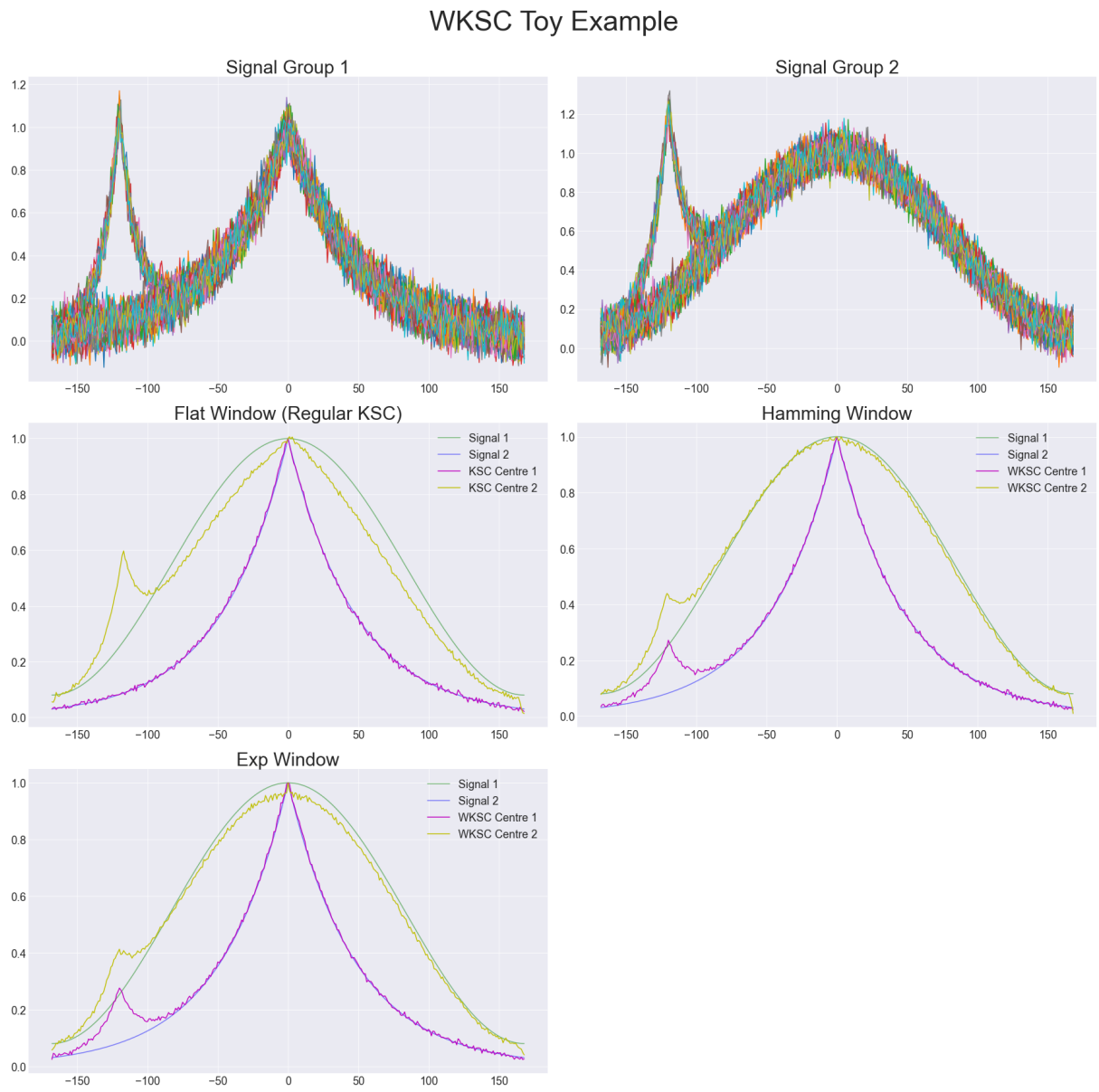

Figure D.1: Examples of the KSC and WKSC algorithms applied to the same set of time series. The WKSC algorithm recovers clusters based on the central peak shape, successfully ignoring the effect of the perturbation. The KSC algorithm is more sensitive to the perturbation, creating a cluster with all of the perturbed time series as well as the Hamming signal time series. The centre of this cluster lies between the shapes of signal 1 and signal 2, and doesn't appropriately respresent any time series.

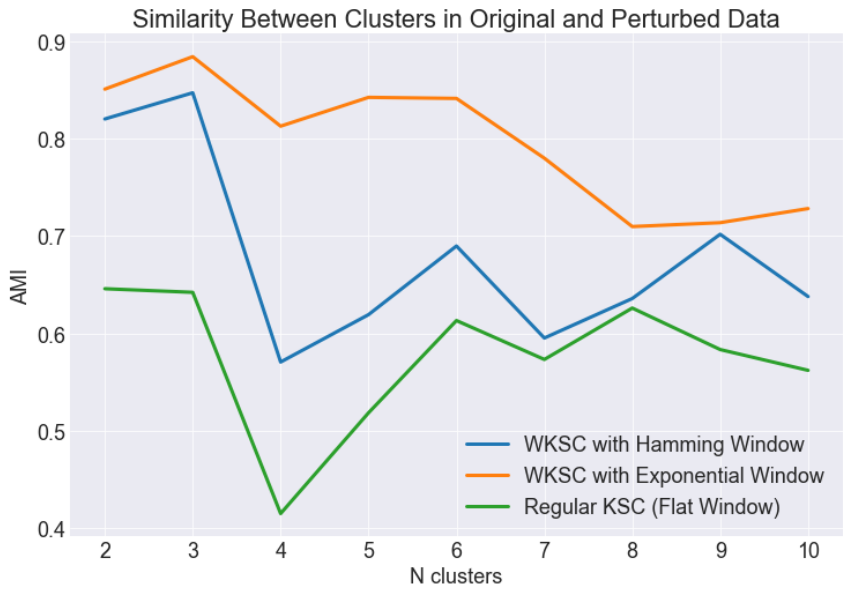

Figure D.2: Clusters in the perturbed data are more similar to those in the original data when using WKSC clustering than when using flatly weighted regular KSC clustering.

D.4 WKSC Model Selection

Model selection is performed using the Hamming window with 2-10 clusters. This is done according to Hartigan index and silhouette score, implementing our own distance measure $\hat{d}$ (Equation 6.2), with results in Figures D.3 pointing towards meaningful separation in a 5 cluster model.

D.5 Solutions to the IDC Model

Given the three governing differential equations of the IDC model,

$$\frac{dA_1}{dt} = \Lambda_1^{\text{in}}(t)A_1 - \lambda_1^{\text{out}}A_1 - \lambda_1^{\text{T}}A_1, \tag{D.6}$$

$$\frac{dA_2}{dt} = \Lambda_2^{\text{in}}(t)A_2 + \lambda_1^{\text{T}}A_1 - \lambda_2^{\text{out}}A_2 - \lambda_2^{\text{T}}A_2, \tag{D.7}$$

$$\frac{dA_3}{dt} = \lambda_2^{\text{T}}A_2, \tag{D.8}$$

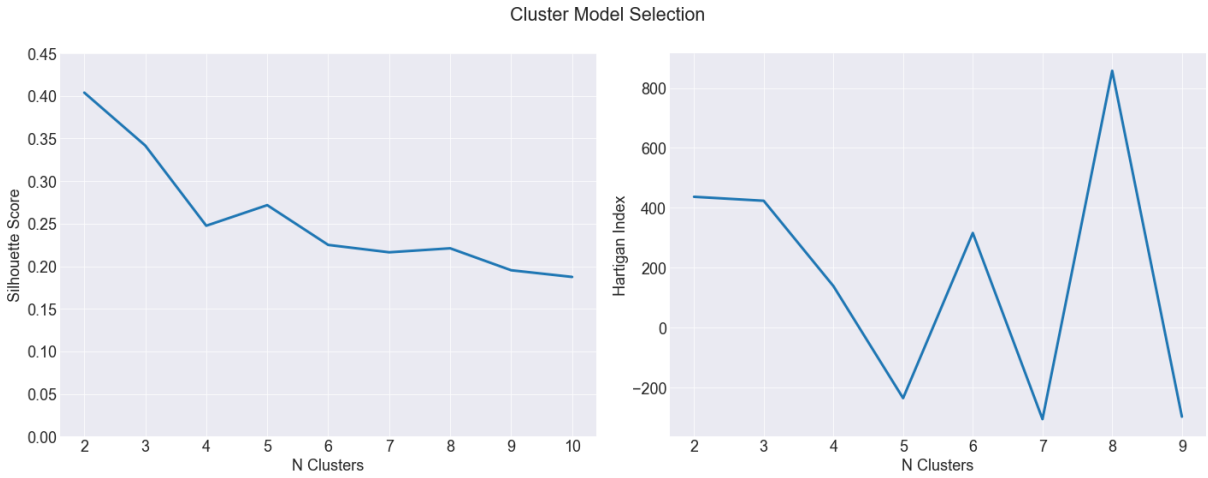

Figure D.3: Local maximum and minimum at $n = 5$ in the Silhouette score and Hartigan index model selection measures suggest proceeding with a 5 cluster model.

we seek to find analytic solutions for $A_1(t)$, $A_2(t)$, $A_3(t)$ with $\Lambda_i^{\text{in}}(t) = H(-t)\lambda_i^{\text{in}}$.

Equation D.6 is separable, and is simply solved for

$$A_1(t) = \begin{cases} A_1^0 e^{\gamma_1 t} & \text{if } t \leq 0 \\ A_1^0 e^{-\alpha_1 t} & \text{if } t > 0 \end{cases} \tag{D.9}$$

where $\gamma_i = \lambda_i^{\text{in}} - \lambda_i^{\text{out}} - \lambda_i^{\text{T}}$ and $\alpha_i = \lambda_i^{\text{out}} + \lambda_i^{\text{T}}$ Substituting into Equation D.7, and multiplying through by $e^{-\gamma_2 t}$, $e^{\alpha_2 t}$ we find

$$\begin{aligned} e^{-\gamma_2 t}\frac{dA_2}{dt} - \gamma_2 e^{-\gamma_2 t}A_2 = \frac{d(e^{-\gamma_2 t}A_2)}{dt} = \lambda_1^{\text{T}} A_1^0 e^{(\gamma_1 - \gamma_2)t} \text{ if } t \leq 0 \\ e^{\alpha_2 t}\frac{dA_2}{dt} + \alpha_2 e^{\alpha_2 t}A_2 = \frac{d(e^{\alpha_2 t}A_2)}{dt} = \lambda_1^{\text{T}} A_1^0 e^{-(\alpha_1 - \alpha_2)t} \text{ if } t > 0 \end{aligned} \tag{D.10}$$

which is solved by

$$A_2(t) = \begin{cases} \frac{\lambda_1^{\text{T}} A_1^0}{\gamma_1 - \gamma_2} e^{\gamma_1 t} + A_2'^0 e^{\gamma_2 t} & \text{if } t \leq 0 \\ -\frac{\lambda_1^{\text{T}} A_1^0}{\alpha_1 - \alpha_2} e^{-\alpha_1 t} + A_2''^0 e^{-\alpha_2 t} & \text{if } t > 0 \end{cases} \tag{D.11}$$

Where $A_2'^0 = A_2(0) - \frac{\lambda_1^{\text{T}} A_1^0}{\gamma_1 - \gamma_2}$ and $A_2''^0 = A_2(0) + \frac{\lambda_1^{\text{T}} A_1^0}{\alpha_1 - \alpha_2}$.

Finally, substituting into Equation D.8, we arrive at

$$\frac{dA_3}{dt} = \begin{cases} \frac{\lambda_1^{\mathrm{T}}\lambda_2^{\mathrm{T}}A_1^0}{\gamma_1-\gamma_2}e^{\gamma_1 t} + \lambda_2^{\mathrm{T}}A_2'^0 e^{\gamma_2 t} & \text{if } t \leq 0 \\ -\frac{\lambda_1^{\mathrm{T}}\lambda_2^{\mathrm{T}}A_1^0}{\alpha_1-\alpha_2}e^{-\alpha_1 t} + \lambda_2^{\mathrm{T}}A_2''^0 e^{-\alpha_2 t} & \text{if } t > 0 \end{cases} \tag{D.12}$$

which is

$$A_3(t) = \begin{cases} \frac{\lambda_1^{\mathrm{T}}\lambda_2^{\mathrm{T}}A_1^0}{\gamma_1(\gamma_1-\gamma_2)}e^{\gamma_1 t} + \frac{\lambda_2^{\mathrm{T}}A_2'^0}{\gamma_2}e^{\gamma_2 t} + A_3'^0 & \text{if } t \leq 0 \\ \frac{\lambda_1^{\mathrm{T}}\lambda_2^{\mathrm{T}}A_1^0}{\alpha_1(\alpha_1-\alpha_2)}e^{-\alpha_1 t} - \frac{\lambda_2^{\mathrm{T}}A_2''^0}{\alpha_2}e^{-\alpha_2 t} + A_3''^0 & \text{if } t > 0 \end{cases} \tag{D.13}$$

where $A_3'^0 = A_3^0 - \frac{\lambda_1^{\mathrm{T}}\lambda_2^{\mathrm{T}}A_1^0}{\gamma_1(\gamma_1-\gamma_2)} - \frac{\lambda_2^{\mathrm{T}}A_2'^0}{\gamma_2}$ and $A_3''^0 = A_3^0 - \frac{\lambda_1^{\mathrm{T}}\lambda_2^{\mathrm{T}}A_1^0}{\alpha_1(\alpha_1-\alpha_2)} + \frac{\lambda_2^{\mathrm{T}}A_2''^0}{\alpha_2}$

The total attention towards a subject, and the shape of the peaks is then

$$\sum A_i(t) = \begin{cases} A_1^0(1 + \frac{\lambda_1^{\mathrm{T}}}{\gamma_1-\gamma_2}(1 + \frac{\lambda_2^{\mathrm{T}}}{\gamma_1}))e^{\gamma_1 t} + A_2'^0(1 + \frac{\lambda_2^{\mathrm{T}}}{\gamma_2})e^{\gamma_2 t} + A_3'^0 & \text{if } t \leq 0 \\ A_1^0(1 - \frac{\lambda_1^{\mathrm{T}}}{\alpha_1-\alpha_2}(1 - \frac{\lambda_2^{\mathrm{T}}}{\alpha_1}))e^{-\alpha_1 t} + A_2''^0(1 - \frac{\lambda_2^{\mathrm{T}}}{\alpha_2})e^{-\alpha_2 t} + A_3''^0 & \text{if } t > 0 \end{cases} \tag{D.14}$$

Which takes dual exponential growth form for $t < 0$, and dual decay for $t > 0$.

References

Adams, J., Brückner, H., & Naslund, C. (2019). Who counts as a notable sociologist on Wikipedia? gender, race, and the "professor test". *Socius*, *5*, 2378023118823946.

Ahn, B. G., Van Durme, B., & Callison-Burch, C. (2011). WikiTopics: What is popular on Wikipedia and why. In *Proceedings of the workshop on automatic summarization for different genres, media, and languages* (pp. 33–40).

Alexa top sites. (2021). Retrieved 2021-06-14, from `https://www.alexa.com/topsites`

Almeida, R. B., Mozafari, B., & Cho, J. (2007). On the evolution of Wikipedia. In *Icwsm.*

Analytics datasets: Clickstream. (2021). Retrieved 2021-06-14, from `dumps.wikimedia.org/other/clickstream/readme.html`

Anderson, J. R. (2005). *Cognitive psychology and its implications.* Macmillan.

Aragon, P., Laniado, D., Kaltenbrunner, A., & Volkovich, Y. (2012). Biographical social networks on Wikipedia: A cross-cultural study of links that made history. In *Proceedings of the eighth annual international symposium on wikis and open collaboration* (pp. 1–4).

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., ... others (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics*, *25*(1), 25.

Bao, P., Hecht, B., Carton, S., Quaderi, M., Horn, M., & Gergle, D. (2012). Omnipedia: Bridging the Wikipedia language gap. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1075–1084).

Bartholomaeus, Trevisa, J., & Berthelet, T. (1535). *Anno. M.D.XXXV. Bertholomeus de proprietatibus rerum.* Londini: in aedibus Thomae Bertheleti regii impressoris.

Bednarek, M., & Caple, H. (2017). *The discourse of news values: How news organizations create newsworthiness.* Oxford University Press.

Bennett, J. C. (2007). *The anglosphere challenge: Why the English-speaking nations will lead the way in the twenty-first century.* Rowman & Littlefield.

Bergsma, F. (1978). News values in foreign affairs on Dutch television. *Gazette (Leiden, Netherlands)*, *24*(3), 207–222.

Black, E. (2010). *Wikipedia The dumbing down of world knowledge.* Retrieved 2021-06-14, from `http://www.thecuttingedgenews.com/index.php?article=12106&pageid=37&pagename=Page+One`

Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital journalism*, *4*(1), 8–23.

Boydstun, A. E., Hardy, A., & Walgrave, S. (2014). Two faces of media attention: Media storm versus non-storm coverage. *Political Communication*, *31*(4), 509–531.

Bright, J., & Nicholls, T. (2014). The life and death of political news: Measuring the impact of the audience agenda using online data. *Social Science Computer Review*, *32*(2), 170–181.

Callahan, E. S., & Herring, S. C. (2011). Cultural bias in Wikipedia content on famous persons. *Journal of the American Society for Information Science and Technology*, *62*(10), 1899–1915.

Caple, H., & Bednarek, M. (2013). Delving into the discourse: Approaches to news values in journalism studies and beyond. *Preprint*. Retrieved from `https://ora.ox.ac.uk/objects/uuid:1f5c6d91-bb1f-4278-a160-66149ecfb36b`

Capocci, A., Servedio, V. D., Colaiori, F., Buriol, L. S., Donato, D., Leonardi, S., & Caldarelli, G. (2006). Preferential attachment in the growth of social networks: The internet encyclopedia Wikipedia. *Physical Review E*, *74*(3), 036116.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., ... others (2018). Universal sentence encoder for English. In *Proceedings of the 2018 conference on empirical methods in natural language processing: System demonstrations* (pp. 169–174).

Chang, T.-K., Shoemaker, P. J., & Brendlinger, N. (1987). Determinants of international news coverage in the US media. *Communication Research*, *14*(4), 396–414.

Channick, R. (2012). *Encyclopaedia Britannica ends print run.* Retrieved 2021-06-14, from `https://www.latimes.com/business/la-xpm-2012-mar-14-la-fi-britannica-ends-print-20120314-story.html`

Chase, M. (2021). *Wikipedia is 20, and its reputation has never been higher.* Retrieved 2021-06-14, from `https://www.economist.com/international/2021/01/09/wikipedia-is-20-and-its-reputation-has-never-been-higher`

Cherubini, F., & Nielsen, R. K. (2016). Editorial analytics: How news media are developing and using audience data and metrics. *Available at SSRN 2739328*.

Ciampaglia, G. L., Flammini, A., & Menczer, F. (2015). The production of information in the attention economy. *Scientific Reports*, *5*, 9452.

Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F., & Flammini, A. (2015). Computational fact checking from knowledge networks. *PlOS One*, *10*(6), e0128193.

Clayman, S. E., & Reisner, A. (1998). Gatekeeping in action: Editorial conferences and assessments of newsworthiness. *American Sociological Review*, 178–199.

Clegg, N. (2019). *Facebook, elections and political speech.* Retrieved 2021-06-14, from `https://about.fb.com/news/2019/09/elections-and-political-speech/`

Cooke, R. (2021). *Wikipedia is the last best place on the Internet.* Retrieved 2021-06-14, from `https://www.wired.com/story/wikipedia-online-encyclopedia-best-place-internet/`

Crane, R., & Sornette, D. (2008). Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, *105*(41), 15649–15653.

Das, S., Lavoie, A., & Magdon-Ismail, M. (2013). Manipulation among the arbiters of collective intelligence: how Wikipedia administrators mold public opinion. In *Proceedings of the 22nd ACM international conference on conference on information & knowledge management* (pp. 1097–1106). New York, NY, USA: ACM.

De Amorim, R. C., & Mirkin, B. (2012). Minkowski metric, feature weighting and anomalous cluster initializing in K-means clustering. *Pattern Recognition*, *45*(3), 1061–1075.

Defining public interest on Twitter. (2019). Retrieved 2021-06-14, from https://blog.twitter.com/en_us/topics/company/2019/publicinterest.html

de Semir, V. (1996). What is newsworthy? *The Lancet*, *347*(9009), 1163–1166. doi: 10.1016/S0140-6736(96)90614-5

Devgan, L., Powe, N., Blakey, B., & Makary, M. (2007). Wiki-surgery? internal validity of Wikipedia as a medical and surgical reference. *Journal of the American College of Surgeons*, *205*(3), S76–S77.

Diderot, D., & Alembert, J. L. R. d. (1754). *Encyclopédie, ou, dictionnaire raisonné des sciences, des arts et des métiers [electronic resource].* Geneve ; Paris ; Neufchastel: Chez Briasson [and others].

Dimitrov, D., Lemmerich, F., Flöck, F., & Strohmaier, M. (2018). Query for architecture, click through military: Comparing the roles of search and navigation on Wikipedia. In *Proceedings of the 10th ACM conference on web science* (pp. 371–380).

Doody, A. (2009). Pliny's "Natural history: Enkuklios paideia" and the ancient encyclopedia. *Journal of the History of Ideas*, *70*(1), 1–21.

ElBahrawy, A., Alessandretti, L., & Baronchelli, A. (2019). Wikipedia and cryptocurrencies: interplay between collective attention and market performance. *Frontiers in Blockchain*, *2*, 12.

Fallis, D. (2008). Toward an epistemology of Wikipedia. *Journal of the American Society for Information science and Technology*, *59*(10), 1662–1674.

Ferron, M., & Massa, P. (2011). Collective memory building in Wikipedia: The case of North African uprisings. In *Proceedings of the 7th international symposium on wikis and open collaboration* (pp. 114–123).

Galtung, J., & Ruge, M. H. (1965). The structure of foreign news: The presentation of the Congo, Cuba and Cyprus crises in four Norwegian newspapers. *Journal of Peace Research*, *2*(1), 64–90.

García-Gavilanes, R., Mollgaard, A., Tsvetkova, M., & Yasseri, T. (2017). The memory remains: Understanding collective memory in the digital age. *Science Advances*, *3*(4), e1602368.

García-Gavilanes, R., Tsvetkova, M., & Yasseri, T. (2016). Dynamics and biases of online attention: The case of aircraft crashes. *Royal Society Open Science*, *3*(10).

Gates, A. J., & Ahn, Y.-Y. (2019). CluSim: a python package for calculating clustering similarity. *Journal of Open Source Software*, *4*(35), 1264.

General data protection regulation. (2016). Retrieved 2021-06-14, from http://data.europa.eu/eli/reg/2016/679/oj

Georgescu, M., Pham, D. D., Kanhabua, N., Zerr, S., Siersdorfer, S., & Nejdl, W. (2013). Temporal summarization of event-related updates in Wikipedia. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 281–284).

German, K. (2021). *In a post-truth world, we need Wikipedia more than ever.* Retrieved 2021-06-14, from https://www.cnet.com/news/in-a-post-truth-world-we-need-wikipedia-more-than-ever/

Gildersleve, P., & Yasseri, T. (2018). Inspiration, captivation, and misdirection: Emergent properties in networks of online navigation. In *International workshop on complex networks* (pp. 271–282).

Giles, J. (2005). *Internet encyclopaedias go head to head.* Nature Publishing Group. Retrieved 2021-06-14, from https://www.nature.com/articles/438900a

Graells-Garrido, E., Lalmas, M., & Menczer, F. (2015). First women, second sex: Gender bias in Wikipedia. In *Proceedings of the 26th ACM conference on hypertext & social media* (pp. 165–174).

Graham, M., Hogan, B., Straumann, R. K., & Medhat, A. (2014). Uneven geographies of user-generated information: Patterns of increasing informational poverty. *Annals of the Association of American Geographers*, *104*(4), 746–764.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297.

Guo, L., Vargo, C. J., Pan, Z., Ding, W., & Ishwar, P. (2016). Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling. *Journalism & Mass Communication Quarterly*, *93*(2), 332–359.

Halbwachs, M. (1950). La mémoire collective. *Paris: PUF*.

Hale, S. A. (2014). Multilinguals and Wikipedia editing. In *Proceedings of the 2014 ACM conference on web science* (pp. 99–108).

Hammer, E. M., & Zalta, E. N. (1997). A solution to the problem of updating encyclopedias. *Computers and the Humanities*, *31*(1), 47–60.

Harcup, T., & O'Neill, D. (2001). What is news? Galtung and Ruge revisited. *Journalism Studies*, *2*(2), 261–280.

Harcup, T., & O'Neill, D. (2017). What is news? News values revisited (again). *Journalism Studies*, *18*(12), 1470–1488.

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, *58*(1), 83–90.

Heath, A. (2021). *Facebook to end special treatment for politicians after Trump ban.* The Verge. Retrieved 2021-06-14, from https://www.theverge.com/2021/6/3/22474738/facebook-ending-political-figure-exemption-moderation-policy

Hecht, B., & Gergle, D. (2009). Measuring self-focus bias in community-maintained knowledge repositories. In *Proceedings of the fourth international conference on communities and technologies* (pp. 11–20).

Hill, B. M., & Shaw, A. (2014). Consider the redirect: A missing dimension of Wikipedia research. In *Proceedings of the international symposium on open collaboration* (p. 28).

History of encyclopaedias. (2021). Retrieved 2021-06-14, from https://www.britannica.com/topic/encyclopaedia/History-of-encyclopaedias#ref32037

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength natural language processing in Python.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.1212303

How the current events page works. (2021). Retrieved 2021-06-14, from https://en.wikipedia.org/wiki/Wikipedia:How_the_Current_events_page_works

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international aaai conference on web and social media* (Vol. 8).

In the news. (2021). Retrieved 2021-06-14, from `https://en.wikipedia.org/wiki/Wikipedia:In_the_news`

Ioffe, S. (2010). Improved consistent sampling, weighted minhash and l1 sketching. In *2010 IEEE international conference on data mining* (pp. 246–255).

Kane, G. C. (2009). It's a network, not an encyclopedia: A social network perspective on Wikipedia collaboration. In *Academy of management proceedings* (Vol. 2009, pp. 1–6).

Kanhabua, N., Nguyen, T. N., & Niederée, C. (2014). What triggers human remembering of events?: A large-scale analysis of catalysts for collective memory in Wikipedia. In *Proceedings of the 14th ACM/IEEE-CS joint conference on digital libraries* (pp. 341–350).

Kaplan, J., & Osofsky, J. (2016). Input from community and partners on our community standards. *Facebook Newsroom*.

Keegan, B. C. (2012). High tempo knowledge collaboration in Wikipedia's coverage of breaking news events. *PhD Thesis*.

Keegan, B. C. (2013). A history of newswork on Wikipedia. In *Proceedings of the 9th international symposium on open collaboration* (pp. 1–10).

Keegan, B. C. (2020). An encyclopedia with breaking news. In *Wikipedia @ 20.* Retrieved from `https://wikipedia20.pubpub.org/pub/dj6frhgz`

Keegan, B. C., Gergle, D., & Contractor, N. (2011). Hot off the wiki: Dynamics, practices, and structures in Wikipedia's coverage of the tōhoku catastrophes. In *Proceedings of the 7th international symposium on wikis and open collaboration* (pp. 105–113).

Keegan, B. C., Gergle, D., & Contractor, N. (2013). Hot off the wiki: Structures and dynamics of Wikipedia's coverage of breaking news events. *American Behavioral Scientist*, *57*(5), 595–622.

Kepplinger, H. M., & Ehmig, S. C. (2006). Predicting news decisions. An empirical test of the two-component theory of news selection. *Communications: The European Journal of Communication Research*, *31*(1), 25–43.

Kerdprasop, K., Kerdprasop, N., & Sattayatham, P. (2005). Weighted k-means for density-biased clustering. In *International conference on data warehousing and knowledge discovery* (pp. 488–497).

Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A*, *115*(772), 700–721.

Kobayashi, R., Gildersleve, P., Uno, T., & Lambiotte, R. (2021). Modeling collective anticipation and response on Wikipedia. In *Proceedings of the international AAAI conference on web and social media* (Vol. 15, pp. 315–326).

Kummer, M. E. (2014). Spillovers in networks of user generated content: Pseudo-experimental evidence on Wikipedia. *ZEW-Centre for European Economic Research Discussion Paper*(14-132).

Kwon, S., Cha, M., Jung, K., Chen, W., & Wang, Y. (2013). Prominent features of rumor propagation in online social media. In *2013 IEEE 13th international conference on data mining* (pp. 1103–1108).

Lamprecht, D., Dimitrov, D., Helic, D., & Strohmaier, M. (2016). Evaluating and improving navigability of Wikipedia: A comparative study of eight language editions. In *Proceedings of the 12th international symposium on open collaboration* (p. 17).

Lehmann, J., Gonçalves, B., Ramasco, J. J., & Cattuto, C. (2012). Dynamical classes of

collective attention in Twitter. In *Proceedings of the 21st international conference on World Wide Web* (pp. 251–260).

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., ... others (2015). DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, *6*(2), 167–195.

Lemmerich, F., Sáez-Trumper, D., West, R., & Zia, L. (2019). Why the world reads Wikipedia: Beyond English speakers. In *Proceedings of the twelfth ACM international conference on web search and data mining* (pp. 618–626).

Lerman, K. (2018). Computational social scientist beware: Simpson's paradox in behavioral data. *Journal of Computational Social Science*, *1*(1), 49–58.

Lerman, K., & Hogg, T. (2010). Using a model of social dynamics to predict popularity of news. In *Proceedings of the 19th international conference on world wide web* (pp. 621–630).

Lester, M. (1980). Generating newsworthiness: The interpretive construction of public events. *American Sociological Review*, *45*(6), 984-994.

List of Wikipedias. (2021). Retrieved 2021-06-14, from `https://meta.wikimedia.org/wiki/List_of_Wikipedias`

Lorenz-Spreen, P., Mønsted, B. M., Hövel, P., & Lehmann, S. (2019). Accelerating dynamics of collective attention. *Nature Communications*, *10*(1), 1759.

Loveland, J., & Reagle, J. (2013). Wikipedia and encyclopedic production. *New Media & Society*, *15*(8), 1294–1311.

Luyt, B. (2015). Wikipedia, collective memory, and the Vietnam War. *Journal of the Association for Information Science and Technology*.

Matsakis, L. (2018). *Youtube will link directly to Wikipedia to fight conspiracy theories.* Conde Nast. Retrieved 2021-06-14, from `https://www.wired.com/story/youtube-will-link-directly-to-wikipedia-to-fight-conspiracies/`

Matsubara, Y., Sakurai, Y., Prakash, B. A., Li, L., & Faloutsos, C. (2012). Rise and fall patterns of information diffusion: Model and implications. In *Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 6–14).

McCombs, M. E., & Shaw, D. L. (1972). The agenda-setting function of mass media. *Public Opinion Quarterly*, *36*(2), 176–187.

McGrady, R. D. (2020). Consensus-based encyclopedic virtue: Wikipedia and the production of authority in encyclopedias. *PhD Thesis*. Retrieved from `https://www.lib.ncsu.edu/resolver/1840.20/38333`

McIver, D. J., & Brownstein, J. S. (2014). Wikipedia usage estimates prevalence of influenza-like illness in the United States in near real-time. *PLOS Computational Biology*, *10*(4), e1003581.

MediaWiki. (2021). *API:Main page — MediaWiki, the free wiki engine.* Retrieved 2021-06-14, from `https://www.mediawiki.org/w/index.php?title=API:Main_page&oldid=3463462`

Messner, M., & DiStaso, M. W. (2013). Wikipedia versus Encyclopedia Britannica: A longitudinal analysis to identify the impact of social media on the standards of knowledge. *Mass Communication and Society*, *16*(4), 465–486.

Messner, M., & South, J. (2011). Legitimizing Wikipedia: How US national newspapers frame and use the online encyclopedia in their coverage. *Journalism Practice*, *5*(2), 145–160.

Mestyán, M., Yasseri, T., & Kertész, J. (2013). Early prediction of movie box office success based on Wikipedia activity big data. *PloS one*, *8*(8), e71226.

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41.

Milne, D., & Witten, I. H. (2008). Learning to link with Wikipedia. In *Proceedings of the 17th ACM conference on information and knowledge management* (pp. 509–518).

Miz, V., Benzi, K., Ricaud, B., & Vandergheynst, P. (2017). Wikipedia graph mining: Dynamic structure of collective memory. *arXiv preprint arXiv:1710.00398*.

Moat, H. S., Curme, C., Avakian, A., Kenett, D. Y., Stanley, H. E., & Preis, T. (2013). Quantifying Wikipedia usage patterns before stock market moves. *Scientific Reports*, *3*(1), 1–5.

Mucha, P. J., Richardson, T., Macon, K., Porter, M. A., & Onnela, J.-P. (2010). Community structure in time-dependent, multiscale, and multiplex networks. *Science*, *328*(5980), 876–878.

Murray, S. A. (2009). *The library: An illustrated history.* Skyhorse.

Nexis UK. (2018). LexisNexis (Firm).

Nicholls, T., & Bright, J. (2019). Understanding news story chains using information retrieval and network clustering techniques. *Communication Methods and Measures*, *13*(1), 43–59.

Nystedt, D. (2007). *Baidu may be worst Wikipedia copyright violator.* Retrieved 2021-06-14, from `https://www.pcworld.com/article/135550/article.html`

Osborne, C., Graham, M., & Dittus, M. (2021). Edit wars in a contested digital city: Mapping wikipedia's uneven augmentations of berlin. *The Professional Geographer*, *73*(1), 85–95.

Osborne, M., Petrovic, S., McCreadie, R., Macdonald, C., & Ounis, I. (2012). Bieber no more: First story detection using twitter and wikipedia. In *Sigir 2012 workshop on time-aware information access* (pp. 16–76).

Overly, S. (2019). *YouTube CEO: Politicians can break our content rules.* Retrieved 2021-06-14, from `https://www.politico.com/story/2019/09/25/youtube-ceo-politicians-break-content-rules-1510919`

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web.* (Tech. Rep.). Stanford InfoLab. Retrieved from `http://ilpubs.stanford.edu:8090/422/`

Pentzold, C. (2009). Fixing the floating gap: The online encyclopaedia Wikipedia as a global memory place. *Memory Studies*, *2*(2), 255–272.

Perez, S. (2020). *Facebook tests Wikipedia-powered information panels, similar to Google, in its search results.* TechCrunch. Retrieved 2021-06-14, from `https://techcrunch.com/2020/06/11/facebook-tests-wikipedia-powered-information-panels-similar-to-google-in-its-search-results`

Piotrkowicz, A., Dimitrova, V., & Markert, K. (2017). Automatic extraction of news values from headline text. In *Proceedings of the student research workshop at the 15th conference of the European chapter of the association for computational linguistics (EACL SRW 2017)* (pp. 64–74).

Portal:current events - Wikipedia. (2021). Retrieved 2021-06-14, from `https://en.wikipedia.org/wiki/Portal:Current_events`

Potts, A., Bednarek, M., & Caple, H. (2015). How can computer-based methods help researchers

to investigate news values in large datasets? a corpus linguistic study of the construction of newsworthiness in the reporting on Hurricane Katrina. *Discourse & Communication*, *9*(2), 149–172.

Randall, E. (2014). *How a raccoon became an aardvark.* Retrieved 2021-06-14, from `https://www.newyorker.com/tech/annals-of-technology/how-a-raccoon-became-an-aardvark`

Ratkiewicz, J., Flammini, A., & Menczer, F. (2010). Traffic in social media I: Paths through information networks. In *2010 IEEE second international conference on social computing* (pp. 452–458).

Ratkiewicz, J., Menczer, F., Fortunato, S., Flammini, A., & Vespignani, A. (2010). Traffic in social media II: Modeling bursty popularity. In *2010 IEEE second international conference on social computing* (pp. 393–400).

Reagle, J., & Rhue, L. (2011). Gender bias in Wikipedia and Britannica. *International Journal of Communication*, *5*, 21.

Rosengren, K. E. (1970). International news: Intra and extra media data. *Acta Sociologica*, *13*(2), 96–109.

Rosenzweig, R. (2006). Can history be open source? Wikipedia and the future of the past. *The Journal of American History*, *93*(1), 117–146.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.

Sande, Ø. (1971). The perception of foreign news. *Journal of Peace Research*, *8*(3-4), 221–237.

Scharkow, M. (2013). Thematic content analysis using supervised machine learning: An empirical evaluation using German online news. *Quality & Quantity*, *47*(2), 761–773.

Schaudt, S., & Carpenter, S. (2009). The news that's fit to click: An analysis of online news values and preferences present in the most-viewed stories on azcentral.com. *Southwestern Mass Communication Journal*, *24*(2).

Schultz, I. (2007). The journalistic gut feeling: Journalistic doxa, news habitus and orthodox news values. *Journalism Practice*, *1*(2), 190–207.

Schulz, W. (1976). *Die konstruktion von realität in den nachrichtenmedien: Analyse der aktuellen berichterstattung.* Alber.

Schwarz, A. (2006). The theory of newsworthiness applied to Mexico's press. How the news factors influence foreign news coverage in a transitional country. *Communications*, *31*(1), 45–64.

Seward, Z. (2009). *Here's the AP document we've been writing about.* Retrieved 2021-06-14, from `https://www.niemanlab.org/2009/08/heres-the-ap-document-weve-been-writing-about/`

Seymour, M. C. (1992). *Bartholomaeus Anglicus and his encyclopedia.* Variorum.

Shoemaker, P. J., Chang, T.-K., & Brendlinger, N. (1987). Deviance as a predictor of newsworthiness: Coverage of international events in the US media. *Annals of the International Communication Association*, *10*(1), 348–365.

Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2018). A comparison of ARIMA and LSTM in forecasting time series. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)* (pp. 1394–1401).

Singer, P., Lemmerich, F., West, R., Zia, L., Wulczyn, E., Strohmaier, M., & Leskovec, J.

(2017). Why we read Wikipedia. In *Proceedings of the 26th international conference on World Wide Web* (pp. 1591–1600).

Smellie, W. (1771). *Encyclopædia Britannica [electronic resource] : or, a dictionary of arts and sciences, compiled upon a new plan. ... Illustrated with one hundred and sixty copperplates. By a society of gentlemen in Scotland. In three volumes.* Edinburgh: printed for A. Bell and C. Macfarquhar; and sold by Colin Macfarquhar.

Solon, O. (2018). *YouTube will use Wikipedia to help solve its conspiracy theory problem.* Retrieved 2021-06-14, from `https://www.theguardian.com/technology/2018/mar/13/youtube-wikipedia-flag-conspiracy-theory-videos`

Steiner, T., Van Hooland, S., & Summers, E. (2013). MJ no more: Using concurrent Wikipedia edit spikes with social network plausibility checks for breaking news detection. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 791–794).

Suh, B., Convertino, G., Chi, E. H., & Pirolli, P. (2009). The singularity is not near: Slowing growth of Wikipedia. In *Proceedings of the 5th international symposium on wikis and open collaboration* (p. 8).

Summers, N. (2013). *Wikimedia Foundation sites hit 500m unique visitors each month.* Retrieved 2021-06-14, from `https://thenextweb.com/insider/2013/04/19/sites-owned-by-the-wikimedia-foundation-now-receive-over-500m-unique-visitors-each-month/#.tnw_GGK22yoV`

Tandoc, E. C. (2014). Journalism is twerking? How web analytics is changing the process of gatekeeping. *New Media & Society*, *16*(4), 559–575. doi: 10.1177/1461444814530541

Thompson, N., & Hanley, D. (2018). Science is shaped by Wikipedia: Evidence from a randomized control trial. *Preprint*. Retrieved from `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3039505` doi: https://dx.doi.org/10.2139/ssrn.3039505

Traag, V. A., Van Dooren, P., & Nesterov, Y. (2011). Narrow scope for resolution-limit-free community detection. *Physical Review E*, *84*(1), 016114.

Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports*, *9*.

TwitterInc. (2020). *Our plans to relaunch verification and what's next.* Retrieved 2021-06-14, from `https://blog.twitter.com/en_us/topics/company/2020/our-plans-to-relaunch-verification-and-whats-next.html`

Twyman, M., Keegan, B. C., & Shaw, A. (2017). Black Lives Matter in Wikipedia: Collective memory and collaboration around online social movements. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing* (pp. 1400–1412).

Understanding media and culture: An introduction to mass communication. (2016). University of Minnesota Libraries Publishing. Retrieved from `https://open.umn.edu/opentextbooks/textbooks/143`

Van Dalen, A., de Vreese, C., & Albæk, E. (2017). Economic news through the magnifying glass: How the media cover economic boom and bust. *Journalism Studies*, *18*(7), 890–909.

Vincent, N., & Hecht, B. (2021). A deeper investigation of the importance of Wikipedia links to search engine results. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–15.

Vinh, N. X., Epps, J., & Bailey, J. (2010). Information theoretic measures for clusterings

comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research*, *11*, 2837–2854.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*, 261–272. doi: 10.1038/s41592-019-0686-2

Voss, J. (2005). Measuring Wikipedia. In *Proceedings of the 10th international conference of the international society for scientometrics and informetrics.* Retrieved from `http://hdl.handle.net/10760/6207`

Vu, H. T. (2014). The online audience as gatekeeper: The influence of reader metrics on news editorial selection. *Journalism: Theory, Practice & Criticism*, *15*(8), 1094–1110.

Waldman, S. (2004). *Who knows?* Guardian News and Media. Retrieved 2021-06-14, from `https://www.theguardian.com/technology/2004/oct/26/g2.onlinesupplement`

Wang, D., Song, C., & Barabási, A.-L. (2013). Quantifying long-term scientific impact. *Science*, *342*(6154), 127–132.

Watt, J. H., Mazza, M., & Snyder, L. (1993). Agenda-setting effects of television news coverage and the effects decay curve. *Communication Research*, *20*(3), 408–435.

Welser, H. T., Cosley, D., Kossinets, G., Lin, A., Dokshin, F., Gay, G., & Smith, M. (2011). Finding social roles in Wikipedia. In *Proceedings of the 2011 iconference* (pp. 122–129).

West, R., Weber, I., & Castillo, C. (2012). A data-driven sketch of Wikipedia editors. In *Proceedings of the 21st international conference on world wide web* (pp. 631–632).

Westerståhl, J., & Johansson, F. (1994). Foreign news: News values and ideologies. *European Journal of Communication*, *9*(1), 71–89.

What Wikipedia is not. (2021). Wikimedia Foundation. Retrieved 2021-06-14, from `https://en.wikipedia.org/wiki/Wikipedia:What_Wikipedia_is_not#Wikipedia_is_not_a_newspaper`

Wikimedia. (2021). *Wikimedia downloads.* Retrieved 2021-06-14, from `https://dumps.wikimedia.org/`

Wikipedia founder Jimmy Wales responds. (2004). Slashdot.org. Retrieved 2021-06-14, from `https://slashdot.org/story/04/07/28/1351230/wikipedia-founder-jimmy-wales-responds`

Wikipedia: Notability. (2021). Retrieved 2021-06-14, from `https://en.wikipedia.org/wiki/Wikipedia:Notability`

Wikipedia: Notability (events). (2021). Retrieved 2021-06-14, from `https://en.wikipedia.org/wiki/Wikipedia:Notability_(events)`

Wikipedia statistics. (2021). Retrieved 2021-06-14, from `stats.wikimedia.org/EN/TablesWikipediaZZ.htm`

Withers, R. (2018). *Amazon owes Wikipedia big-time.* Slate. Retrieved 2021-06-14, from `https://slate.com/technology/2018/10/amazon-echo-wikipedia-wikimedia-donation.html`

Wu, F., & Huberman, B. A. (2008). Popularity, novelty and attention. In *Proceedings of the 9th ACM conference on electronic commerce* (pp. 240–245).

Yang, D., Halfaker, A., Kraut, R., & Hovy, E. (2016). Who did what: Editor role identification in Wikipedia. In *Tenth international AAAI conference on web and social media.*

Yang, J., & Leskovec, J. (2011). Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on web search and data mining* (pp. 177–186).

Yasseri, T., & Bright, J. (2014). Can electoral popularity be predicted using socially generated big data? *it-Information Technology*, *56*(5), 246–253.

Yasseri, T., & Bright, J. (2016). Wikipedia traffic data and electoral prediction: Towards theoretically informed models. *EPJ Data Science*, *5*(1), 1–15.

Yasseri, T., & Menczer, F. (2021). Can the Wikipedia moderation model rescue the social marketplace of ideas? *arXiv preprint arXiv:2104.13754*.

Yoshida, M., Arase, Y., Tsunoda, T., & Yamamoto, M. (2015). Wikipedia page view reflects web search trend. In *Proceedings of the ACM web science conference* (pp. 1–2).

Zaccone, G., Karim, M. R., & Menshawy, A. (2017). *Deep learning with TensorFlow*. Packt Publishing Ltd.

Zhao, Z., Chen, W., Wu, X., Chen, P. C., & Liu, J. (2017). LSTM network: A deep learning approach for short-term traffic forecast. *IET Intelligent Transport Systems*, *11*(2), 68–75.

Zlatić, V., Božičević, M., Štefančić, H., & Domazet, M. (2006). Wikipedias: Collaborative web-based encyclopedias as complex networks. *Physical Review E*, *74*(1), 016115.