Distributed Moderation systems - an exploration of their utility and the social implications of their widespread adoption.

by

Richard A. Mills

MRes (Applied Social Statistics), Lancaster University (2009)

MSc (Advanced Psychological Research Methods), Dundee University (2007)

BSc (Psychology), University of Abertay (2006)

Thesis submitted for the degree of Ph.D at the University of Lancaster

June 2013

Distributed Moderation systems - an exploration of their utility and the social implications of their widespread adoption.

by

Richard A. Mills

MRes (Applied Social Statistics), Lancaster University (2009)

MSc (Advanced Psychological Research Methods), Dundee University (2007)

BSc (Psychology), University of Abertay (2006)

Abstract

The present research introduces and investigates Distributed Moderation systems - in particular, sites where the votes of users are aggregated in order to rank or grade items of content. The primary subject of this research is `reddit.com` - a 'social news' website where users vote to collectively determine the level of visibility which will be afforded to submitted items of content.

This research is investigative in nature - at its inception there was little published research on Distributed Moderation (DM) systems. The question which has guided the research is "what can we learn about these systems through observation and the interrogation of data which they naturally produce and store in their day-to-day operation?". There are Chapters of the thesis which investigate how DM works in practice (Chapter 5) and how/why individual users participate (Chapter 6). The research also devotes considerable attention to the social implications of producing information resources in this fashion (Chapter 7) - how do the resources which are produced using DM systems differ to those produced in a more conventional manner?

At the outset of this research `reddit` was a relatively little-known website - over the course of the research it has become much more widely recognised and in the process it has changed considerably. Chapter 8 considers `reddit` from a longitudinal perspective; observing its development has offered insight into both the potential and the limitations of this particular application of DM. The final Chapter re-visits research questions and considers how one might go about adapting DM to other domains, with an emphasis on the political.

Declaration

This thesis results entirely from my own work and has not been offered previously for any other degree or diploma.

Signature -

Acknowledgements

Contents

List of Figures

Chapter 1

Introduction

The Internet is possibly one of the biggest drivers of social change today. It allows anyone who is online to communicate with anyone else who is also online. National borders and geographical barriers are of little consequence to these communications. The Internet also allows any individual who is connected to publish information at the click of a button - be this in the form of text, picture, audio or video. 'Web 2.0' sites (O'Reilly, 2005) such as `Youtube` have further lowered the barriers to entry as an information producer, through a website like `Youtube` any individual who creates a user account can 'publish' their own videos with no special technical expertise required. These 'published' videos are then viewable by any individual who can access `Youtube`, either through the `Youtube` website itself or by being embedded in another web page.

The ease with which an individual can become an information (or 'content') producer online is directly implicated in the rapidly increasing quantity of information available on the World Wide Web. As the quantity of information available online continues to increase exponentially our capacity to navigate and make sense of this vast information repository becomes increasingly important. This problem is often referred to as 'Information Overload' - and providing solutions in this area has been a route to success for a number of high profile Internet-based companies. Google is the obvious example of a hugely successful company which has gained its success by allowing people to navigate the web more efficiently. Indeed a whole industry (Search Engine Optimisation) has sprung up whose sole purpose is to increase the prominence of a client's websites through search engines such as Google - one indication of the importance now placed on gaining the attention of Internet users.

Social Network based web applications (such as `Facebook` and `Twitter`) can also be thought of as solutions to the problem of information overload. Where a search engine like Google provides information in response to search queries, Social Networking sites provide information on the basis of personal relationships (or 'network ties'). These websites allow a user to monitor information

feeds produced by, or addressed to, selected people known to the user - with information in these systems flowing along network ties.

Websites which employ Distributed Moderation and filtering systems offer a different approach to information overload, one which has been steadily gaining traction in recent years. It is these Distributed Moderation (or DM) web systems which are the subject of the present research. The principle behind DM systems is that some dimension of user activity is taken as a measure of an item's quality or relevance - this measure being used to rank and sort items, with those items which rank more highly being placed in positions of greater visibility.

There are a growing number of websites which are built on the principles of Distributed Moderation and in recent years several such websites have seen a surge in their popularity. The specifics of how Distributed Moderation works varies from website to website. Some of these websites derive the measure of an item's quality or relevance from user actions which serve a purpose external to the task of ranking and sorting content. A simple version of this mechanism uses the number of times an item has been viewed as an indication of its quality or 'interestingness'. For example, `Youtube` uses the number of times a video has been viewed in this manner, and several news providing websites (e.g. BBC News) offer lists of the most viewed or forwarded content on their website.

'Social Bookmarking' websites (e.g. `Delicious.com`) employ more sophisticated systems based on similar principles. On these websites a user's decision to bookmark an item increases its visibility for other users - and the tags which a user attaches to these bookmarks are aggregated such that they can be used by other users to navigate the website and search for content of particular types. For example, a user who wished to learn about photography could search for the key words 'photography' and 'tips', and be presented with all of the bookmarks which have been tagged with these words, with the most often bookmarked items appearing at the top of the list.

On Social Bookmarking websites the measure of an item's quality is still based on actions which serve a purpose external to the ranking procedure; but whether a user chooses to bookmark an item or not seems like a better indicator of quality than whether they viewed the item. Every user who wishes to form an opinion about the item must view it first, if the metric of item quality is based on view counts then even those users who find the content objectionable will increase the salience of that item of content. The tagging systems employed by Social Bookmarking websites allow their users to go beyond merely ranking the content submitted to the website. By aggregating these tags the Social Bookmarking website's users can collectively impose a classification scheme on the content submitted there - making it easier for users to navigate this content.

On other websites (such as the 'Social News' websites which serve as a focal point for this research) users participate directly in the ranking and sorting of content through a voting system. These voting systems generally allow for both 'up' votes which increase an item's score (and there-

fore visibility), and 'down' votes which decrease an item's score. This type of website will be referred to here as 'Social News'. Popular exemplars of this form are `reddit.com`, `Digg.com` and `Slashdot.org`. These websites allow anyone to create a user account, and anyone who is signed into a user account can both submit items of content or comments, and vote on the items or comments submitted by other users. These votes are used to dynamically rank the items of content and comments, and the order these are displayed in is determined by their rank.

The items submitted to these websites are most often hyper-links to externally hosted web resources, although `reddit.com` also allows for the submission of text posts created on the website directly by the submitting user. Much of activity on Social News websites takes the form of up/down voting and these websites have a 'Front page' which serves as the focal point of user attention and voting activity. Through the voting system very large numbers of users make a collective decision about which items will gain a place on the website's Front page - where the items will be seen by the greatest number of users and visitors to the website.

We can draw an analogy between Social News websites and conventional news resources such as newspapers. Users who submit content can be thought of as journalists, and users who vote collectively fulfill the role of an editor. In this analogy the website's Front page represents the editor's decision about which of the many submitted items warrant publication. The major disparities between the workings of social and conventional news are as follows:

- On a Social News website many of the items have been found rather than produced by the submitting user, with the submitting user merely giving the item a title and deciding which category to submit it to.

- On a Social News website every user can act as both journalist and editor, by submitting and voting respectively.

- On a Social News website each user acts autonomously - the 'editor' cannot dictate what the 'journalists' submit, although at times text posts about the kind of items which do or do not belong on the Front page have been quite common.

Research on Social News websites is warranted for a number of reasons. These websites represent a novel 'crowdsourcing' approach to information overload. The tasks of content discovery and filtering/ranking of this content are separated and opened up to a very large number of participants. The principles behind these websites are also gaining traction elsewhere on the web. Websites such as `AmericaSpeakingOut.com` and `YourFreedom.hmg.gov.uk` apply similar systems in the domain of political policy. The principle of up/down voting has also been applied in a question & answer format by websites such as `Stackoverflow.com` - where items are questions and the comments on these are intended to be answers, with up/down voting being used to rate these questions and answers. The same system can also be employed in reverse to produce a mediated interview (where

comments are questions and the highest scoring questions are answered) - for example the 'Digital Debate' conducted with the leaders of the main political parties in the run-up to the 2010 General Election in the UK (Guardian, 2010a).

All of the applications of up/down voting listed above are being employed on `reddit.com` and it appears that `reddit`'s users have been pivotal to the creation or popularisation of some of these. `Reddit.com` also seems to have opened more aspects of its functioning up to distributed decision-making than other websites (see section 3.1.1). These factors would seem to make `reddit.com` a particularly good venue for research on the workings of Distributed Moderation systems.

The basic principle of up/down (or 'like/dislike') voting has been proliferating quite rapidly in recent years. Social Media applications like `Youtube` and `Facebook` have recently applied this mechanism to submitted items of content and comments thereon - and a number of British newspapers now also employ similar systems for user comments on their articles (or have at one point experimented with this approach). It seems that any website which deals with a large number of user contributions and wishes to sort these based on their quality could benefit from Distributed Moderation. If a website receives 10,000 unfiltered contributions per day, and wishes to sort these such that the best items are most visible, then allowing their users to register their judgments on these items seems like a logical step to take. Up/down voting allows users to register such judgments (albeit on one dimension only) quickly and easily, and these judgments are immediately and automatically made use of by the algorithms the website uses to sort content.

On the modern-day frontier that is the World Wide Web new methods of communication are continually being introduced and evolving. The web is still very much in its infancy, and the way in which its use is developing is highly unpredictable and at the same time highly significant for the future. The Internet is bound to have far-reaching social implications, but without an understanding of where this technology is heading it is hard to predict what the ultimate consequences for society might be.

Distributed Moderation systems seem to be on the rise. Websites built on these principles are themselves growing, and the mechanisms these websites rely on are increasingly being adopted by websites which exist for some other purpose. These Distributed Moderation systems have sociological importance because of their unique approach to the problem of Information Overload, but this is not the only quality which identifies them as a particularly interesting development in digital communications technology. The use of voting-based Distributed Moderation systems on 'Social News' websites has the fascinating side-effect of allowing very large numbers of individuals to participate in collective decision-making and discussion. Furthermore, unlike older bulletin boards and discussion forums, Distributed Moderation systems do not seem to have an upper limit to the number of participants they can sustain in a single discussion or decision. A user who joins one of these 'democratically mediated' conversations late in the day (e.g. when several hundred

individuals have already made contributions) can express their opinion by voting to change the score of pre-existing contributions - influencing the discussion without adding to the quantity of text of which it is comprised.

Social News websites are also of particular interest because they operate in the domain of 'news'. An informed citizenry is critical to a functioning democracy (Mattson, 1998) and these websites have a unique approach to the way in which 'the news' is determined. It may be the case that this socially driven process of determining what 'the news' is has some benefits when compared to conventional sources like newspapers and television news. This is an issue which will be addressed in Chapter 2.

As Social News allows every user to be involved in voting to determine what 'the news' is, there may also be psychological effects associated with participation on these websites. Where the reader of a newspaper or viewer of television news is passive with respect to the production of that resource (although probably not with respect to their interpretation thereof) - Social News users are engaged to varying degrees in the production of the specific Social News resource they consume. This may produce an effect where these users feel more empowered and confident in their opinions. For further discussion see section 3.3.

In summary, the key reasons for studying Distributed Moderation systems (and Social News websites in particular) are as follows:

- They offer a novel, people powered, approach to the problem of information overload.

- Social News in particular offers a novel approach to the coverage of current events - an approach which may have different strengths (and weaknesses) when compared to the mass media.

- This approach is gaining traction - Social News websites are becoming more popular, and the mechanisms they use are being adopted in other domains (most notably politics).

- Voting systems allow for a previously impractical number of individuals to participate in producing a single decision or discussion.

- The effects these systems have on the resources which they produce, and on the individuals who participate, are largely unknown.

On the surface, Distributed Moderation systems seem to perform well on their task of selecting, from thousands of submissions, the cream of the crop - and displaying these to the broadest audience. It is presumably a rare event when a Social News website's Front page plays host to content which the majority of its users find to have no merit. Indeed, if this were a regular occurrence it seems unlikely that the user-base of a website like `reddit.com` would be growing so quickly. Furthermore, the nature of the items of content submitted (whether they are news

articles, humerous pictures, videos, or ideas for political policy) seems immaterial to the Distributed Moderation system - all of these content types seem equally amenable to Distributed Moderation. These systems also seem to have no obvious problem dealing with large and ever-increasing volumes of user activity and content submissions. [1]

A Utopian outlook on the potential of Distributed Moderation would suggest that these systems can be deployed in any context and with any user population size. A nation could, for instance, deploy such a system to collect, rank and sort its population's political policy ideas; allowing every member of the population to contribute, with the most popular ideas rising to the top in an organic fashion and without the need for gatekeepers - from here going on to form the basis of public political discourse.

This scenario is of course highly speculative. We do not know whether DM systems reliably rank the 'best' items most highly, or whether this process has a large element of randomness or social influence (e.g. when the first vote an item receives is negative does this influence the perception of subsequent voters?). We do not know whether DM systems can deal well with content which is political or controversial in nature - for example it is quite possible that small groups of highly motivated users can coordinate their actions in some way to sideline content which they find objectionable. We do not know how individuals react to seeing content they find personally objectionable judged positively through these systems (e.g. how would one feel if one's peers consistently up-voted racist proposals or perspectives to high-visibility locations?). We do not know whether these systems can in fact handle very large numbers of contributors as easily as they appear to, perhaps beyond a certain user population level the random element or the role of social influence increases in importance. It also seems likely that these systems require a minimum number of users to function as intended, and without this critical mass do not reach their potential.

Given that what we do not know about DM systems far outweighs what we do know, it is somewhat alarming that these systems have already been deployed in the context of national politics. Both the American Republican party (`AmericaSpeakingOut.com`) and the British coalition government (`YourFreedom.hmg.gov.uk`) have already applied some form of Distributed Moderation to gather (rated) political policy ideas from the general population.

The present research aims to fill some of these gaps in our understanding of Distributed Moderation - by looking in detail at some instances of DM which have been growing and developing for years I aim to understand the relative strengths and weaknesses of these principles - and in so doing inform decisions about whether such systems should be trialled in important domains like politics - and if so how they might be configured to maximise their utility and minimise potential disadvantages.

The rapid pace of change in Internet use generally and Distributed Moderation in particular results

---

[1]This introduction was written near the start of the research in 2010, by its conclusion in 2013 the 'first impression' which `reddit` makes is quite different.

in an unusual problem for this thesis whereby the subject of the research has changed considerably over the course of the research. This PhD research began in October 2009, when `reddit.com` (the primary website of interest) was a relatively unknown Social News website - widely perceived as being secondary in size and importance to `Digg.com`. Writing now in 2013, Reddit has grown substantially to become a widely recognised website (e.g. Obama participated in an 'Ask Me Anything' post in 2012) hosting levels of user activity which are orders of magnitude greater than was the case in 2009 - while `Digg` has suffered a steep decline before being sold and re-launched as an almost unrecognisable site. This thesis has been written in incremenets over the duration of this 3.5 year period and consequently certain portions of it are already out-dated. For example, on `reddit` in 2009 there was a strong sense that the 'Front page' was central to the website and the same for everyone - whereas in 2013 `reddit`'s long-standing users appear to have migrated to smaller subreddits, the larger 'default' subreddits of 2009 appear to have suffered a decline in the quality of material which is highly ranked, and the subreddits which feed into `reddit`'s Front page by default have also been changed. If one was to browse the `reddit.com` 'Front page' for the first time now in early 2013 it would likely not be immediately apparant why the term 'News' has been incorporated in a description of the website.

Chapter 8 will give an account of how the websites which were studied here have changed over the course of the research. For other chapters it is helpful to know when they were written:

- Chapters 2, 3 & 6 were written largely in 2010.

- Chapter 7 was written largely in 2011.

- Portions of Chapter 5 were written in 2010, with much of it being re-written with new analyses in 2012.

- Chapters 4, 8 and 9 were written largely in 2012/2013.

This research is exploratory in nature and the questions which have guided it are general ones: How does Distributed Moderation work in practice? How and why do people participate on DM websites? How do the resources produced by DM websites differ from their more conventionally produced alternatives? What are the social implications of producing information resources in this fashion? How has `reddit.com` develeoped and changed over the course of this research (a longitudinal perspective on DM)? What other purposes might the principles of Distributed Moderation be applied or adapted to?

The thesis is therefore structured as follows:

- Chapter 2 - a review of some relevant background literature on the Internet, largely from the disciplines of sociology and political science.

- Chapter 3 - a review of literature which is more directly related to Distributed Moderation.

- Chapter 4 - an overview of the methods employed by this research - chiefly a description of how data from websites of interest was collected and transformed.

- Chapter 5 - How does Distributed Moderation work in practice?

- Chapter 6 - How and why do people participate on DM websites?

- Chapter 7 - The social implications of Distributed Moderation - how does the interaction on DM websites, and the resources which they produce, differ from more conventional approaches to the production of information?

- Chapter 8 - How has `reddit`, the website which this research has focused on, changed over the course of the research?

- Chapter 9 - Discussion and Conclusions.

In addition, we might consider this research as having one all-encompassing background question: what can we learn about Distributed Moderation through the analysis of the procedural data these websites generate in their day-to-day operation? We are entering an era where Internet use is increasingly ingrained in peoples' daily lives - Social Media applications keep highly accurate and detailed records of users' behaviour by necessity, they need this data to function. Analysis of this ecologically valid data has the capacity to yield rich understanding of this increasingly prominent aspect of social life.

Chapter 2

A review of relevant literature from the Social Sciences

This chapter will review the social science literature which is relevant to the present research. It will begin with a brief overview of some sociological theories concerning the Internet; proceeding to discuss the more recent development of 'Web 2.0' - relating this to pre-existing theory about the Internet, discussing research which explicitly addresses Web 2.0 applications, and describing some types of research which have been conducted on the new 'social' web.

2.1 A brief history of the Internet - and the birth of the Web

The 'Information Technology Revolution' was born in the 1970s (Castells et al., 2000) - a decade that saw the invention of the micro-processor and micro-computer, and also the development of the first electronic communications network (ARPAnet) and subsequently TCP/IP, the method of connecting networks to each other which is still employed by the Internet to this day. It was however not until the early 1990s that the World Wide Web came into being, with the development of `HTML` and the first Web browser.

The pre-Web Internet was chiefly the domain of academics and the military as it required a computer to access and a fair degree of technical expertise to utilise effectively. The birth of the Web (and concurrent rapid decline of the cost of computers) marks the point where the Internet became much more widely accessible.

It is worth noting at this point that the 'birth of the Web' did not require any significant alteration of the Internet's infrastructure, it was merely a case of developing a standard method of 'encoding'

documents (`HTML`) and referencing these documents (URLs), coupled with a piece of software which could decode said documents (the Web browser). These elements sat atop the existing Internet, and while the Web is so widely used that it is almost synonymous with the Internet for many people, in a technical sense it is just one way of using the Internet. The Web itself was in a sense created by Internet users, it required no formal approval or mandate to be put into effect and its success is due to its utility and popularity with users. These are themes which we will re-visit when we come to discuss the development of more advanced Web applications (collectively referred to as 'Web 2.0' applications (O'Reilly, 2005)).

First, let us consider some sociological theories on how the Internet would affect society which were based on the early Web of static web pages and e-mail.

2.2   Sociological perspectives on the Internet

The surge in Internet use in the early 1990s which accompanied the birth of the Web triggered much sociological theorising about the social implications of this new communications medium - in particular its capacity to bring about social change. A brief overview of some of these theories is offered below.

In fact such sociological theories pre-date widespread Internet use. Bell (1977) was the first sociologist to envision the coming of digital communications technology and theorise about the effects its widespread use would have on society. More recently, Castells et al. (2000) has established himself as a leader of sociological thought on the Internet's capacity to transform aspects of society. Castells et al. (2000) argues that we have entered the 'Information Age', that new digital communications media are transforming every aspect of society - facilitating the emergence of a global economy, transforming the nature of work and influencing our ideas about the Self.

Sociologists of a Marxist persuasion would tend to view the Internet as another communications medium which could be co-opted by the elite to enhance control of politics and production (Davis et al., 1997). From a Weberian perspective the Internet is a point-to-point medium which could advance rationalisation by reducing limits of time and space (Collins, 1979). Both Marxists and Weberians place particular emphasis on inequality of access to the Internet - commonly referred to as the 'Digital Divide'. This divide exists on a number of levels. Internet accessibility is unevenly distributed across the globe - with the international divide between countries where the Internet was available and affordable being highly correlated with national wealth in the 90s (Hargittai, 1999). Within nations, access to the Internet and the capacity to utilise it effectively are unevenly distributed - with education level and income found to be key factors by Robinson et al. (2002), and Norris (2001) finding that younger people, men, the highly educated and highly affluent were the most likely demographics to be online. Norris (2001) also compares the diffusion of Internet

use with that of previous communications technologies and finds that the spread of Internet access follows a similar pattern to previous communications technologies. If the Internet is to become central to an individual's capacity to obtain information/eduction, build and maintain networks of social support, and engage in political dialogue - then inequality of access is likely to reinforce existing social inequalities related to income and education.

Technological determinists, unsurprisingly, would see social change induced by the Internet as a result of enabling new communications practices with corresponding skill-sets; as with previous socially transformative technologies (such as the printing press) those individuals who are able to adapt well to these new communications practices will have an advantage (Eisenstein, 1979). Some early scholars of new digital communications media foresaw the coming of an 'information society' which would replace the 'industrial society' (Bell, 1977). Critical theorists were chiefly concerned with the effect of new communications technologies on public deliberation and the integrity of civil society (Habermas et al., 1989; Calhoun, 1998).

While broad macro-sociological theories such as these are informative for the present research it is beyond the scope of this research to evaluate these; although some of these perspectives will be re-visited below. Rather we have opted to focus on a particular aspect of Internet use ('Social Media' which employ Distributed Moderation systems) and develop a detailed understanding of this with empirical observations. DiMaggio et al. (2001) highlighted some opportunities for sociological research afforded by the Internet, emphasising the opportunity to study this new communication medium while it develops, and the opportunities for new kinds of research afforded by the nature of the Internet. This call seems to have fallen largely on deaf ears, recent reviews by Farrell and Petersen (2010) and Murthy (2008) documenting a lack of sociological research which utilises primary data collected on the Internet.

While sociologists have appeared reluctant to engage in new kinds of research made possible by the Internet, other groups have showed no such hesitation. Private companies, Google in particular, have created a whole industry built on utilising data stored on the Web - chiefly to create profiles of users which are useful for market researchers. Google's acquisition of `Youtube` is particularly interesting in this regard (Van Dijck, 2009). Google paid $1.6 billion for `Youtube`, a website which was tremendously popular but not profitable. Google already had their own (arguably superior) video sharing website (Google Video); so it seems unlikely that they were attracted by `Youtube`'s technological infrastructure. What Google got for their money was the `Youtube` user-base, and one of the ways they have endeavored to turn a profit from `Youtube` has involved making use of the immense quantities of information `Youtube` users generate in their everyday use of the website.

Academic researchers, and in particular social scientists, have lagged behind private industry in their willingness to embrace the collection and analysis of the vast array of 'procedural' data which Social Media websites produce on a daily basis. Addressing this gap in the sociological literature

is but one reason to study newly emerging forms of interaction on-line. There are further, and arguably more compelling, reasons concerning the specifics of how Internet use has changed over the last decade.

2.3   Political perspectives on the Internet

The idea that the Internet would have a radical impact on politics is almost as old as the Internet itself and pre-dates the widespread adoption of the World Wide Web. For example, Rheingold (1993), Naisbitt (1991) and Corrado and Firestone (1996) all make grand claims about the democratizing potential of the Internet based on early observations of discussion spaces such as bulletin boards and `Usenet`. There are several dominant themes to much of the research on the Internet from a political perspective. Chief among these is the tendancy for scholars to fall (or be placed into) one of two camps, those who believe that the Internet will revolutionise politics such that it is barely recognisable (most often thought to involve a shift to a much more direct or participatory form) and those who believe that familiar political institutions and procedures would adapt to this new technology resulting in 'Politics as Usual' (Margolis and Resnick, 2000). Much of the political science research on the Internet can be thought of as trying to answer this question of whether or not it would cause a 'revolution' in politics.

Wright (2012) puts forward a highly compelling argument that this 'Revolution'/'Normalization' dichotomy has actually hampered our understanding of how the Internet and politics interact. To borrow some of Wright's examples, this dichotomy has lead to the dismissal of changes in public behaviour because they are not as significant or instantaneous as the term 'Revolution' implies (e.g. a finding that 44% of Americans read political blogs several times per year or more becomes "More Than Half of Americans Never Read Political Blogs" - and "Only 11% of bloggers in one survey said that the primary topic of their blogs is politics or political issues" (Davis, 2009: 4) is the summary given to a finding that implies there are 14.5 million primarily political blogs). Conversely, those who believe a revolution is underway may exaggerate the significance of minor developments - for example *The Guardian* newspaper citing Barack Obama's plan to email and text supporters directly about his choice of running-mate as an example of "the power of technology to transform electoral politics" (Wright, 2012).

A second theme to political science research on the Internet is that it has tended to focus on aspects of Internet use which are explicitly political in nature.

> "Virtually all research has, until very recently, focused on established political events (e.g. elections), institutions (e.g. parliament/party web-sites), activities (e.g. government-run online consultations) and actors (e.g. elected representatives blogs)." (Wright, 2012, p. 255)

Wright sees this as a result of the prevalence of the revolution/normalization frame which much of the research has adopted, and calls for more research on what he terms 'Third Spaces'. These 'Third Spaces' are online discussion fora "with a primarily non-political focus but where political talk emerges within conversations. The key link between participants is not (normally) their location but specific issues or topics" (Wright, 2012, p. 255).

If we consider contemporary political actors' presence on the Internet it is characterised by the uptake of tools, platforms and repetoires of behaviour which have already been established elsewhere, often long before they are adopted by the government or politicians. If the Internet is to cause a 'revolution in politics' I would argue that this is most likely stem from forms of communication which are in their infancy or have not yet been devised. The web is still bustling with experimentation and innovation - but governments and established political parties seem unlikely to be the source of a communication technology which could 'revolutionise' politics.

The present research addresses the political impact of the Internet not as its primary subject but at a tangent. Social Media which employs DM voting systems were selected *a priori* as the subject of this research because they appeared to the researcher as a novel and interesting form of mass communication and there was a dearth of research aiming to understand how they function. The prominence of 'voting' as a means of interaction on these websites immediately suggests a link to the political, but it was never the intention of this research to determine whether Distributed Moderation would be part of a revolution in politics - the revolution/normalization debate is not one which will be entered into here. Rather this research confines itself to describing how current examplars of this form may be interacting with the political - and speculating on how this technology might be adapted for more explicitly political purposes in future.

The Social News website `reddit.com` seems to fit with the description of a 'Third Space' quite neatly - it is not a website dedicated to politics but does see a considerable amount of political discussion. I would argue that `reddit` is a particularly interesting 'Third Space' to study for two broad reasons. The first of these is the centrality of voting and the novel form of 'democratically mediated discussion' which takes place on `reddit`. The second relates to `reddit`'s output, a crowd-sourced 'News' type resource which is conceptually interesting given the prominent role of News media in democratic societies - and which during the course of this research became almost directly involved in the political process of the United States (through the campaign against the 'Stop Online Piracy Act', see chapter 7).

2.4   The changeable nature of Internet use

'Internet use' is itself not a stable concept, the profile of individual users' online behaviours is quite different now to what it would have been ten years ago when Internet use was largely characterised

by the use of e-mail and accessing static web-pages, often through privately owned and run 'portals'. The term 'Web 2.0' has been coined (O'Reilly, 2005) to describe a range of new applications which are 'native' to the Web (i.e. could not conceivably function outside a medium such as the Internet) - by 2013 'Social Media' seems to have replaced 'Web 2.0' as the blanket term for such websites. Many of today's most popular Web 2.0 applications (e.g. `Youtube`, `Wikipedia`, `Flickr`) are entirely dedicated to 'user-generated content'. This represents a major departure from the earlier Web, where in order to become an information producer one had to have a web site or page, and placing content on these websites was usually limited to their owners or administrators. Figure 2.1 illustrates a rapid shift in user traffic away from web portals (`MSN` and `Yahoo`) and towards user content driven sites (`Facebook` and `Youtube`).
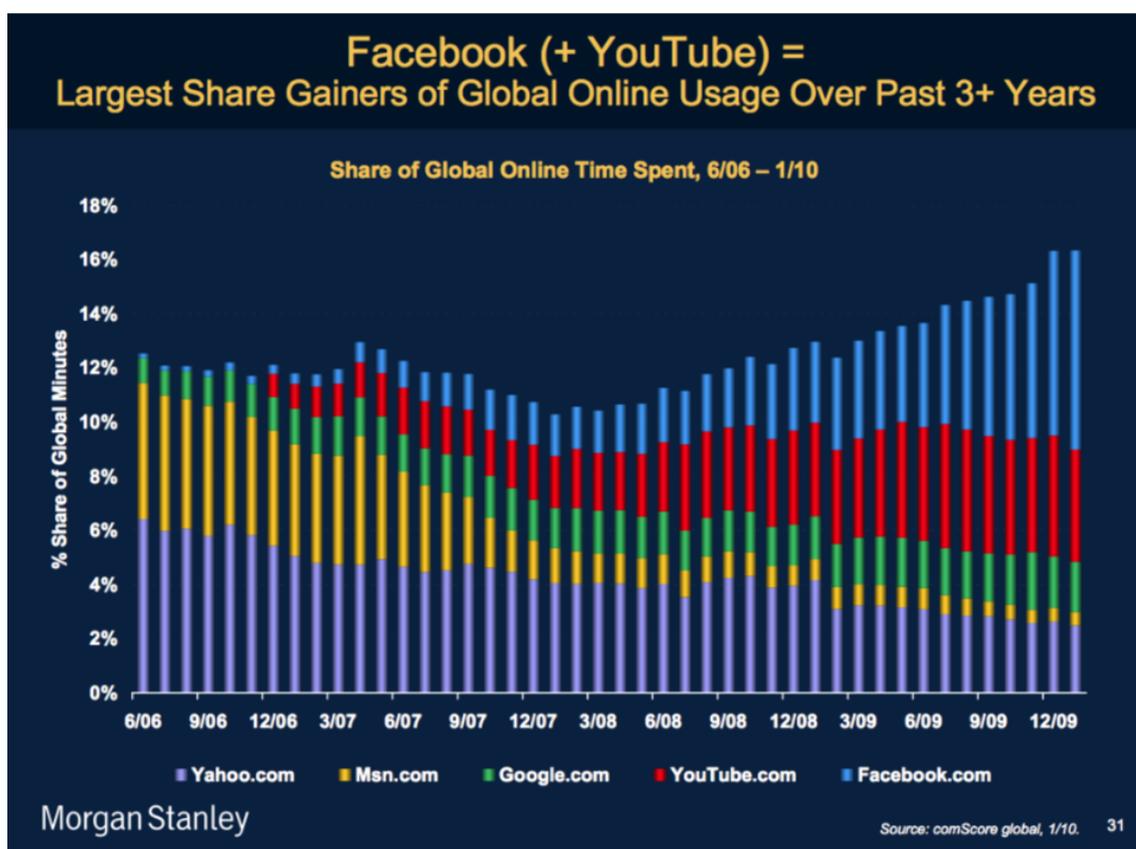
Figure 2.1: The share of global Internet users' time spent on five major websites between 2006 and 2010; taken from Meeker (2009)

Furthermore, it seems likely that the nature of Internet use will continue to evolve in unpredictable ways in the future as people continue to experiment with new methods of interaction which the Internet makes possible. Let us take `Twitter` as an example of this; the service was launched in 2006 but didn't begin to 'take off' until 2007, when 400,000 tweets per quarter were being sent, this increased to 100 million tweets per quarter in 2008, 2 billion tweets per quarter by the end of 2009 and 4 billion per quarter at the start of 2010 (Beaumont, 2010). In 2010 `Twitter` is becoming increasingly ingrained not only in our web use but in television broadcasts; live programmes now

often invite their viewers to talk back through `Twitter` and typically read some of these messages during the programme.

Services such as `Twitter` and `Facebook` are widely perceived as being integral to several recent political uprisings in the middle east (Stone and Cohen, 2009; Grossman, 2009; Gaffney, 2010). It is a hallmark of the speed at which new Information Communication Technologies (ICTs) are developing that a web service like `Twitter` can move from its infancy to widespread adoption in the space of just five years; and from there proceed to play a role in the shaping of a nation's politics.

The adoption and diffusion of new technologies is itself the subject of a considerable volume of research. Concepts such as the technology adoption lifecycle (Rogers, 1966) and Gartner's hype cycle (Linden and Fenn, 2003; Fenn et al., 2009) incorporate some form of adoption threshold, beyond which the ultimate success of a given technology is highly likely. Furthermore, a new technology which has passed this threshold can reach a kind of critical mass; at which point it is highly likely that said technology will become firmly embedded within the societies where it has been adopted. Such concepts are also applicable to the uptake of new web applications, including the various forms of Social Media.

If we consider academic research on these new Social Media; this tends to be conducted and published when a given application (e.g. `Twitter`) is well on its way to widespread popularity, or has already reached the point at which its widespread adoption and subsequent embedding in society is all but assured. If this research uncovers potentially negative effects related to the uptake of its subject application, by the time these are reported it is effectively too late to halt or slow the widespread adoption of the application. For example, Turkle (2010) cautions strongly against heavy use of Social Media as a replacement for face-to-face human interaction, citing a number of negative side effects. However, by the time this book was published many individuals had incorporated these applications into their lives to such an extent that they are not easily abandoned.

If we are to consider the Internet's long-term impact on society it seems prudent to pay particular attention to the newest forms of social interaction which are emerging online. This is one of the primary reasons for focusing on Distributed Moderation systems in the present research. When this research began `reddit.com` was a relatively small website; eighteen months later it had grown to become one of the top 100 websites globally in terms of web traffic (Alexa, 2010). This is indicative of the problems social scientists face when attempting to keep pace with the rapid development of novel web applications. 'Social News' applications have probably already passed the threshold where they are 'here to stay' - before anyone has developed a solid understanding of how they operate, how this affects the content they produce and the individuals who participate.

In addition to the youth of these websites, there are several bodies of literature which serve to highlight the importance of studying Distributed Moderation systems; and in particular their use in the context of 'Social News'. Firstly, we will consider the distribution of attention on the Web, with particular reference to the problems of 'Information Overload' (section 2.5). We will then consider Web 2.0 applications which foster 'Peer Production' efforts (section 2.6). It is now possible for a text to be produced by a group without any formal hierarchical organisation, and this text can be widely distributed without the need for capital (beyond an Internet-connected computer). These two features of the modern Web stand in stark contrast to the 'industrial' method of production and information dissemination through the mass media. A further perspective of utility here is to look at the convergence of 'old' and 'new' media systems; Jenkins (2006) argues in favour of research which studies the intersection of old and new media, rather than comparing and contrasting each as separate entities. Again, Social Media applications which employ Distributed Moderation systems are of particular interest here (section 2.7). Social News websites process content which originates from old and new media, produced by professionals and amateurs. All of this content is 'processed' by Social News websites in the same way, irrespective of its source, and the voting system through which it is processed is fundamentally social in nature - being powered by the voting behaviour of individuals, and being accompanied by commentary from these individuals; this commentary itself also being filtered and sorted through the voting system such that it represents a kind of democratically mediated discussion.

2.5   Attention distribution on the World Wide Web

2.5.1   Links on the World Wide Web

Hyper-links are integral to the functioning of the World Wide Web, and have been the subject of studies by researchers from a variety of disciplines (e.g. Gonzalez-Bailon (2009); Adamic and Adar (2003); Barabasi and Albert (1999); Albert et al. (1999); Flake et al. (2002)). Links can be either internal (linking to another page on the same website, e.g. a navigation bar) or external (linking to a page on a different website). Of these two types of link external links have been the subject of more research; where links are referred to these are external links unless otherwise stipulated.

At their most basic level links provide a path from one web page to another; by using a link the user is transported from the source page to the destination page. In this way links serve as the sign-posts of the web - if a page on Website A links to a page on Website B then individuals who view Website A will learn of the existence of Website B - and be presented with a path to Website B which they can easily follow. This means that every visitor to the page on Website A bearing the hyper-link has the potential to also become a visitor to Website B. This is one of two major

ways in which links influence the distribution of user attention between websites on the Internet.

External links also serve a second purpose in the functioning of the web - they are used by search engines such as Google (Brin and Page, 1998) and others (Tomlin, 2003) as an indication of a website's quality or importance. A website with more in-links will appear in a more prominent position than a website with fewer in-links; for a given search term that matches both websites.

The algorithms behind most of the popular search engines weight these in-links according to the centrality of the sites on which the links appear (Tomlin, 2003). For example, if `Yahoo.com` links to a page on Website A this should significantly increase the prominence with which Website A appears in search results for relevant terms. In contrast, a link to Website A from an individual's low-traffic home-page will have very little effect on Website A's search prominence. The importance of a website's ranking by search engines is illustrated in a study (Silverstein et al., 1999) which reported that 85% of users only looked at the first page of search results provided in response to their query.

The function of links on the web has been likened to that performed by citations for academic papers (Gonzalez-Bailon, 2009). Each citation brings with it the possibility that a reader will "follow" the citation and subsequently read the paper which is referenced. Each citation also adds one to the recipient article's citation count - with this citation count being used as a proxy for article importance or quality in some cases.

The distribution of citations among authors (Newman, 2001), and of links between websites (Barabasi and Albert, 1999; Barabási et al., 2000; Huberman and Adamic, 1999); have both been shown to be scale-free with a long tail. In both cases the distribution is highly skewed such that a very small number of authors or websites receive a disproportionately large number of citations or in-links. The Power Law is one distribution of this highly skewed type which has received a lot of attention from researchers in recent years - due to its utility in describing a range of natural (e.g. earthquake magnitude - Gutenberg and Richter (1944), number of species in biological taxa - Willis and Yule (1922) , solar flare intensity - Lu and Hamilton (1991)) and man-made (e.g. city populations - Newman (2005), individuals' annual income - Pareto (1896), frequency of word use in human language - Zipf (1949)) phenomena (for review see Mitzenmacher (2004); Newman (2005); Clauset et al. (2009)).

Power Laws and the distribution of attention on the World Wide Web

The pure power-law distribution, also known as the zeta distribution, or discrete Pareto distribution is expressed as

$$p(k) = \frac{k^{-\gamma}}{\zeta(\gamma)},$$

where:

- k is a positive integer usually measuring some variable of interest, e.g., number of links per network node;

- $p(k)$ is the probability of observing with value k;

- γ is the power-law exponent;

- $\zeta(\gamma)$ is the Riemann zeta function defined as $\sum\limits_{k=1}^{\infty} k^{-\gamma}$.

Many research articles have been published which document observations of the power-law in a range of Internet-related measures. Early examples of such work focused on the distribution of links between websites (Barabasi and Albert, 1999; Barabási et al., 2000; Broder et al., 2000). More recently the same skewed distribution of links has been observed within certain sub-sections of the web: in the blogosphere (Kumar et al., 2005), and in social commerce networks (Stephen and Toubia, 2009). Power law distributions have also been observed in a range of Internet phenomena unrelated to links; such as the size of files transferred (Willinger and Paxson, 1998) and the size of e-mail address books (Newman et al., 2002).

A power law distribution has also been observed in the distribution of hits (individual page views) between websites for `AOL` customers on a single day (Adamic and Huberman, 2000). In principle, hits seem to offer a better measure of user attention than links because they count the number of times a page is accessed; whereas the presence of a link does not guarantee that any specific number of individuals will follow it. There are however several problems with the use of hits as a measure of user attention. A hit is generated every time a page is loaded, irrespective of whether a person or web-crawling bot made the request, and irrespective of how many times the same person or bot has previously requested the page.

To circumvent these problems, a measure of the number of unique IP addresses to visit a page is frequently used, in place of hits, as a measure of attention. This measure has its own problems however; the same individual is counted multiple times if they visit the page from different networks, and conversely all individuals to access a page from within a single network can be counted as a single unique visitor. For these reasons, research relating to the distribution of attention on the Internet frequently relies on measures other than hits to infer attention.

Much of the research which has observed and described Power Law distributions in nature and on the Web has used a method involving linear regressions on log-transformed variables (frequency counts, for example the number of times a word appears in a corpus) to determine whether they follow the power law, and to estimate its parameters.

Criticisms have recently been leveled at the use of graphs and linear models to demonstrate that certain data-sets follow the power-law. Goldstein et al. (2004) and Clauset et al. (2009) both argue that this methodology is biased with regard to detecting power law distributions and estimating their parameters. Both of these papers call for a shift to Maximum Likelihood Estimation in research dealing with the power-law. Clauset et al. (2009) in particular offers several pieces of convincing evidence that Maximum Likelihood produces more reliable estimates of power law parameters.

The power law is relevant to this research in that it describes many aspects of web use - often referred to as the 80/20 rule (Trueswell, 1969) or the long tail (Anderson, 2008). This research is concerned with developing an understanding of what this kind of highly skewed distribution means for the functioning of the Distributed Moderation websites in question. By extension, this research deals more with the social implications of power law distributions in man-made phenomena than with determination of whether a given data-set follows a power law distribution or not (and how this distribution might be produced). The highly skewed distribution of user attention between websites on the Internet has implications for its use and development; irrespective of whether the power law or some other highly skewed distribution provides the best fit to data representing this trend.

There follows a short description of some mechanisms and models which have been proposed as potentially underlying a power law distribution of links on the web.

Mechanisms underlying a Power Law distribution of links on the Web

The network topology of the World Wide Web has been the subject of many studies - some of the most influential have applied graph theory to create models which produce a power law distribution of links. Barabasi and Albert (1999) proposed a mechanism termed 'preferential attachment' whereby new websites added to the network would preferentially link to existing websites with an already high number of in-links. This kind of 'rich get richer' mechanism is enough to produce a power law distribution in a network of constant growth such as the Internet.

The Barabasi and Albert (1999) model suggests a systematic relationship between a website's age and its number of in-links; and by this model it is very unlikely that a new website added to the Internet in 2010 would stand much chance of being seen and linked to. While a relationship between website age and in-links does exist (Huberman and Adamic, 1999) it is not as strong as the model of Barabasi and Albert (1999) would suggest. In a study of internal links Huberman and Adamic (1999) found that adding an intrinsic growth rate as an attribute of each website significantly improved the fit of their model. A model with both a preferential attachment mechanism and intrinsic growth rates for each website seems to offer the best explanation of the how the Internet's

link topology emerges.

While graph theory has produced robust models which can explain how a skewed distribution of links between websites could arise - these do little to illuminate the factors which underly individual link formations. From a graph theory perspective the motivations for an individual author on one website to create a link to another website are a black box. One study to address this was conducted by Gonzalez-Bailon (2009) with a sample of websites from the `.org` domain. Gonzalez-Bailon (2009) looked at attributes of the organisations which owned the websites in their sample and assessed them in terms of their access to resources and perceived status - revealing a positive relationship between both resources/status and number of in-links.

Other studies (Kumar et al., 1999; Flake et al., 2002; Adamic and Glance, 2005) have looked at linking patterns within and between 'communities' of websites; and used the presence of such linking structures to infer the existence of communities. Kumar et al. (1999) and Flake et al. (2002) found that they could reliably identify communities by looking at densely inter-connected regions of the web. Adamic and Glance (2005) looked at the political blogosphere and found that both liberal and conservative blogs were much more likely to link to blogs on the same side of this distinction than across the divide.

While hyper-links embedded in static web pages have been the subject of considerable research to date; social link sharing (via e-mail or Social Media) has received relatively little attention. It seems highly unlikely that this kind of personal link dissemination will follow the same (almost strategic) principles that govern when one organisation or blog will link to another. With the development of 'Web 2.0' (O'Reilly, 2005) social link sharing is likely to become more significant. A service such as `Twitter` enables its users to share a link to any web resource with all of their followers quickly and easily. Indeed several 'url shortening' services have sprung up to facilitate link sharing within `Twitter`'s 140 character per tweet limit (Wauters, 2009).

Social News websites also facilitate the sharing of links and channeling of attention; but they do so in a way which is quite distinct from the social networking approach of `Twitter` or `Facebook`. The users of a Social News website, through its Distributed Moderation (voting) interface, make a single collective decision about which of the many links submitted will reach the Social News website's Front page - and consequently a large proportion of its users.

2.5.2 Social significance of the web's attention distribution

The biggest difference between the Internet and the mass media is that the barrier to entry as an information producer is orders of magnitude lower on the Internet. The nature of mass media is such that there is a clear distinction between producers and consumers - with only a tiny fraction of the population able to act as information producers. To become an information broadcaster in the

mass media requires ownership or control of expensive broadcasting equipment; and there is often an upper limit on the number of broadcasters who can exist in a given medium (e.g. the number of analogue television channels which could exist in a given region was limited by the technology itself). This produced a situation whereby the owners of broadcast media acted as gatekeepers to their audience.

This distinction between producers and consumers of information is much less useful in describing Internet use. From the Internet's birth it has been relatively easy to become an information producer (by hosting one's own website). The financial and technical barriers to doing this have always been relatively low, and they have decreased steadily throughout the Internet's development (e.g. the advent of blogging software).

The more recent development of 'Web 2.0' (O'Reilly, 2005) has blurred this distinction between information producers and consumers further still. Many popular websites now allow members of the public to contribute in certain areas (these contributions are most usually comments). Furthermore, a lot of the most successful new websites of recent years (e.g. `Youtube`, `Wikipedia`, `Twitter`, `Facebook`) have been entirely dedicated to content produced by end-users rather than staff or administrators. This development is heavily implicated in the rapidly increasing quantity of information available on the Internet - estimated at 8 million pages in 1999 (Lawrence and Giles, 2000) and 11.5 billion pages in 2005 (Gulli and Signorini, 2005).

In such an environment of information abundance - attention becomes the more important resource (Goldhaber, 1997). One of the key criticisms of the Internet's democratising effects centres around this issue of 'Information Overload' (Noam, 2003) - when everyone speaks at once nobody can be heard; and there is no value in being able to speak if nobody can hear you. There are several variants of this particular critique; each with a different view on the consequences of Information Overload.

Noam (2003) suggests that in this environment of information abundance money would come to be the determining factor in whether one's own contributions to the web would be seen; those with the capacity to pay for advertising would come to dominate the attention of Internet users. The implications of this perspective are that despite the radically different structure of the Internet, in practice its usage patterns would be quite similar to those of the mass media. If money is required to reach an audience then the capacity to reach an audience will be limited to the individuals or organisations with finances to pay for advertising; a route to an audience which is also available in print and broadcast media.

Sunstein (2002) cautions against the dangers of excessive filtering of the information we receive - and argues for the societal importance of individuals encountering information which they would not have sought in advance. The Internet allows individuals to decide, with a great degree of

specificity, which information resources they wish to consume - and with this capacity comes a number of pit-falls for a democratic society. Firstly, it is assumed that many individuals who tailor the information (materials, topics, positions) they are exposed to will do so in a way that fits with their existing interests and opinions. This is likely to result in many enclaves of like-minded individuals, and within these enclaves polarization is likely to occur whereby the opinions of group members become more extreme in the absence of dissenting voices (Myers and Lamm, 1976). A second problem with the tailoring of information consumption is that it will decrease the extent to which members of society have shared experiences - these shared experiences being important to promoting a sense of common identity among members of a society.

Sunstein (2002) draws on the historical importance of 'public forum doctrine', whereby individuals have a right to speak freely in public locations, in ensuring that people were exposed to opinions and ideas which they would not have sought out for themselves. Mass media is also seen as important in the role of 'general interest intermediary'. When there were only three television channels, television itself served to strengthen the social fabric by ensuring that a large number of individuals were paying attention to the same sources of information, the content of which is determined by the 'general interest intermediary' and does not necessarily attempt to match the pre-existing interests or opinions of viewers. The proliferation of many and varied digital television channels has weakened this unifying aspect of television - but the Internet has the potential to facilitate fragmentation to a much greater degree, depending on how individuals use it. The vast array of highly specialised websites afford new opportunities for individuals to tailor the information they receive, 'the Daily Me' is a term which refers to an individual's highly customised set of information inputs - it is akin to a newspaper which only shows readers a set of stories which interest them specifically and which match their pre-existing opinions. Furthermore, the Internet allows us to not only read information from such a tailored set of sources, but to join discussion groups consisting of like-minded individuals.

The implications of fragmented public discourse are overwhelmingly negative for democracy and the public sphere. A lack of common experiences and familiarity with the views of others will make it more difficult for individuals to relate to others who do not share their views/experiences. At a societal level if there is no shared understanding of what the issues of the day are, this will severely limit the public's capacity to rally around some cause and bring about social change.

2.5.3 Social implications of the highly skewed distribution of attention on the web

A second generation of criticisms focused on the way people actually used the Internet. It has long been established that user attention on the Internet is highly concentrated (Adamic and Huberman,

2000) and many studies (Barabasi and Albert, 1999; Barabási et al., 2000; Broder et al., 2000) suggest that the distribution of attention between websites is highly skewed and may follow a power law. These usage patterns go some way to refuting concerns about the fragmentation of public discourse (at this stage) - but this high degree of centralisation suggests that in practice the Internet may not be very dissimilar to the mass media model of information dissemination.

A power law distribution of attention between websites has both advantages and disadvantages in terms of the Internet's potential to democratise society. A major advantage is that websites at the very high end of the user attention distribution serve as focal points; two randomly chosen individuals can have a somewhat shared understanding of 'what's on the web' because they will most likely visit some of the same websites - this is important if the web is to serve as a platform for public discourse. This concentration of user attention can also be thought of as a strategy to deal with the problem of information overload (one looks at the same sites as other people); although in these terms it is a somewhat basic approach.

The major disadvantage of a power law distribution of user attention is that many sites and pages on the web receive little or no attention; in practice the Internet is more similar to mass media than its open structure would suggest. These usage patterns do nothing to refute the concern (Noam, 2003) that money would ultimately be required to reach an audience on the web. There are however reasons why this situation may be more desirable than the mass media. These centre around the role performed by websites in the medium-to-low end of the user attention distribution, and the mechanisms whereby the attention a website receives increases and decreases.

Although there are very few websites which receive attention from a lot of users, there are many websites which receive attention from a modest number of users. According to Benkler (2006, p242), at an aggregated level these moderately attended-to websites may be performing a critical role:

> "Filtering, accreditation, synthesis and salience are created through a system of peer review by information affinity groups, topical or interest based. These groups filter the observations and opinions of an enormous range of people, and transmit those that pass local peer review to broader groups and ultimately to the polity more broadly, without recourse to market-based points of control over the information flow." (Benkler, 2006, p242)

Money can be used to buy attention on the web in various ways (e.g. banner advertisements and 'AdWords'), but financial backing is not the sole determinant of whether a given web page will be seen. It is possible for a new website with no advertising budget to garner a lot of attention quickly if it is linked to by any established site and is sufficiently interesting that early visitors use word-of-mouth (or e-mail) to tell those they know of its existence. The Blogosphere seems particularly

suited to performing this role of spreading the word about new and interesting websites (Benkler, 2006, p243); and as we shall see, performing this role closely matches the original purpose of Social News websites.

The highly skewed distribution of attention on the web means that highly attended-to content can still form a basis for public discourse. The open nature of the web means that anyone can participate in this discourse as an information producer - and while many of these contributions may go unseen, the web allows many more people to reach an audience than the mass media. The processes which determine whether a given website will climb to the high end of the user attention distribution are themselves quite complex and worthy of further study. Money can certainly play a part in this but its role seems much more limited than in the mass media.

While the highly skewed distribution of user attention has persisted throughout the web's history to the present; the last five years have seen a marked shift in the type of websites which exist at the high end of the user attention distribution. The early Internet was dominated by 'Portals' (e.g. `Yahoo` and `MSN`) and these have close parallels to the TV stations/networks of the mass media - in that the owner of each portal decides exactly what appears there; and has capacity to charge companies for placement on their portal and the user attention this would generate. In the last five years these portals have seen a steady decline in the user attention they garner; portals have been supplanted at the highest end of the user attention distribution by websites which are themselves social in nature (e.g. `Facebook` and `Youtube` - Meeker (2009)). The capacity these social websites offer users to easily become information producers is directly implicated in the rapidly increasing amount of content on the web (e.g. there are 24 hours of video content uploaded to `Youtube` every minute, with about 50% receiving at least one comment or rating) (Daisley, 2010).

A website like `Youtube` does little to combat the problem of information overload; if anything it exacerbates this problem. There are however other forms of Social Media, such as Social News sites (and also new services like `Twitter` - Kwak et al. (2010)) which can be used to mitigate information overload. A Social News website's purpose is not to add more content to the web; rather their primary function is one of signposting existing information which users might not otherwise see. The short message nature of `Twitter` means that it is also frequently used as a means of directing one's followers' attention towards some existing web resource. The role these new types of website may come to play in directing user attention on the web is one aspect which makes them of interest to social scientists.

Services like `Twitter` and Social News websites already cause a problem for the kind of network topology models which concentrate on the presence or absence of links (e.g. Barabasi and Albert (1999)). As the owner of hypothetical Website A, I can create an account on both `Twitter` and a Social News website, and within minutes have links to Website A appear on these high-traffic sources. The problem with this approach is that unless I have 'followers' (Twitter), or my link

receives a positive voting response (Social News) - it is not likely to be seen by any significant number of people. This poses a problem for our understanding of the distribution of attention on the web, but it also raises an opportunity - because the process whereby a link becomes widely seen (or neglected) on these services is both formally and rigidly defined in the website's software infrastructure.

2.5.4 Social Media and the channeling of their users' attention

Some forms of Social Media map quite neatly unto the theories of Internet-enabled public discourse outlined above. Social Networking based applications like `Facebook` and `Twitter` seem, on the surface, more likely to lead to a fragmentation of public discourse. This is because an individual user receives information only from those users they have decided to 'friend' or 'follow'. One `Twitter` user might choose to follow the feeds of celebrities, while another might choose to follow the feeds of scientists - the same service can provide radically different types of information to different users. If an individual's pre-existing perspective informs their decision about which other users they will receive information from through these applications, then this may produce a feedback or reinforcement effect - the individual is presented with opinions which tend to agree with their own and this may make their own opinions more extreme (Myers and Lamm, 1976).

Social News websites seem, in principle, more similar to broadcast media; in that every user at a given point in time is presented with more or less the same information (i.e. the contents of the website's Front page). The major departure from mass media is that the decision about which items will be 'broadcast' rests on social (the voting behaviour of users) rather than economic (the capacity to pay or otherwise generate revenue) factors. However, this rather simplistic overview glosses over some important subtleties of Social News.

Firstly, Social News websites allow their users to subscribe to or unsubscribe from categories of content. If for example a user finds `reddit.com`'s political coverage objectionable, said user can unsubscribe from the 'politics' category while still enjoying humorous images in the 'pics' category; or the user could seek out more right-leaning political discourse in the lower-traffic 'libertarian' category. Secondly, there is more than one Social News website. The hypothetical user above who found `reddit`'s political coverage objectionable could seek out an alternative Social News website where popular content in the 'politics' category is more in line with their own opinions. Determining whether such behaviours occur among Social News users is of central concern if we are to asses the impact of these websites on public discourse.

In principle, one can see reasons why DM as applied on Social News may have the potential to work against the 'fragmentation of public discourse' (Sunstein, 2002). The voting system operates to select a small number of items to make highly visible, distilling 'shared experiences' from the web for

their 'community' of users, which in turn serve as a shared basis for future discourse. Determining the extent to which a website like `reddit` has its own sense of community is important for this research, because the presence of a strong community among its large and expanding user-base would itself be evidence that its 'vote-to-broadcast' model works against fragmentation. However, as with all communications technology, the effect on public discourse will be determined by the way in which people use it over and above the nature of the technology itself (Sunstein, 2002). Social News websites also offer users a means of tuning out content which they are not interested in - and if users engage in this type of fine-tuning extensively Social News may prove to be a 'distributor of the Daily Me'.

A related and equally important question concerns the effect of Distributed Moderation on the discourse which takes place on Social News websites. Every item of content on these websites is subjected to the voting system, and to reach any substantial audience an item must fare relatively well through this system. Therefore, an important question to ask is whether these voting systems are capable of representing conflicting or minority points of view. The importance of discussion spaces incorporating discussants with multiple and contrasting views has been well established (Sunstein, 2002; Habermas et al., 1989). Social News websites admit any user and consequently all points of view to the discussion, but their users collectively have the potential to exclude or silence perspectives which dissent from the majority.

Let us take the example of `Wikileaks`' release of diplomatic cables in late 2010, a subject which saw considerable coverage through conventional and Social Media. If 60% of `reddit`'s users are pro-`Wikileaks`, theoretically this majority group could down-vote any submissions which they perceive as anti-`Wikileaks`; with the result that items appearing in high-visibility spaces would be uniformly positive or at worst neutral in regard to `Wikileaks`. In this hypothetical situation would the 40% of users who are anti-`Wikileaks` continue to participate in the discourse on `reddit.com`? It seems unlikely that users would continue to post anti-`Wikileaks` content once it became clear that such content always received a negative rating, and consequently never reached a substantial audience. Would these users go further and stop using the 'politics' category, or stop using `reddit` entirely?

To put these questions in a more general form:

- Are Distributed Moderation systems capable of representing views held by a minority of their users?

- Do users of a Distributed Moderation system expect to agree with all or most of the judgments made through the system?

- Do users who disagree with these judgments drop out or migrate to an alternative website?

The precise manner in which users' opinions and behaviours influence the composition of a Social

News Front page is of particular importance from a fragmentation of discourse perspective. Revisiting the `Wikileaks` example, if the 60% majority of `Wikileaks` supporters eliminate competing views from highly visible locations then reddit is effectively a pro-`Wikileaks` echo-chamber, with all of the negative consequences thereof (Sunstein, 2002), *despite* the fact that its users are not homogenous in that regard. `Reddit` is not, *a priori*, a pro-`Wikileaks` website - but depending on how its users vote it may appear and behave as such.

Initial observations suggest that `reddit`'s voting system is capable of expressing more than one side of an argument; and that while minority opinions may be marginalised to some degree they are not filtered out completely as the above example hypothesises. The effect of Distributed Moderation on discourse will be thoroughly investigated in Chapter 7; and due to the reasons cited above this constitutes one of the most important areas regarding the social significance of Social News.

2.6 Distributed Moderation as Peer Production

The term 'Peer Production' is one which has seen a recent rise in popularity among scholars of the Internet; and it relates to one of the more sociologically exciting uses of the contemporary web. Peer Production encompasses any endeavour where a (potentially large) number of individuals work autonomously together on some project (usually the production of an information resource). These endeavours lack a hierarchical structure to their organisation; individuals who participate in the production effort do so ostensibly as equals, hence the term *peer* production. Successful Peer Production endeavours are remarkable because they accommodate individuals' freedom to adopt and switch between various roles, even drop in and out of the production effort. Shirky (2008) goes as far as suggesting that this emerging phenomenon of Peer Production represents a newfound capacity for people to organise without an organisation; and without the overheads associated with a hierarchical organisation (i.e. layers of management).

In recent years a 'new consensus' has been emerging among scholars of Peer Production (Kreiss et al., 2011). The position taken by scholars such as Benkler (2002, 2006), Jenkins (2006), Leadbeater and Powell (2008), Gillmor (2004), Tapscott and Williams (2008), Van Dijck (2009), Shirky (2008) and Surowiecki (2005) - is that social production and the technology which enables it has the potential to democratise aspects of culture and society; and threatens hierarchical and bureaucratic methods of organisation (for critique see Kreiss et al. (2011)).

A fundamental tenant of this 'new consensus' is that new digital communications technology is enabling collaboration between large numbers of individuals without the requirement of a hierarchical organisation or contractual commitments. Open Source software and `Wikipedia` are two commonly cited examples of Peer Production. These examples demonstrate that an assortment of loosely connected individuals can produce information goods (software and an encyclopedia re-

spectively) which rival or surpass those produced on a for-profit basis by conventional companies (e.g. Giles (2005); Kittur and Kraut (2008)) . The most salient advantage of this production method is its cost-effectiveness - peer produced resources cost so little to produce that they are usually offered to the public for free.

This characteristic of Peer Production takes on even greater significance when we consider Peer Production in the domain of News. An informed citizenry is critical to a functioning democracy (Mattson, 1998), and many scholars have expressed criticisms of the mass media in this regard (e.g. Benkler (2006); Herman and Chomsky (1988)). Peer Production, due to its low cost and high number of participants, has the potential to be much more resistant to the influence of money than production which is for-profit and hierarchically organised. For instance, it is commonly accepted that newspapers adopt and promote positions on the stories they cover, and that these positions are determined at quite a high level of the organisation (Glasgow Media Group (1976)). This has most recently been illustrated by Rupert Murdoch's statement to the Leveson inquiry that "If politicians want my views they should read Sun editorials" (Guardian, 2012c). This statement supports the notion that the owner(s) of a conventional News organisation have power to determine aspects of how that News organisation covers certain stories.

Peer Production endeavours also tend to be driven by amateurs rather than professionals, and again in the domain of News this is likely to result in a product which bears some systematic differences to that which has been produced by a conventional News organisation. Within conventional News organisations professional journalists determine which stories are to be covered as News and produce the articles or segments which embody this coverage. This class of professional journalists have been trained to interpret and represent the world in a particular way - and all conventional News is produced along these lines. Broadcast News media outlets in particular strive to produce News coverage which appears balanced and objective (Glasgow Media Group (1976)). The importance of News coverage to a democratic society means that this 'journalistic perspective' exerts a strong influence on the framing of social issues. Social News websites, by allowing amateurs to enter the domain of News production, may produce News coverage which deviates from the conventional News media in both the decisions about which stories warrant coverage and the nature of this coverage. The specifics of how News coverage produced by amateurs might differ to that produced by professionals are not well understood because before the advent of the Internet and Social Media it was simply not possible for an amateur to produce News coverage and have this coverage reach a significant audience.

Social News websites differ from other forms of Peer Production in that they focus more on collective decision-making than collaboration. Each post or comment submitted to a Social News website is 'owned' by its author; only the original author can edit or delete the item. Rather than editing each other's content Social News users vote to dictate the prominence each item will be afforded.

As such, the activity is better characterised as distributed decision-making (through a democratic interface). Social News websites do however share other hallmarks of Peer Production. Users are autonomous in deciding whether/when/how they will participate, and all user accounts (with the exception of a small number of moderators) are capable of the same actions (voting, submitting and commenting); with votes from all users carrying the same weight. Social News websites also produce a resource; a ranked and sorted list of the content which has been submitted, epitomised by the Front page - and this resource is accessed by many individuals who do not participate. `Reddit.com` recently passed something of a milestone in this regard, serving more than one billion page views in the month of January 2011 (Reddit-Blog, 2011a).

It is helpful at this stage to contrast 'Social News' websites with conventional news resources such as newspapers. Users who submit content can be thought of as journalists, and users who vote collectively fulfill the role of an editor. In this analogy the website's Front page represents the editor's decision about which of the many submitted items warrant publication. The major disparities between the workings of social and conventional news are as follows. On a Social News website many of the items have been found rather than produced by the submitting user, with the submitting user merely giving the item a title and deciding which category to submit it to. On a Social News website every user can act as both journalist and editor, by submitting and voting respectively. On a Social News website each user acts autonomously - the 'editor' cannot dictate what the 'journalists' submit; although text posts about the kind of items which do or do not belong on the Front page are quite common on some Social News websites.

The power to decide which of the submitted items are 'newsworthy' is distributed between all of a Social News website's voting users. Conceptually, this should make it more difficult for vested interests to surreptitiously influence the coverage of a given story on a Social News website. For example, a newspaper could be persuaded not to cover a given story or to cover it in a certain way, by persuading the newspaper's editor. This task is made much more difficult when the role of editor is being fulfilled collectively by thousands of loosely connected voting users. The task in this case is more akin to persuading a population - where that population can only be reached on their 'own turf', and no special consideration will be paid to the fact that an individual represents a large company or other powerful entity.

This does not mean that a Social News website's decision about which stories to cover is immune to external influence, and it certainly does not imply that vested interests will not attempt to do so. The most obvious way to have a Social News website drop a particular story would be to approach the website's administrators; as they have the power to delete or hide items. There are several instances where this has occurred.

In May 2007 the encryption code for HD-DVDs was cracked; and some `Digg.com` users began submitting posts which linked to the encryption key. `Digg`'s administrators were approached

and asked to remove these posts on the grounds that they represented copyright infringement (Greenberg, 2007). `Digg`'s CEO at the time acquiesced to these demands, and wrote a blog post explaining that any posts making reference to the encryption key would be deleted, in order to avoid opening `Digg` up to lawsuits which could ultimately bring the website down. `Digg` users however did not comply with this request, instead flooding `Digg` with new submissions about the encryption key and voting posts about it up to the `Digg` Front page. Eventually `Digg`'s administrators conceded that they could not prevent their users from posting and publicising the encryption key.

A similar story unfolded on `reddit.com` in August 2009 - a `reddit` user discovered an exploit on the `Sears.com` website whereby the url could be modified to change the description of products (Masnick, 2009). Sears contacted `reddit`'s administrators and requested that the offending post be removed - `reddit`'s administrators complied with this request. The removal of this post did not go unnoticed, and `reddit` users began submitting posts which were highly critical of both `Sears.com` and the `reddit` administrators; these posts occupying most of the website's Front page for several hours. In this instance `reddit`'s administrators did not have to back down, the issue eventually subsided when `Sears.com` fixed the exploit on their website.

These examples serve to illustrate the power which Social News users wield, and the extremely negative reaction which occurs among these users when their power is seen to be diminished. When a website like `Digg` cannot adequately respond to a request to remove illegal content from its Front page, this suggests that influencing a Social News website's coverage of a story through the administrators is an ineffective strategy and that the chances of doing so 'behind closed doors' are minimal. If a Social News website's administrators struggle to comply with a request backed by the threat of litigation it seems unlikely that they would respond to such requests to influence coverage where laws are not being broken. Even if an administrator wished to comply with such a request, it seems that to do so they would risk a back-lash from users and significant harm to the long-term prospects of their website.

The robustness of a Social News website's coverage to external influences does however depend on the equality of its users' actions through the voting system. While influencing the behaviour of the thousands of users involved in voting to determine a story's prominence seems futile - if some of these users held much more influence over the website's collective decisions than others, then these users might represent a weak point which could be targeted to more subtly influence the coverage on the Social News website. The use of exactly such a strategy has in fact been documented on `reddit`, one 'Social Media Marketing' company has approached high-profile `reddit` users to solicit their involvement in 'gaming' reddit's voting system, offering payment for doing so (redd.it/elqgs).[1]

---

[1]This is a short-form url which links directly to a post's comments page on `reddit`, the 5 or 6 character string following the slash is the post's unique ID code.

Reddit's community of users are aware of such strategies, and there are several instances of 'citizen journalism' where a `reddit` user has uncovered attempts to manipulate the website's voting system (e.g. redd.it/gepnl ; redd.it/glgy8 ; redd.it/gjt26 ... and `reddit.com/r/reportthespammers` is a whole subreddit devoted to this kind of activity). These issues will be returned to in greater detail in section 3.4.

2.7 The convergence of 'old' and 'new' media systems

Jenkins (2006) has an interesting take on the convergence of old and new media. Many scholars who contrast the characteristics of old and new media seem to assume that new media will supplant old media, whereas Jenkins devotes a lot of thought to the interplay between old and new media. Jenkins' focus is on popular culture, and he identifies many contemporary examples where fans take mass media cultural products and appropriate these for their own purposes (e.g. Harry Potter fan fiction, amateur Star Wars movies). It is common for Internet communities to form around such activities, and Jenkins recounts stories about several of these. One particularly interesting example is the 'Survivor Sucks' bulletin board, where a group of fans began by discussing the show and sharing predictions of what would happen in future episodes. Over time this 'spoiling' activity became more popular, and the group would collaborate (by sharing evidence or tips) on this task of predicting events as the series unfolded. Individual group members took on different roles and varying levels of involvement, and the group developed informal protocols to direct how the 'spoiling' activity should be conducted. This group exhibited a form of Collective Intelligence (in that the group was much more effective than any of its individual members) facilitated by Bulletin Boards, a relatively simple method of online communication that has been available for almost as long as the Internet itself.

It is Jenkins' contention that such knowledge sharing and collaborative processes are equally applicable to the political sphere. These practices arise in the cultural domain because they have intrinsic entertainment value - for a fan of Harry Potter or Survivor they are first and foremost a way of extending the cultural product's domain, reaping more entertainment from something which is already entertaining. Whether this activity produces anything of merit is in a sense a side issue, people 'play' in these spaces because it is fun. Jenkins argues that by engaging in this kind of play we (particularly younger people) are learning how to participate in online collectives, and also what these collective endeavours are capable of. The capacity to gather, share, and make sense of evidence as a group has obvious applications in the political sphere - it seems logical to suggest that when people are confident in this kind of participation we will begin to see more groups who apply their collective intellect to more political issues.

Much more recently, and with the benefit of seven years to watch these trends develop, Chadwick

(2013) re-visits many of the themes of 'convergence culture', or in his formulation the 'hybrid media system'. Chadwick observes that some of the assumptions of 'convergence' thinking have proved (at this stage) erroneous. Firstly, the assumption that all media will inevitably be drawn into a single delivery platform has thus far not been validated, people still consume media through a range of platforms (e.g. television, print, personal computer, mobile phone). The 'hybrid media system' sees these platforms and their media as co-evolving to compete with, compliment and sit alongside one another, rather than 'converging' into a single delivery platform.

Secondly, Jenkins (2006) often equates 'online' with 'grassroots', whereas recent years have been characterised by the emergence of 'online media elites', the involvement of these elites in the production of older media, and the increasing presence of older media entities online. Thirdly, where Jenkins saw a shift from simultaneous consumption (when an item was broadcast) to asynchronous participation (consumers viewing an item and subsequently engaging in discussion or other forms of related online participation at the time(s) of their choosing) - instead the recent trend is in the opposite direction. That is to say that 'viewers' of live programming now often begin discussing the content as it is being broadcast (Anstead, 2011), simultaneously consuming the broadcast and producing a response (on another screen/device/window) - raising the possibility that their contributions can be fed back into the live broadcast, and offering a viewer-produced commentary alongside the live broadcast. For example, during the first live Prime Ministerial debate of the UK General Election is 2010, 211,000 Twitter messages about the debate were produced by 47,420 distinct users - with these tweets being subjected in real-time to a 'sentiment tracker', and some of them being shown on screen as part of ITV's coverage (Chadwick, 2013, p.76) and/or being re-tweeted to reach ultimately large audiences themselves.

> "But when the real-time communication affordances of social media suddenly emerged in the mid-2000s, this pushed things in a different direction, and in ways which simultaneously enhance the popularity and legitimacy of both online and broadcast media forms." (Chadwick, 2013, p. 58)

Chadwick (2013) focuses almost exclusively on politics, and what the 'hybrid media system' means for political discourse and the shaping/exercising of political power. Chadwick makes extensive use of the concepts of 'political information cycles' and the capacity for older and newer media actors to "routinely engage in loosely coupled assemblages in the pursuit of new information that will propel a news story forward and increase its newsworthiness" (Chadwick, 2013, p. 88)

> "Political information cycles work on the basis of cross-platform iteration and recursion. This serves to loosen the grip of journalistic and political elites through the creation of fluid opportunity structures with greater scope for timely intervention by online citizen activists... The combination of news professionals' dominance and the integration of non-elite actors in the constuction and contestation of news at multiple points in

a political information cycle's lifespan are important characteristics of contemporary political communication." (Chadwick, 2013, p. 64)

Chadwick (2013) is also careful to distinguish the 'pockets of intense engagement' where the interplay between older and newer media is at the forefront, from concepts like 'crowdsourcing' or 'the wisdom of the crowd'. The hybrid media system as it relates to politics is still dominated by older media - and where newer media play a role it is most often through elite actors from this realm (e.g. well known bloggers) interacting with those of older media. The hybrid media system does allow (some) non-elite actors to be involved in 'making the news', but the capacity to do so requires one to have some special insight into a story and, crucially, to capitalise on this insight at an appropriate time. In the examples Chadwick (2013) cites where this has occurred, these non-elite actors have had their impact only through the information they provided being put before an elite actor (e.g. a tweet about a blog post which was circulated on twitter and then seen by a Channel 4 news reporter).

"At the same time, however, we should not lose sight of the fact that ordinary citizens, using digital technologies that enable them to cross from the outside to the inside of the elite political-media nexus, may now, on occasion, affect the meanings and flows of political information in new ways." (Chadwick, 2013, p. 88)

The other side of this is relationship is realised by elite actors from the conventional media who now devote some of their attention to the latest social media content - either to gauge reaction to an item of breaking news (Chadwick, 2013, ch8) or to glean tip-offs about a developing story from the investigative work of a non-journalist blogger - as seems to have occurred between Krishnan Guru Murthy and Sarah Nuttall (via a re-tweeting intermediary) (Chadwick, 2011, p.13). These events occurred in 2011/2012, and by then certain bloggers were well known as influencers and `Twitter` had become a firmly embedded branch of new media.

As this research has unfolded, the Social News genre and `reddit` in particular may well have been making its debut in the hybrid media system - unfortunately we are at this stage without the excellent and insightful type of behind-the-scenes view of the way older media elites perceive their latest new media rival (which Chadwick amply provides through interviews with elite media actors).

Chadwick (2013) also discusses organizational hybridity and the conflicts of 'old media logic' with 'new media logic', particularly in the context of the Wikileaks organization and its collaboration with the *Guardian, New York Times, Der Spiegel* and a range of other elite old media actors. This will be considered further in section 7.1, where a case study is presented concerning `reddit`'s coverage of the Wikileaks release of diplomatic communications in 2010.

2.8 Distributed Moderation, the Commons, and non-zero-sum games

The DM systems considered by the present research are fundamentally public entities - practically all of the information submitted to these websites is publicly accessible, and members of the public have the capacity to participate in the running of the DM system. In the case of Stackoverflow, user question and answer submissions are actually created under a Creative Commons license (Stack Overflow, 2011).

There is a considerable body of research on the use and administration of Common Pool Resources - much of this concerns physical resources which can be depleted through over-use, the 'tragedy of the commons' (Gordon, 1954). More recently, Ostrom (1990) has led the development of a whole interdisciplinary field of research into the management of Common Pool Resources (CPRs). Ostrom (1990) provides evidence of communities who successfully manage their CPRs (e.g. irrigation systems, forrests) to avoid depletion without an external authority.

Ostrom identifies several characteristics of communities who were able to successfully manage their CPR:

- Group boundaries are clearly defined.

- Rules governing the use of collective goods are well matched to local needs and conditions.

- Most individuals affected by these rules can participate in modifying the rules.

- The right of community members to devise their own rules is respected by external authorities.

- A system for monitoring members' behaviour exists; the community members themselves undertake this monitoring.

- A graduated system of sanctions is used.

- Community members have access to low-cost conflict resolution mechanisms.

Ostrom (1990) also called for the newly emerging field of CPR research to work towards the "empirically-supported theory of self-organizing and self-governing forms of collective action" (p. 25). It is through this angle that CPR research has relevance the newly emerging Social Media ecosystem - where the resources in question are not physical and easily depletable but intangible. Other researchers have also begun extending CPR research to human-created CPRs emerging on the web and facilitated by technological systems (Hess and Ostrom, 2003).

The first question we must ask in applying CPR thinking to Distributed Moderation is 'what are the common pool resources at the centre of DM systems?'. In the case of Social News there are

two candidates. The Front page, representing the outcomes of that day's content discovery and rating by users, is very much the resource which Social News is designed to produce, but can it be considered as a Common Pool Resource? The Front page does not offer a good conceptual fit with a CPR because it is the *product* of the Social News site, and the raw materials of its production (submissions and votes) can be supplied without limit by users who wish to participate in the peer production effort. While the 'quality' of this resource will be determined by the specifics of user participation, as long as there are users who participate there will always be a Front page because it is produced automatically by the site's software infrastructure.

Rheingold (2003) suggests that the CPR at the heart of many-to-many communications online is human attention - and while this makes a lot of sense on Social News sites it has some counter-intuitive corollaries. In this view, the Social News Front page can be thought of as a well of human attention - when a user submits a post which appears on the Front page this receives (or consumes) a proportion of the attention currently focused there. Each submission can be thought of as an attempt to draw from this well, with the success of the attempt being determined by other users through the voting system.

In this framing the 'free riders' of Social News would be individuals who repeatedly submit posts (in an attempt to 'consume' attention) without paying attention to others' posts (not putting anything back into the well) or contributing to the administration of the resource (by voting). This framing sees 'lurkers', who only browse and do not participate, as contributing to the CPR without making any attempt to consume the resource themselves. Such a positive appraisal of 'lurkers' is counter-intuitive, as in many other contexts it is these lurkers who are seen as the most egregious free riders.

With human attention as their CPR, Social News sites are clearly host to a 'non-zero-sum' game. Each additional user which these websites attract increases the sum total of available attention. Even users who come to a Social News site with the intention of 'free riding' and consuming the attention of others will invariably contribute some quantity of their own attention to the common pool. The manner in which Social News sites apportion their attentional resource is determined by their software infrastructure (with the centrality of the Front page indicating an almost 'winner takes all' solution) - the recipients of this attention (or 'winners') are determined by the collective voting behaviour of users on a daily basis.

For Question & Answer sites like Stackoverflow, which generate and store knowledge, the CPR is likely of a different nature. In this case the resource the site produces is the capacity to find the answer to a question. Stackoverflow is producing an archive of programming knowledge and as this grows so too does the likelihood that a user will find that their question has already been asked and answered. In addition, there are a certain number of users actively answering new questions on the site - the question-answering capacity and time of these users is the CPR, one

which could potentially be depleted. If the number of users with questions to ask exceeds the number of users who are answering these questions this would represent depletion of the CPR. In this sense, Stackoverflow's capacity to build a navigable archive of previously answered questions is important as a strategy to avoid depletion of the CPR.

2.9 The Democratisation of the Public Sphere

The public sphere is an aspect of social life in which the informal discussions of small groups shape the formation of public opinion. The concept of a public sphere was coined by Habermas et al. (1989) in relation to the interactions taking place in 18th-century coffee shops and salons. Here individuals (of a certain class and gender) would meet to engage in critical-rational discussion of issues of common interest (political, cultural and economic). An important characteristic of these discussions was that they were free to contradict or criticise the official position of the state in which they took place. The outcomes of these discussions would circulate among the coffee shops and beyond, and the prevailing opinion in these venues would come to be seen as representing the public's opinion on a matter. Once public opinion had been established in this fashion it could then go on to influence the decision-making of the state. A functioning public sphere is therefore an important part of a participatory democracy.

Since its original conception our understanding of the public sphere has become more nuanced. Dahlgren (2005) breaks the public sphere down into three dimensions: structures, representation and interaction - and he describes it broadly as follows:

> "In schematic terms, a functioning public sphere is understood as a constellation of communicative spaces in society that permit the circulation of information, ideas, debates (ideally in an unfettered manner) and also the formation of political will (i.e., public opinion). These spaces, in which the mass media and now, more recently, the newer interactive media figure prominently, also serve to facilitate communicative links between citizens and the power holders of society." (Dahlgren, 2005, p. 148)

Other scholars have proposed sub-types of public sphere, with Eriksen (2005) describing three variants of the public sphere which have been cited by other scholars. The first of these, the general (sometimes referred to as weak) public sphere is most similar to the original Habermasian public sphere. It is characterised by informality of communication, ambiguity of purpose and openness of participation - although external actors such as the state may place limits on or guide this public towards certain topics, or even dictate who can participate or at what level individuals can participate.

A second type of public, the strong public, is quite different in nature to the general public sphere.

A strong public sphere has a direct and sometimes explicitly defined relation to the political system (e.g. parliamentary debates or working groups could be described as strong publics). These publics are characterised by purposeful and often regulated deliberation, they are generally not openly accessible to all and the right to participate in the deliberation is limited to a select few.

The third type of public which Eriksen refers to is a segmented (or transnational) public - these are publics which 'spring up' in response to specific issues (and might also be termed 'issue publics'). Eriksen (2005) is writing in the context of the European Union, and gives examples including the European campaigns against racism and the scandal about the handling of BSE as events which have created transnational public spheres. Eriksen sees this as a form of "*elite communication, where the experts and the well-educated speak to one another and stage communicative noise and protest*" (Eriksen, 2005, p. 13).

A simplistic reading of Eriksen (2005) is that the general European Union public sphere is so weak as to be almost non-existant, spaces for this discourse are limited and under-utilised. This lack of a general public sphere causes problems of legitimacy when we consider the European Union as a democratic society. Ideally the strong publics who in effect run the EU would take guidance and input from the general European public sphere - but this general public sphere does not exist in a useful form for this purpose. In its absence the strong public which is the European Parliament lacks accountability and when it does take direction from external sources it is from segmented public spheres which "fall way short of complying with the democratic proviso of openness and equal access" (Eriksen, 2005, p. 13).

The public sphere is therefore the weaving together of many constituent spaces which themselves may bear little similarity - incorporating both the physical and the virtual. One critical component of a functioning public sphere is the capacity for large-scale deliberation, and it is here where the Internet may be of particular utility. Many scholars have discussed the Internet's potential (specifically the potential of discussion fora) to meet this need - with some seeing great promise (Coleman and Gotze, 2001) and others fearing that the Internet will cause new problems (Sunstein, 2002). Bohman (2004, p. 57) argues that online discussion fora in their present form constitute 'weak' or general publics, but that they could become 'strong' publics if they can "exercise influence through institutionalized decision-making procedures with regularized opportunities for input".

Wright and Street (2007) point out the folly of lumping 'discussion fora' together - arguing that the design of these fora is critical in determining whether they will foster deliberation or result in cacophony. Wright suggests that the structure of a discussion forum can shape the nature of discussions which take place there in the same way that the design of physical spaces is thought to affect the discussions they host. Wright then pursues this question of how design influences discussion by looking at the European Union's 'Futurum' discussion forum - a space created with the goal of fostering pan-European debate about a European constitution. The nature of discussion

on `Futurum` was compared to `Usenet` and `AOL` discussion spaces and found to embody more of the characteristics of deliberation (i.e. discussants engaged more in seeking information and replying to the messages of other discussants) than the `AOL`/`Usenet` discussion spaces. This difference is then traced back to differences in the design of `Futurum` as compared to the other discussion spaces.

A second article by Wright (2007) concerning `Futurum` examined it in relation to the concept(s) of a public sphere and found that it did not fit neatly into any of the three categories proposed by Eriksen (2005). Messages on `Futurum` were 'pre-moderated' (i.e. they had to be approved by a moderator, in this case an employee of an EU institution, before they appeared to other users) and this kind of censorship means that it does not fit into the category of 'general' public. `Futurum` also had no direct relationship to the drafting of the actual EU constitution which occurred concurrently, and therefore it cannot be considered a strong public. The discussants on `Futurum` were thought to be un-representative of EU citizens (most likely giving voice to an already vocal minority) as a whole and this would raise issues with incorporating the discussions into the drafting of the actual constitution - but may also be one of the reasons why the discussion on `Futurum` was found to be more deliberative in nature than other venues considered.

Social News websites add an interesting dimension when thought of in the context of the public sphere. Poor (2005) argues that `Slashdot` (a Social News website by our definition) meets the criteria for a public sphere by Habermas' (1962) original definition of the term. Poor identifies four criteria for a public sphere - "that they are spaces of (often mediated) discourse, that they allow new (previously excluded) discussants, that the issues discussed are often political in nature, and that ideas are judged on their merit rather than the standing of the speaker". `Slashdot` is found to meet these criteria, and Poor concludes that it is best described as an 'issue public' because its users tend to be technologically adept and the topics discussed reflect this.

`Reddit` also meets these criteria of a public sphere, and it does so at a scale which is orders of magnitude greater than a coffee shop or salon or any other corporeal context in which this kind of discourse might take place. Let us consider `reddit` as analogous to a coffee shop, and each post as a discussion. Any individual can propose a topic (submit a post) and through the voting system users collectively determine which topics are worthy of discussion. For topics which are popular lively discussion takes place, over the course of a number of hours hundreds of individuals make comments and reply to the comments of others. These comments and the threads they are situated in are also subject to the voting system, with those comments which are more popular being placed in positions of greater visibility. If we include users who vote on these comments as discussants, then the number of individuals who participate in a discussion frequently numbers in the thousands. Furthermore, popular discussions are played out in front of an audience numbering in the tens of thousands, with all of these onlookers having the capacity to become directly involved in the discussion if they wish.

Given that the number of participants and onlookers for a discussion is much larger on `reddit` than in a coffee shop, we might therefore expect that the outcome of a particular discussion on `reddit` will have a more immediate and pronounced impact on 'public opinion'. This is not to say that `reddit` influences public opinion directly, but that the perspectives and rhetorical devices which rise to prominence on `reddit` will reach a sizeable audience very quickly - and from this point may or may not spread throughout the population in a more conventional fashion.

Suppose that a major and controversial news story breaks today which is widely discussed and has some ambiguity to how it should be interpreted. People discuss this story in a variety of fora, offline and online, including `reddit`. The discussion which occurs on `reddit` will have several thousand participants and several hundred thousand readers. Within a matter of hours this particular discussion will have concluded and those people who witnessed it may have been influenced by its outcome or some point or other which was made - going on to spread these points in further discussions they are involved in. Similarly, individuals may lift talking points which they hear from friends or in some other context and then submit these as part of the discussion on `reddit`. Depending on how well `reddit`'s voting system works in practice, the 'best' points which are made should rise to the top of the discussion page. The content of popular comments will in this manner be amplified two-fold - the individual's comment will first be broadcast to those `reddit` users who are interested in the discussion, and those individuals may then re-iterate the comment or a component thereof in another setting.

If we consider the same scenario without the Internet or Social News then it is likely that a point which is originally made in a coffee shop will take much longer to be circulated to a large enough audience that it can influence public opinion. Print or broadcast media will likely play a critical role in the spread of a unit of rhetoric, but to be broadcast in such a manner the point must first reach an individual with access to these media and their audiences (e.g. a professional journalist). If the gatekeepers of broadcast media who observe this unit of rhetoric were to disregard it then its chances of influencing public opinion would be minute.

If we compare the structure of Social News websites with discussion fora in general the most salient difference is the presence of a voting system on Social News sites. The application of a voting system is likely to result in Social News websites being able to operate at much larger scales than conventional discussion fora. Discussion boards typically do not weight the contributions to a discussion but instead present these in chronological order. While it is technically possible for hundreds of individuals to be involved in a discussion on an Internet forum, the thread which is produced will be very large and difficult to interpret - insightful contributions are mixed with the irrelevant and the reader must determine which is which. The same can be said for Internet chat-rooms, they could serve as arenas of the public sphere but do not allow for a significantly larger number of participants than their off-line equivalents.

There are of course many ways besides Social News websites and discussion fora in which information can be circulated on the Internet. Blogs are now thought to play a significant role in public discourse. While blogs usually allow readers to leave comments they cannot be thought of as a Habermasian public sphere because there is an inequality of participants in the discussion, a clear distinction between the blogger(s) and the readers.

Social network based applications such as `Twitter` and `Facebook` are also now thought to play a rapidly expanding role in public discourse, and these platforms are closer match with what would be expected of a public sphere than blogs. There is equality of participants in the sense that all `Twitter` accounts are created equal and discussions take place in public 'twitter space' rather than on one user's patch. In practice however, the visibility of a user's tweets are determined by their number of followers and this measure is highly skewed with a small number of participants having a very large number of followers. Hashtags mitigate this effect to some degree - for example if one wants to see discussions related to the Occupy Wall Street movement one can search for #OWS and be presented with all of the tweets which include this. However, these results do not constitute a single discussion but rather a stream of disjoint tweets. In fact it is debatable whether `Twitter` actually facilitates discussion in the conventional sense; contributions are limited to 140 characters and while it is possible to refer to other users with @ tags many tweets are simple statements or signposts addressed to the user's followers in general.

Social News websites are certainly interesting from the perspective of the public sphere and seem to offer something new in terms of the arenas in which the public sphere operates. The potential of Social News websites to augment or improve the public sphere rests on the practicalities of how Social News websites operate. If Social News websites really do allow thousands of people to participate in a discussion and simultaneously highlight the best parts of that discussion (as judged by the participants) then their impact could be immense. However it is perhaps more likely that there is a random element to the decision as to which contributions are promoted to salience in a discussion - perhaps good contributions go unnoticed or early contributions have a better chance to be seen.

To determine how useful Social News websites might be as an arena of the public sphere research into how they operate in practice is required. We can however speculate that Social News websites are characterised by the speed at which they operate, large-scale discussions are conducted in hours and broadcast as they occur. It is likely that with this frantic pace something is lost as compared to the bourgeoise public sphere of 18th century coffee shops or a discussion forum like `Futurum`. The smaller scale of the coffee shops and slower pace at which a consensus on an issue would spread in this kind of network likely allowed for greater deliberation and perhaps more reasoned outcomes. It may transpire that Social News websites are organs for the amplification of whatever opinions the majority of their users hold, rather than spaces where deliberation occurs and consensus is

reached.

2.9.1 Distributed Moderation and Politics - beyond the Public Sphere

Aside from the possible role of Social News websites as venues of the public sphere, Distributed Moderation systems may interact with the political system in other ways, two of which are outlined here.

The first of these relates to the importance of the press in the political process. Social News may give voice to constituents and perspectives which are not represented in mass media news outlets. Privately owned news companies have obligations to their owners or parent companies, most notably the imperative that they be profitable; but a privately owned news company may also be discouraged from running stories that harm the interests of their owners or parent companies (Glasgow Media Group (1976)). The hierarchical nature of a conventional news organisation facilitates the meeting of these obligations, there are individuals within the hierarchy with the responsibility and power to ensure that the goals of the organisation are being met. This stands in contrast to Social News, where everyday decisions about which stories warrant attention rest with the undifferentiated mass of users. These users do not rely on the company which owns the Social News website for their income, they are not beholden to the Social News website in any way. We might therefore expect that these users are free to make their decisions based on a personal assessment of the content's merits, without considering the possible repercussions of placing a certain story on the website's Front page for the Social News website's owners.

Conventional news organisations also rely almost exclusively on professional journalists to produce their content, and therefore all of the content which is broadcast tends to come from this class of professional journalists (Glasgow Media Group (1976)). This is in contrast to Social News websites, where content produced by professional journalists and published by conventional news organisations sits alongside opinion pieces or suggestions produced by the SN website's users.

The second way in which Distributed Moderation may democratise politics relates to the application of DM principles in the domain of political policy. Political parties are an important facet of western democracy, but these parties tend to be oligarchical rather than democratic in nature. Michels (1915) produced a well-regarded study of early 20th century political parties which focused on the German Social Democratic Party. Through this study Michels found that any large organisation, no matter how democratic its ideals, was forced to adopt a system of bureaucracy to maintain efficiency. This need for bureaucracy emerged from the fact that the organisation needed to make many decisions in its day-to-day operation, and that it was not practical to call on the party's members to vote on all of these decisions. Therefore it was necessary to appoint certain individuals to make these decisions (a centralisation of power); and to define the roles of

and relationships between these appointed individuals (with a bureaucracy). This had the effect of elevating appointed officials above rank-and-file members, forming an elite which was alienated from ordinary party members. As the party grew, and the number of decisions that had to be made increased, the roles of appointed officials became more specialised and in turn these officials delegated tasks to sub-ordinates whose experience led them to become even more specialised. This specialisation had the effect of furthering the distinction between party officials and members, entrenching leaders in their positions. As the party's bureaucracy became more complex, knowledge of how this bureaucracy operated became a necessity if one was to achieve anything through the party. This knowledge of how the bureaucracy operated was limited to those with first-hand experience, namely the officials who had been appointed to positions in which they operated within the bureaucracy.

Once these individuals had been elevated to a position of power within the party they were motivated, at least in part, to maintain or advance their own position - as opposed to thinking solely about what course of action would facilitate the meeting of the party's aims. These elite members had a degree of control over which information was disseminated to ordinary party members, and with this control they could push the party in a direction which suited their own personal ambition. The small group of appointed officials also had the power to decide which party members would subsequently be elevated to official positions within the party. They tended to promote individuals whose outlook matched their own and who would serve as allies within the party's organisation. Through these mechanisms the elite could effectively maintain control of the party beyond their own tenure. (Michels, 1915)

Michels believed that these characteristics of the German Social Democratic Party were universal to all political parties and in fact to all organisations, leading him to coin the term 'Iron Law of Oligarchy' - "Who says organization, says oligarchy." The Iron Law of Oligarchy implies a bleak prognosis for western democracy. It is not possible for a political party to accurately represent the will of its members. Instead the political party represents the will of its oligarchs who, believing themselves more knowledgable than the rank-and-file because politics is their profession, seek to shape rather than respond to the outlook of ordinary members.

Distributed Moderation systems address one of the key problems which necessitates the adoption of a bureaucracy in the first place; namely the problem of making daily decisions on behalf of a large and disorganised populace. To take `reddit` as an example, every day millions of votes are cast by ordinary users, the website's software aggregates these votes and through this process collective decisions are made about which of the many submitted items are to be placed in positions where they will be highly visible. The presence of bureaucracy is minimal, the same basic principle (that any item of content or comment thereon can be voted on by any user, and that these votes can be either positive or negative) applies across the whole website. There is also only a limited hierarchy

among users - with a very small number of users having the role of moderator or administrator, and these roles being limited to the enforcement of a number of generally-accepted rules. The vast majority of the activity which powers `reddit` consists of users with equivalent status casting votes through the ubiquitus voting interface.

`Reddit` avoids the path to oligarchy which is taken by conventional organisations because it is not an organisation in any conventional sense. Here we must distinguish between `reddit` the company and `reddit` the community of users. `Reddit` the company is very much a conventional organisation, but members of this organisation have the task of maintaining and developing the software infrastructure of `reddit.com` - providing a set of tools for `reddit`'s community of users. `Reddit` the community is characterised by a lack of organisation at the level of the individuals involved, with the voting interface being used to impose organisation or structure on the content which is submitted - without the necessity of allocating roles and positions to the users themselves. Despite this lack of organisation `reddit`'s users produce a ('news' type) resource which is more usually associated with conventional organisations. Furthermore, `reddit`'s rapid growth over the course of this research suggests that the resource which it produces is of some value or holds some appeal for a large number of readers and contributing users.

The relative simplicity of `reddit`'s decision-making infrastructure and the parity of its users can be counted as strengths in that they eliminate some of the mechanisms through which an oligarchy might emerge. However, one might imagine that this simplicity and lack of organisation comes at a cost. A simple voting system which applies to all content and which all users access as equals might be suited to making one kind of decision ('is this content worthy of the attention of `reddit`'s users?') - but can the same system be used to make higher-order decisions (e.g. about what `reddit`'s purpose is, or how it might be altered or improved)? Can `reddit`'s disorganised users make the same kind of strategic decisions which are required of political parties? Without hierarchical organisation could `reddit`'s users act as a unitary social entity which has the capacity to alter its behaviour in response to such a decision?

There is also the question of whether an elite of influential users might emerge on `reddit` despite its lacking strong forms of bureaucracy and hierarchy. Michels also cites psychological reasons for the emergence of an elite, with the masses desiring strong charismatic leaders. Might `reddit` users seek to elevate individuals to the position of leader despite working within a framework which does not require officially recognised leaders?

We conceive of `reddit`'s voting system as supplanting the role of a more complex bureaucracy, imposing order on what would otherwise be chaos at the level of items of content. In the same way that an organisation's bureaucracy priviliges those individuals with knowledge of how it operates - might `reddit`'s simplistic form of bureaucracy be subject to exploitation by those users who are more knowledgable of its workings?

Questions relating to the presence of elites on websites which employ Distributed Moderation will be addressed in section 6.1. The capacity of Distributed Moderation facilitated communities to behave as unitary social entities will be addressed in section 7.4. It is however worth observing here that Distributed Moderation systems are in their infancy, and that `reddit` or any other Social News website was not created for the purpose of facilitating this kind of higher-order decision-making. Any such capacity which these websites have to facilitate higher-order collective decision-making is merely a by-product of a system which was designed for a different purpose (ranking and sorting content). Efforts have been made to apply Distributed Moderation in the domain of political policy (see section 3.2.1) but these endeavours have either transplanted Social News mechanisms wholesale or provided a more limited set of tools to their users. When we consider Distributed Moderation's potential with regard to the democratisation of politics it is from a perspective of what these novel systems get right and how they might be tweaked so that they can be fruitfully deployed for this purpose in the future.

2.10 The Democratisation of Culture

Many of the arguments made here in relation to the Internet, Distributed Moderation and the production of News or information apply equally to the production of culture. The Internet has allowed many more idividuals to become producers of cultural artefacts (e.g. art, music, videos) and to make their creations available to others. A Social News website like `reddit` can act as a method of channeling attention for these amateur productions.

There are many examples where individuals have posted some of their work or something they have produced on `reddit` and this has gone on to receive a lot of attention (examples are scattered throughout the thesis). In one case a user posted about their game of 'Civilization II' (an empire-building strategy game) which they had been playing for ten years (redd.it/uxpil). This post appeared on the `reddit` Front page and because it was shown to such a wide audience there, and captured the imagination of some, several conventional news organisations subsequently published pieces on this individual's game - e.g. Jordison (2012). This individual has sparked a discussion about the future (specifically about the possibility of an Orwellian dystopia) which has not only permeated `reddit` but appeared in a number of highly recognisable news outlets - all from their observations on a game which they had been playing. It is difficult to imagine this eventuality (where so many people were thinking about and discussing what Civilization had to say about an Orwellian future based on one individual's save-game) *without* a website such as `reddit` in the mix, allowing anyone to submit anything and picking out the 'most interesting' submissions.

Reddit has also helped to promote services with a more explicit economic component. The 'Humble Indie Bundle' (where a selection of independently developed games are bundled together and offered

on a 'pay what you like' basis) has had several very successful iterations and relied initially on `reddit` (where a post about it appeared on the Front page) as its primary means of generating publicity (source: email correspondence).

There are also many subreddits which cater to various types of 'user-generated content' - for example image macro 'memes' (where users take a standard background which is associated with a particular type of message and overlay their own text) on /r/AdviceAnimals, ghost stories on /r/nosleep, 'rage comics' on /r/fffffffuuuuuuuuuuuuu/. Subreddits such as these receive thousands of submissions every day and many more votes, with the votes serving to determine which of the day's submissions are 'the best' and will be displayed for all to see at the top of the subreddit. For a subreddit like /r/fffffffuuuuuuuuuuuuu/, which deals with a form of product (the 'rage comic', a short comic strip recounting some frustrating incident from the creator's life) that was itself popularised on `reddit`, the subreddit also serves as a place to discuss the specifics of how one makes a 'rage comic', share the materials from which a 'rage comic' is composed, discuss the characteristics of a good 'rage comic' and decide on which criteria a post must have to be considered a valid 'rage comic' (e.g. redd.it/n1kuw).

Over the course of this research the prevalence of user-generated content on `reddit` has increased substantially. In terms of the democratisation of culture, Reddit serves as an alternative route between producers and consumers of culture and in this sense is performing a role similar to that of a publishing company. Any individual can submit a link to their 'product' on `reddit` and its merits will be judged by other users through the voting system, with those which achieve the best voting response being shown to a very large number of people. In contrast, it may require pre-existing status or connections to even have one's product reviewed by a conventional publishing company - with items that *are* reviewed being judged by a small number of professionals and those which are rejected failing to reach the public at all. On `reddit` this process is much more open and results in a non-binary outcome - many users who submit an item of their own work will have it seen by a reasonable number of users and receive some feedback from this audience, even if their post does not ultimately appear on the Front page.

2.11   Is 'Democratisation' a good thing?

The capacity for individuals to participate as producers of information and culture is generally seen as positive among researchers of related disciplines - the position of authors cited in this work tends to be one of optimism. In some cases this optimism is guarded, with concerns that top-down or bottom-up processes could nullify the Internet's potential or lead to undesirable outcomes. For example, Sunstein (2002) cautions against the Daily Me and fragmentation of public discourse resulting from individuals' freedom to choose and filter the information they receive.

Others (Rheingold, 2003; Benkler, 2006) worry that the Internet's potential will be lost through the corruption of its open and accessible architcture (over-regulation by government, control exercised by Internet Service Providers). There are also works of exhuberant optimism, for example Reese (2013)'s "Infinite Progress: How the Internet and Technology Will End Ignorance, Disease, Poverty, Hunger and War".

It is however important to devote some consideration to the opposing perspective - that the Internet (and Social Media in particular) will erode support for conventionally produced information and culture, and that this will be supplanted by a vastly inferior resources produced by amateurs. One popular work to espouse such a position is Keen (2007)'s "The Cult of the Amateur" - wherein "democratization, despite its lofty idealization, is undermining truth, souring civic discourse, and belittling expertise, experience and talent" (Keen, 2007, p. 15).

This book contains some valid points, in particular about the affect of anonymity online and the erosion of established information and culture production industries when these are forced to compete with resources produced by amateurs and offered for free. The death of newspapers in particular would be a troubling development because it is not clear that the amateur 'citizen journalists' and bloggers who might take their place would have the capacity to provide resources of comparable quality.

To its credit, the book is also one of the first to remark upon Social News websites like Digg and Reddit.

> "Reddit is a mirror of our most banal interests. It makes a mockery of traditional news
> media and turns current events into a childish game of Trivial Pursuit" (Keen, 2007,
> p. 6)

Keen's awareness of `reddit` pre-dates my own, and I cannot speak to whether this quote is an accurate description of the site in 2006/2007. In reddit's defence, I would say that 1) `reddit`, especially at that time, was not attempting to produce a news resource which would serve as a viable alternative to traditional news, 2) the site is fundamentally democratic, it shows users the items they collectively choose as worthy of attention, 3) with the proliferation of subreddits, and the increasing number which have considerable levels of user activity, `reddit` is capable of providing a resource similar to the conventional news media (or in fact a highly specialised news resource about scientific disciplines, sports or games) if users choose their subreddit subscriptions to derive such a resource, 4) the configuration of the default Front page is not an attempt to provide a 'news' type resource but to showcase the many and varied forms of intereaction which the site caters to.

Keen returns to Social News websites later in the book, decrying these as "limit[ing] our access to fair and balanced information" (p94).

"For all their claims to be more democratic and honest, these supposedly editor-free social news sites are actually creating a more oligarchic and corrupt media" (Keen, 2007, p. 94)

In support of this claim Keen refers to a *Wall Street Journal* article which reported small groups of 'influencers' who successfully manipulated Social News - in Digg's case one third of observed Front page postings were submitted by 30 highly infulential users. This is a valid concern, because the domination of Social News websites by small groups of users would suggest that these users had found an effective means to manipulate the site(s) - thus seriously diminishing their democratic appeal. If Social News sites are merely mouthpieces for small groups of users who have found a way to exploit the voting system to their own ends then they lose much what makes them interesting - with the 'right to speak' through a Social News site merely being allocated on a basis other than ownership or profession. This issue is re-visited in the context of research on Social News in section 3.4.1, and is addressed by the present research with data from `reddit` in section 6.1.6.

However, Keen (2007)'s "The Cult of the Amateur" and the arguments therein are weakened by an obvious bias in favour of the traditional model of information/culture production and against new/social media and the amateurs who produce these (terminology including 'monkeys' and 'fools'). Some examples:

"few of us have special training, knowledge, or hands-on experience to generate any kind of real perspective." (Keen, 2007, p. 30)

One of many examples where Keen dismisses the perspectives of those who are not professionals in a field as having no value.

"Todays editors, technicians, and cultural gatekeepers the experts across an array of fields are necessary to help us sift through whats important and whats not, what is credible from what is unreliable, what is worth spending our time on as opposed to the white noise which can be safely ignored." (Keen, 2007, p. 30)

Nowhere does Keen acknowledge that gatekeepers, as they "sift through whats important and whats not" actually exercise a great deal of power by setting the agenda for the public sphere, or that the agenda determined by professionals working for privately-run for-profit companies might be distorted in any way.

"Unlike professionally edited newspapers or magazines where the political slant of the paper is restricted to the op-ed page, the majority of blogs make radical, sweeping statements without evidence or substantitation." (Keen, 2007, p. 53)

This is another example of a naieve perspective on conventional news - where a newspaper's choice of which stories to cover or the position these stories take on an event/issue might also

be influenced by the organisation's political bias. It is also itself a 'sweeping statement without evidence or substantitation', dismissing 'the majority' of blogs without providing any source or evidence. Keen also does not allow for the possibility that conventional news organisations, in an effort to compete with social media in the fast pace of the '24-hour news cycle', will begin to take social media content as the inputs for their own news production, and might themselves become more susceptible to the publishing of unreliable information (Chadwick, 2013). Chadwick (2011) provides one example where the 'tables were turned', as conventional news producers reported uncritically on accusations from the 'National Bullying Helpline' - an organisation which, *upon investigation by bloggers*, was found to have questionable motives, this information subsequentially being fed back to conventional news producers and discrediting the organisation in their eyes.

> "But the problem is that the Web 2.0 generation is taking search-engine results as gospel." (Keen, 2007, p. 93)

Another issue with Keen's work is that he does not allow for the possibility that information consumers have the capacity to evaluate sources of information - or for the possibility that we might collectively adapt to the new information environment we inhabit and learn to better evaluate information's credibility. As a member of the Web 2.0 generation myself, it did not take long to figure out that search engine results could not be taken as gospel.

Social News takes the task of determining what 'the news' is away from professionals and delegates it to the sites' users collectively - thereby broadening participation to any individual who wants to participate. This task brings with it both work/obligation and power. Keen is concerned that the sites' amateur users do not have the capacity to filter out bad content, and that we are all better off when professional gatekeepers perform this task because they will do a better job, and it is a waste of our time (Keen, 2007, p. 46). Keen does not make any allowances for the new-ness of this approach - Social News websites are novel, and their voting systems are some of the first attempts to perform this kind of task democratically, there may well be room for subsequent iterations to improve upon the process greatly. The task of critically appraising the merits of a wide range of items is also likely novel to many users of Social News as this was previously the domain of experts - there may also be potential for users to learn to fulfill this role more effectively, and in the process perhaps become more critical thinkers themselves.

If 'the crowd' can learn to perform this task collectively the pay-off is 'the power to determine the news' - and if this power can be realised in a genuinely distributed manner that could be a form of 'democratisation' with positive outcomes.

2.12 Research Questions

Distributed Moderation systems represent a novel form of many-to-many communication; and one which offers a unique response to the problem of Information Overload. These systems appear to hold much potential to transform the public sphere; whether they can reach this seeming potential depends on the answers to a number of questions about the way in which these systems function and the affects they induce in their users.

This research is exploratory in nature. At its commencement `reddit`, the website which has come to be its principle subject, was a largely unknown website - the information about it which was available was limited and scattered. There were however signs that `reddit` could become a significant entity on the web, and so the research set out to build an understanding of `reddit` through observation and the interrogation of data which `reddit` naturally produces in its daily operation. This itself constitutes an initial research question: what can we learn about DM websites through the analysis of data which they naturally collect and store?

In addition there are a several broad questions which lie at the heart of this research, and corresponding Chapters which attempt to address these questions:

- How do the websites 'work'? What purpose does the voting system serve and how are the actions of individuals aggregated to produce an end resource? Chapter 5

- Why and how do individuals participate? Is there a standard form that this participation takes or do individuals specialise in different types of activity? Chapter 6

- What are the broader social implications stemming from websites which utilise DM heavily? In the context of Social News how does this diverge from conventionally produced news? Chapter 7

- How have the websites under scrutiny changed over time? Chapter 8

A set of more specific research questions were also adopted in the early stages of this research. In most cases it was not initially clear how these questions could be answered. It has been possible to address some of these with specific analyses (these have section references). Some other questions have not been addressed directly in a particular section but within the thesis relevant information has been uncovered and this is discussed in the final Chapter. There are also some questions which it was not possible to address with the available resources (i.e. data and/or time).

Some of these more specific research questions are as follows:

- How well do Distributed Moderation systems fulfill their task of picking the 'best' items of content to place in the most highly visible locations? (Section 7.3)

- How reliable are these decisions? Will a given population of users always rank a set of items

in the same order, or is this process susceptible to social influence? (i.e. do early voters influence the voting behaviour of subsequent voters?) Salganik and Watts (2009) observed strong social influence effects in experimental cultural markets (music), Social News websites likely exhibit similar phenomena and in principle these could be investigated.

- Do the decisions reached by DM systems accurately reflect the opinions of their contributing users? (Section 7.1) Are the opinions and judgments of all users weighted equally? Or is it possible for individual users or groups to 'game' the system to induce their desired outcome? It has been suggested that Social News websites are host to small coteries of 'influencers' whose submissions dominate high-visibility areas (Keen, 2007) - if this is the case then it implies the voting system may be easily gamed and therefore a poor imitation of a democratic system. (Sections 6.1.4 & 6.1.5)

- To what extent are highly active minorities or 'superparticipants' (Graham, 2013) present on DM websites, and what are the characteristics of their behaviour? (Sections 6.1.4 & 6.1.5)

- Social News websites focus a large amount of their users' attention on a very small sub-set of the items that are submitted. What are the consequences of this from a public discourse perspective? Do Social News sites offer a means of deriving shared experiences from the web for their users? Or are they likely to lead to further fragmentation of public discourse? (Sunstein, 2002) The presence of a default Front page where a small set of items judged to be 'the best' are displayed to all suggests a unifying effect. However, on `reddit` the proliferation of many specialised subreddits means that users who manage their subscription choices may see very few of the same items - indeed a highly specialised set of subreddit subscriptions would be a close conceptual match with the Daily Me. Can users configure a Social News account such that information they would not have sought out and which they have no interest in is entirely filtered out of their experience on the site? Do they choose to do so?

- Are the users of Social News websites homogenous or heterogenous with respect to their outlook and opinions? Does there tend to be a dominant majority opinion on any given topic? Do users whose opinions tend to diverge from the majority persist on the Social News website or move on to another site? Where minority opinions are voiced can they be heard? The voting system seems to offer the majority a mechanism through which they could silence dissenting voices - is it used in this manner? Sunstein (2002) espouses the importance of exposure to a variety of opinions on a topic, can Social News provide this variety of opinion in highly visible locations? Are these sites echo chambers, or could they reach the lofty status of 'general interest intermediary'?

- Can we conceive of Distributed Moderation systems as mechanisms for the management of

Common Pool Resources (Ostrom, 1990)? If so, how do they perform in this regard? Are they managing their CPRs to avoid a tragedy of the commons or are they susceptible to 'free riders'?

- How do Social News websites relate to the 'hybrid media system'? (Chadwick, 2013)

- Do DM systems have strengths and weaknesses in terms of the types of content they process? Do they fare well with controversial content that divides opinion (e.g. politics); or are they better suited to subjects which are more objective or emotionally neutral (e.g. the best way to solve a programming problem)?

- Are there any significant differences between the various forms of Distributed Moderation? Which of these differences are the result of differences in the software being utilised, and which are the result of different user-bases? Are certain features or characteristics of DM software more suited to certain contexts or content types?

- How much activity can DM systems cope with adequately? It seems likely that for a DM system to function well it requires a minimum number of participants or activity level, but is there also an upper limit to the level of activity a DM system can benefit from? Does the performance of the DM system plateau or decrease if this limit is exceeded? Is there an ideal number of participants, submissions, votes, or a proportional relationship between these, which produces the best results when using DM?

- Do DM systems influence the opinions and perceptions of their users? Can these systems be said to promote groupthink or a 'hive mind'? Do they encourage users to think critically about the content they are voting on? Or do they encourage users to rely on the judgments of other users that have previously been registered through the voting system?

- Can the comments on a Social News (SN) website be thought of as 'deliberative discussions' (Habermas et al., 1989)? Or are they better suited to quickly determining the group's consensus and then broadcasting that message?

- Can the users of DM systems use them to make higher-order decisions about what the endeavour's purpose is or how its users should behave to achieve their collective goals? Are the users of SN websites capable of behaving as a unitary social entity? (Section 7.4.1)

- Are there any ways in which DM systems might be improved or adapted to work well in contexts where the current exemplars of the form struggle?

The selection of these particular research questions represents the confluence of what is interesting about these websites (as determined through the literature) and what is amenable to study (as determined by an understanding of the available data and some other research methods which can be employed).

Chapter 3

A review of the literature on Distributed Moderation Systems

Distributed Moderation and filtering systems are becoming increasingly common on the web. The principle behind these systems is that some dimension of user activity is taken as a measure of an item's quality or relevance - this measure being used to rank and sort items, with those items which rank more highly being placed in positions of greater visibility.

The present research focuses on DM systems in which users vote to reflect their appraisal of items of content and will consider Social News websites (where up/down voting is used to determine the merit of posts or 'stories') and a Question & Answer website (where up/down voting is used to rank the questions and also submitted answers).

The present research excludes systems which derive their measure of an items worth from user behaviours which serve a purpose external to the ranking of items. For example, we do not consider Social Bookmarking websites (e.g. `Delicious`) which determine the prominence of items based on the number of users who have bookmarked them and aggregate the tags which users give to these bookmarks to aid navigation. (Yi, 2012)

This chapter will begin by describing the websites which will serve as the subjects of the present research - focusing on the principles and mechanisms these websites employ, and the way in which they are structured.

Every social web application needs a user-base in order to exist as anything other than an abstract concept, and we will also consider the people who power these Distributed Moderation systems through their active participation.

3.1 Social News websites

The present research has opted to focus in the first instance on a class of website referred to here as 'Social News'. To be included in our definition of 'Social News' a website must meet the following criteria:

- Users explicitly judge the quality or relevance of items through a voting or rating system.

- These judgments are aggregated to determine the salience that will be afforded to individual items.

- The website must have a 'Front page' which acts as a focal point of user attention, where the most positively rated items appear.

- A time penalty is employed such that more recent voting activity is given greater weight - with the result being that the website's Front page is occupied by content that is relatively new.

At a fundamental level these websites can be thought of as 'Socially Driven Hyper-Link Aggregation Systems'. Any registered user can submit a link to any web page or resource they wish. Registered users can also vote to indicate their support or lack thereof for the relevance of other users' links. These votes, coupled with a time penalty, are used to determine each link's rank. Ranks are used to determine the prominence with which links will be displayed on the Social News website - with the highest ranking links at a given moment in time appearing on the website's Front page. Links which appear on the Front page of a popular Social News website are likely to bring a significant volume of traffic to the web page they point to.

We use the term Social News to refer to these websites because they are a socially driven means of processing recently-submitted content and determining the salience afforded to each item. A Social News website's collective user-base performs a role which can be likened to that of a news-room editor - both select from a wide range of possibilities which items to acknowledge (or publish) in the limited space available, and which to ignore. While a newsroom editor deals with possibilities from journalists or press feeds, a Social News website works with whatever its users have been submitting. An editor only has so many minutes or pages to fill, and a Social News website only has a certain number of spaces on its Front page.

The similarities between news-rooms and Social News websites are much less apparent when we consider the content which each is processing and publishing. Social News Front pages rarely resemble newspapers - the most obvious difference being that much more space on Social News websites is afforded to content which is humerous or entertaining in nature. There are however active areas on Social News websites which do more closely resemble conventional news media. Sub-categories dealing with topics like politics, science, world, technology and business news -

tend to contain a high proportion of links to articles from generally well-respected sources.

This chapter will introduce the Social News websites which are or have been the subject of academic research. The larger Social News websites (by the criteria outlined here) are `reddit.com`, `Digg.com` and `Slashdot.org`. Particular attention is devoted to `reddit.com`; as this is the website we have elected to concentrate on (for reasons explained below). All of these websites allow any user to submit links to any resource they wish, contain a variety of categories or sections for different types of content, offer a quantitative method whereby users rank these links, and use these rankings (coupled with a time penalty) to dictate how prominently links will be displayed in high-attention areas of the website (e.g. the Front page).

Websites such as `Stumbleupon.com` and `Delicious.com` are excluded from the category of 'Social News' because the service offered by these websites is better described as 'Social Bookmarking'. A Social Bookmarking website derives its measure of an item's quality from an action (that of bookmarking) which serves a purpose external to the ranking and sorting of content. Social Bookmarking websites also devote considerable activity to the classification of content through a tagging interface. This enables the users of a Social Bookmarking website to specify the type of content they wish to be presented with. Social Bookmarking websites place much more emphasis on a user's ability to find the type of content they are looking for, as such their 'Front page' does not serve as a focal point for user attention in the same way that a Social News website's Front page does. Social Bookmarking websites place more emphasis on categorising content, Social News websites are more concerned with the ranking and sorting of content.

Social News websites are still very much in development, they have not yet reached the point where they are known of or used by the majority of Internet users, but over the last few years several have been steadily moving in this direction (Alexa, 2010). One of DiMaggio et al. (2001)'s three main motivators for Internet-related research is the potential to study the development of this communications platform from its early stages. The relative youth of these websites is therefore another aspect which makes them interesting to social scientists - they are still very much in development. In the 3 years `reddit.com` has been under observation it has undergone considerable changes in both the mechanisms it employs and the nature of the content it 'processes' using these mechanisms. Furthermore, it seems that much of this development has been directed by (or in direct response to the expressed desires of) the website's collective user-base.

Social News websites are also quite complex - they combine several modalities of communication and a variety of content types within the same web space. Another of DiMaggio et al.'s (2001) three main reasons that Internet research is important to sociology is as follows: "The Internet is unique because it integrates both different modalities of communication (reciprocal interaction, broadcasting, individual reference-searching, group discussion, person/machine interaction) and different kinds of content (text, video, visual images, audio) in a single medium" (DiMaggio et al.,

2001, p308). Social News websites incorporate all of these communication modalities and content types, and integrate them within the same coherent web space. This complexity, coupled with the rich and ecologically valid data these websites collect, use and store in their day-to-day running - marks Social News websites out as a potentially very fertile field of study for social science research pertaining to the Internet.

3.1.1 `Reddit.com` - description and terminology

`Reddit.com` (henceforth `reddit`) is a Social News website which has previously described itself, somewhat ambitiously, as "The voice of the Internet – News before it happens". `Reddit`'s current description is "The Front page of the Internet". There are a number of differences between `reddit` and other Social News websites, some of which identify it as a good place to start studying Social News. The software behind `reddit` is Open Source - this means that it is possible to use this software to set up similar websites, and also that it is possible to probe the inner workings of this software in great detail (`reddit`'s anti-spam/cheating code is the only exception, it is excluded from the open source code because to make it public would be to compromise its integrity).

The way `reddit` functions is also relatively straightforward and based largely on an up/down voting system. Every item of content on `reddit` has a score (which starts at 1) - by clicking an up or down arrow relating to these items a user can increase ('up-vote') or decrease ('down-vote') the item's score by 1 point. Up/down voting operates on both primary units of content and their comments and in both cases these votes are used to determine the prominence with which the item or comment is displayed. Voting on `reddit` occurs on a single 'dimension' (i.e. an item has just one score) but this dimension could be described as 'bi-polar' (in that votes can increase or decrease an item's score, and scores can be negative values).

`Reddit` lacks more obfuscated features like the 'share' and 'bury' mechanisms of `Digg`. `Reddit` also makes much more limited use of social networking features (like the ability to add another user as a 'friend') - these having a significant effect on `Digg`'s functioning (Wu et al., 2009). If we wish to study the workings of Distributed Moderation it is beneficial that `reddit` does not complicate the picture by adding network effects. Also, `Reddit`'s administration team have a reputation for a 'hands-off' approach - only removing content from the system when it breaks the law or a set of well-defined guidelines. These characteristics would seem to make `reddit` a good place to start when studying the emerging phenomenon that is Social News.

Posts

All primary units of content on `reddit` (henceforth 'posts') can be divided into two types - link posts and text posts. A link post consists of a title and URL, when a user clicks on the title they are

directed to the resource at the end of the URL. These links can direct a user to any web resource - posts which link to articles, blogs, pictures and videos are common. When `reddit` was created all posts were link posts (i.e. the inclusion of a URL was mandatory), `Digg.com` retained such a policy. The second class of post, text posts (referred to by `reddit`'s software as 'self' posts), were added to `reddit` some time after its initial creation - to facilitate the development of a `reddit` community. A text post consists of a title and optional body of text created by the submitting user through a text entry box on `reddit`. When a user clicks on a text post's title they are transported to the comments page for that post, where the (optional) body of text is displayed underneath its title at the top of the page.

Reddit employs a nested 3-level structure consisting of Subreddits, Posts and Comments. Subreddits can be thought of as categories of content; every post submitted to `reddit` must be submitted to a specified subreddit. Posts are the primary unit of content, each post has its own aggregate score and rank, and it is posts which have the capacity to appear on `reddit`'s Front page. Every post has its own comments page, and a comment can only be entered on the page of a specified post.

Subreddits

`Reddit`, like other Social News websites, uses a categorisation system which provides different sub-categories (henceforth subreddits) for different types of content. For example, there are subreddits for political content (/r/politics), world news (/r/worldnews), humerous content (/r/funny), pictures (/r/pictures) and videos (/r/videos). These subreddits are often referred to by `reddit` users by pre-fixing the subreddit's name with '/r/', as this is the way the website denotes subreddits in its urls (e.g. `www.reddit.com/r/politics` is the Main page for the politics subreddit).

In contrast to other Social News websites this subreddit classification scheme is dynamic and open to users. Any registered user can start their own subreddit, and automatically becomes the moderator of any subreddits they start. Individual users can also choose which subreddits they subscribe to, and when a user is signed into their user account they will only see posts from the subreddits they are subscribed to on their `reddit` 'Front page'. There is however a list of 10-12 'default' subreddits (expanded to 20 in November 2011) - people who visit `reddit` without being signed into a user account will be presented with posts from these default subreddits, and when an individual creates a new user account they will be automatically subscribed to this list of subreddits. Consequently, visitors to `reddit` who are not signed into a user account will all be presented with the same 'Front page', whereas the content of this Front page is variable for users who sign into an account and manage their subscriptions.

`Reddit` is now comprised of many thousands of these subreddits. The distribution of subscriptions

(and consequently attention) between these is highly skewed such that the bigger and more active subreddits have hundreds of thousands of subscribers[1], while there are thousands of subreddits with just one or two subscribers. Subreddits move up and down on this subscription distribution when registered users choose to add or remove them from their Front page. If a relatively young subreddit gains enough subscriptions and becomes widely known `reddit`'s administrators may add it to the 'default subreddit' list. The main effect of this is that popular content submitted to this subreddit becomes visible on the Front page for (the very large number of) visitors who aren't signed into an account, and that the subreddit becomes a default choice for new users who register an account. It is assumed that both of these mechanisms combine to produce a substantial increase in the attention received by posts on a subreddit at the point where it becomes a default choice. The question of what happens to a subreddit when it becomes a default choice will be considered in Chapter 8.

Comments

Commenting is an important aspect of the activity on `reddit`. Up/down voting is applied to these comments in the same manner as for posts - with a score being generated for each comment that determines its rank, and this rank dictating how prominently it will be displayed on the comments page for the post it relates to.

When one visits the comments page for a post on `reddit`, the comment at the top of the page will be the 'top-level' comment with the highest rank (within the last three years the default method of ranking comments has been changed, see section 5.1.7). Comments are threaded such that the second comment vertically (and indented) is the most popular response to the first comment, and the third comment on the list is the most popular response to the top second-level comment, and so on and so forth until the maximum number of comments per thread is reached. There is no hard limit on the depth of these threads but by default comments which are beyond the 6th level will not be displayed unless the reader clicks a link to 'load more comments'. Below this first top-level comment's thread a second one begins, starting with the top-level comment that has the second-highest rank.

This logic is applied to the sorting of all comments on the website, it allows a great number of individuals to contribute to a discussion while maintaining a level of coherence and readability not usually associated with discussions involving hundreds/thousands of people. It is not uncommon for a single post to have several thousand comments, and each of these comments can receive hundreds of votes - much higher levels of activity than could feasibly be handled by a more traditional bulletin board system where text posts are the only available method of contributing to a discussion. By adding a quantitative dimension to the discussion, `reddit` allows users to express their opinion

---

[1]In 2013, millions

about the value of a particular comment without adding to the corpus of text which makes up the discussion.

Comments with a score below a certain threshold (-4 by default) are hidden and replaced by a 'comment below threshold' message, but can be expanded by clicking a button next to this message. In this manner `reddit`'s voting system performs the task of removing inappropriate contributions from a discussion. Most online discussion fora delegate this task to a group of users who are appointed as moderators and given power over other users' contributions. `Reddit` also appoints certain users as moderators, but some of the conventional responsibilities of the moderators have effectively been delegated to the collective use of the voting system by users. [2]

Pages

`Reddit` provides four tabs which select posts to display according to different criteria. All of these tabs (and sub-tabs) operate in both aggregated form, and for individual subreddits. In aggregated form these tabs display content from all of the subreddits a user has subscribed to, or from the default subreddits for users who are not signed into an account.

The first of these tabs is titled 'hot' or 'what's hot', and it selects popular posts based on their score, with a time penalty for older votes. `Reddit`'s Front page is simply the aggregated form of the 'hot' tab. The 'hot' tab for a particular subreddit is referred to here as the subreddit's 'Main page' (because it is displayed upon navigating to the subreddit).

`Reddit` also offers a 'New' tab which has two sub-pages - the 'New' page shows the latest submissions ordered by their submission time, the 'Rising' page shows new submissions which have attracted early positive votes.

The remaining two tabs offer a drop-down menu which allows the user to specify which time interval they wish to look at (e.g. posts from today, this week, this month, this year). The 'controversial' tab displays posts which have attracted a similar number of positive and negative votes, while the 'top' tab shows the posts which have the highest score for the selected time period.

3.1.2 `Reddit` users are influential in the website's development

Ordinary `reddit` users have influenced the website's development in a number of ways. First and most obviously through the up/down voting interface itself. On the Front page `reddit` users are presented with a selection of very successful posts - this likely acts as a feedback loop influencing

---

[2]It should be noted that moderators have a higher profile now in 2013 than they did in 2010 - and that the extent to which moderation occurs varies considerably between subreddits. This is somewhat counter-intuitive because as the number of active (voting) users increased one would have expected that the power of the communities to self-moderate should also have increased. See Chapter 8 for a discussion of this trend.

users' perception of the type of post which the `reddit` community appreciates, and through this their decision about which posts to submit.

The choice of which subreddits an individual user subscribes to also influences `reddit` at an aggregate level. `Reddit` users frequently create new subreddits, and these are often intended to be used for particular types of content or endeavours which do not fit neatly into an existing subreddit. When such a subreddit becomes popular the result is that many `reddit` users then use the website to peruse content of a type which it was not previously associated with. Most of the subreddits which have seen the most prolific growth in the time that the present research has been observing `reddit` cater to particular types of text posts. AskReddit (155,931 subscribers as of 27/04/10, 1,355,410 subscribers as of 12/03/12, 2,930,047 subscribers as of 31/01/13) is a subreddit where users make Text Posts which are questions, and receive responses on the post's comments page. More recently an 'AskScience' subreddit has risen to prominence (516,045 subscribers as of 12/03/12) - this follows a similar format to AskReddit with the difference that questions (and in particular answers) are intended to be of a scientific nature and answers must provide links to some sources of evidence. IamA (58,152 subscribers as of 27/04/10) is a subreddit where Users identify themselves as holding some sort of occupational or life experience (or sometimes as a well known individual) - with the intention being that other users post questions for them on the comments page and the 'Original Poster' (OP) answers the most popular of these. IamA has been particularly successful in the last two years, growing to 253,080 subscribers (as of 28/04/11, and 2,664,914 subscribers as of 31/01/13) and earning a place on the default subreddit list in March 2011. Section 5.1.12 looks specifically at how the voting system is used to conduct IAmA interviews. Both `Digg` and `reddit` run (or have run in the past) 'official' versions of a similar idea, where users make comments containing questions for some well-known individual and the website's staff record a video interview where these questions are put to said individual - it seems that this particular idea first appeared on `Digg` in the form of 'Digg-Dialogg'.

All of these novel uses of the existing `reddit` system have been pioneered on the site by ordinary `reddit` users, and have come to prominence through the subscription choices of other ordinary users. Subreddit subscriptions are a second example of how `reddit` users seem to collectively shape the website's perceived purpose or identity.

The fact that `reddit`'s software is open source, and that it provides a publicly accessible Application Programming Interface (API), allows `reddit`'s users to be involved in the website's development on an even more fundamental level. Due to `reddit`'s open source software ordinary users can find and report bugs to the administrators, and write scripts which add to the functionality of the website. For example, the currently default method of sorting comments was developed by a `reddit` user, and many `reddit` users utilise a 'greasemonkey' script that displays the number of up and down-votes a comment has received (by default `reddit` only reports com-

ments' aggregate scores). For further examples of popular user-produced greasemonkey scripts see `http://reddit.honestbleeps.com`. [3]

Reddit users have also developed a number of external websites which take data from `reddit.com` and apply particular presentation formats or use it for some other purpose. There follow some examples:

- `codereddit.com` displays `reddit` in a format which looks like Java software, enabling `reddit` users who are programmers to browse the website surreptitiously while at work.

- Vizeddit (`http://erqqvg.com/vizeddit/?v=2.0`) offers a real-time visualisation of activity on `reddit`'s Front page, showing which posts are being up-voted and down-voted, and also when new comments are added.

- karmawhores.net ranks `reddit` users along several dimensions of their `reddit` karma (`reddit`'s reputation system)

- `MetaReddit.com` offers a number of ways of browsing for new subreddits. `Reddit` itself offers very limited means of navigating its thousands of subreddits, this website aims to make this task much easier for `reddit` users.

- `Imgur.com` is an external image-hosting website created by a `reddit` user which has since become popular among `reddit` users and in its own right (redd.it/7zlyd).

- In 2013 there are now a plethora of such sites, for example: `stattit.com` (showing usage statistics for subreddits), `unedditreddit.com` (allowing users to see the text of deleted comments), and `redditarchive.com` (showing popular posts for a specified date).

3.1.3 `Reddit.com` - communication modalities and content types

"The Internet is unique because it integrates both different modalities of communication (reciprocal interaction, broadcasting, individual reference-searching, group discussion, person/machine interaction) and different kinds of content (text, video, visual images, audio) in a single medium" (DiMaggio et al., 2001, p308).

On `reddit.com` all of the communication modalities and content types referred to by DiMaggio et al. (2001) are present on the same website. Content nodes on `reddit` can be either text directly from the submitting user (both statements and questions are common), or a link to a resource which is a web page, video, picture or (less commonly) audio.

Reciprocal interaction takes place when users vote or comment on each other's posts (and when they vote on each others' comments). Group discussions are also present in the comments threads

---

[3] This now re-directs to the 'Reddit Enhancement Suite' - a very popular suite of add-ons which add a lot of functionality not present on the `reddit` website itself.

of many posts. It is also quite common for a popular comment to be one where an individual has sought reference for the original post's content (often for the purpose of exposing flaws in the source).

Person/machine interaction is undoubtedly at the core of these Social News websites. Users collectively determine the prominence of content on the site with their individual votes - but it is the software on the server which turns these votes into rankings and applies them to the display of content. An individual cannot decide whether a certain content node should appear on the website's Front page, but they can vote to increase or decrease the likelihood that this will happen.

The communication modality on `reddit` which is of greatest interest is broadcasting - `Reddit` broadcasts the 25 most popular posts at any given moment to a wide audience of visitors and users. Any individual could create a new account on `reddit` today and submit a link to something that they've found (or produced) which they think is worthy of attention - if the voting response of other users is sufficiently frequent and positive (i.e. the post appears on the Front page) then this item of content will be seen by many thousands of individuals within a few hours.

It is still possible for users to interact with a Front page post in terms of voting and commenting. However, a post which reaches the Front page will likely receive considerable levels of voting and commenting activity, meaning that new user votes and comments are somewhat diluted by the creation of many concurrent votes and comments. This has the result that by the time a post reaches the Front page its comments thread will likely be relatively stable. It also seems that the large increase in attention received by a post on the Front page usually results in an increase in voting activity for that post. This up-surge in voting activity is often enough to keep a post on the Front page for several hours, despite the effect of a time penalty steadily eroding the weight of its existing votes. Chapter 5 explores these issues through the analysis of data from `reddit`.

3.1.4 `Digg.com`

`Digg` was, for a number of years, the largest and most widely known Social News website. In August 2010 `Digg` made significant alterations to its software which resulted in a revolt of its users (see section 3.3.1), and since then it has failed to keep pace with `reddit`'s rapid growth. The research which we consider that relates to `Digg` was conducted before the change to 'version 4' in August 2010, and therefore `Digg` is described as it was before this change.

One major difference between `Digg` and `reddit` is the lack of bi-directional voting on `Digg`. On `Digg` users could 'Digg' a post to increase its score but there was no way for a user to reduce this score. Instead `Digg` offered a 'bury' button - if enough users choce to 'bury' a post it would effectively disappear from `Digg`. `Digg` did introduce bi-directional voting for comments after `reddit`, but the scores of comments were not used to determine their display order.

A further difference between `reddit` and `Digg` is that `Digg`'s structure is static and controlled by the website's owners - there are a set number of categories and sub-categories for content and unlike `reddit` users cannot create their own categories. Moderation on `Digg` is also controlled by employees of the company which owns the website, and it seems that these moderators may have more influence over the promotion of stories than moderators on `reddit` (Thomas, 2008). On `reddit` moderators are ordinary users, the person who starts a new subreddit is its first moderator and they can add other users as moderators of that subreddit. Each subreddit lists its moderators at the side of the page and links to their account histories. In contrast `Digg` obfuscates the identity and role of its moderators, in fact the very existence of moderators is rarely if ever mentioned on `Digg`'s official pages.

`Digg` also utilises social networks to a much greater degree than `reddit`. Users on `Digg` are encouraged to follow other users and these choices affect the posts which a user will be presented with. This may have facilitated the emergence of 'power users' who have a large number of followers and who are therefore much more likely to have their posts promoted to the `Digg` Front page (Arrington, 2006). `Digg` also now integrates with `Facebook` and `Twitter`, it is possible to log in to `Digg` with a `Facebook` or `Twitter` ID and to import one's friends/followers from those applications.

This research intended to compare the functionality of `reddit` and `Digg` through the analysis of data from both websites. However, within one month of the data-collection software for `Digg` being produced, `Digg` switched to 'version 4' of its software and changed its API (Application Programming Interface) considerably - with the result that the data-collection software for `Digg` no longer worked. As it became clear not long after this 'upgrade' that `Digg` was in a steep decline in terms of its usage statistics, the decision was taken not to invest the time to create new data-collection software - and to drop consideration of data from `Digg` from the research altogether.

3.1.5 `Slashdot.org`

`Slashdot` pre-dates `reddit` and `Digg` and has persisted with an active community for many years without growing to the size of `reddit` or `Digg`. `Slashdot` also operates in a way which is distinct from `reddit` and `Digg` in a number of important ways. Perhaps most importantly, voting on `Slashdot` restricted to pseudo-randomly selected users who receive a limited number (five) of votes which they may use at their discretion. Votes are cast on a 5-point scale and it is possible to add a tag (e.g. "interesting") which is intended to describe the post or comment. These votes (or 'moderations', as they are referred to on `Slashdot`) are also subject to meta-moderation - if one does not agree with a particular rating there is an avenue through which this can be expressed. `Slashdot` has not been the subject of direct scrutiny in the course of this research, but has been

the subject of a considerable proportion of published research on 'Social News'.

3.2   The Social Implications of Social News

The method of directing attention employed on Social News websites stands in stark contrast to the mass media model - where one must either pay for consumers' attention in the form of advertising, or convince the medium's gatekeepers that the item is worth covering on its own merits. On a Social News website there is no clear distinction between the people who submit content, the people who decide which items of content are widely publicised, and the consumers of the resulting resource - an individual can easily participate in all of these roles concurrently.

The interaction taking place on a Social News website like `reddit` is also distinct from that which takes place on a 'Social Networking' website like `Facebook` or `Twitter`. On these websites an individual's submissions are directed at their network of 'friends' (`Facebook`, Lewis et al. 2008) or 'followers' (`Twitter`). In the case of the 'Blogosphere' one's blog entries are only visible to individuals who read the blog in question.

Arenas such as `Twitter` and the Blogosphere are capable of disseminating information to a wide audience (Kwak et al., 2010), but the way in which this occurs is better described as augmented word-of-mouth than broadcasting. On `Twitter`, individual messages (tweets) are frequently repeated (re-tweeted) so that they are circulated beyond the original author's network of followers. `Twitter` also uses the prevalence of keywords (or 'hash-tags') to highlight the most popular terms being discussed at any given moment ('Trending topics') - thus providing a short-cut to the topics under most active discussion, roughly analogous to 'news'. Similarly, blogs tend to link to each other quite frequently (Adamic and Glance, 2005) and in this manner an individual blog entry can be disseminated to a much wider range of individuals than regularly read said blog . There are also aggregation methods in place on certain websites which augment this process (e.g. `boingboing.net`).

Social News websites differ from these other methods of disseminating information in some important ways. Firstly, the potential audience for any item of content is determined by the area of the website which it is submitted to - the submitter's personal contacts within the network have no direct bearing on who will see their submission initially. As noted above some Social News websites (`Digg` in particular) have made use of Social Networking principles, and these incorporate an option to browse content submitted by friends - a basic version of this kind of feature was added to `reddit` relatively recently. `Reddit` allows users to designate others as 'friends', and provides one page on which a user can see the recent submissions of those users they have designated as friends.

Secondly, a lot of the activity taking place on these websites (i.e. voting) pertains directly to the decision about which items of content the website will make most prominent. Users make these choices as individuals, but their individual choices are aggregated and expressed as group decisions. The website's users collectively determine what will be displayed most prominently on the website for all immediately subsequent users and visitors.

Social News websites, in their original conception as 'Social Hyper-Link Aggregation Systems', are an augmentation of the existing web - they specialise in highlighting the most attention-worthy content and do not produce new content. Social News websites may offer a way for individuals to collaborate on this task of exploring and assessing the quality of information on the web. If one comes across an item of content on the web which they believe is worthy of public attention it is both quick and easy to submit a link to this on a Social News website. Indeed some popular websites (e.g. BBC) have begun incorporating direct links to the submission pages of an array of social news sites - making this process even easier. [4]

When an individual browses the content which has already been submitted to a particular Social News website - all of this content comes with scores and values indicating its quality as perceived by other users of the website. When browsing to the page of any link (or text) post on `reddit` one can immediately see how many users have up-voted or down-voted it, both as absolute values and as a percentage who 'like it'. If one looks at this post's comments one can see what the most popular response to the post's content has been so far - as determined by the same up/down voting system used for posts, and providing the same variety of output values. If I am convinced that this hypothetical link post is worth a visit then I can click through to see its target page. If I wish I can then feed my opinion on this post's quality back into the system with a vote. If I see a comment which captures the essence of my opinion on the post's content I can endorse this in the same way.

When an individual visits the Front page of a Social News website they are effectively reaping the rewards of all the recent content discoveries and assessments performed by its collective user-base.

3.2.1 The spread of Distributed Moderation principles

Two forms of quantitative web interaction pioneered on Social News websites have recently begun to gain traction outside the arena of Social News. The application of a publicly accessible up/down voting interface for comments has recently spread to a variety of websites which are not Social News in nature (e.g. `Youtube.com`, `Facebook.com`, `dailymail.co.uk`). The "Digg-Dialogg" style interview - where the interviewee takes questions that have been voted on by members of the public - has also started to appear on more widely accessed websites. For example, during the 2010 UK

---

[4]Looking back at this in 2013 - it was something of a fad and in many cases the button relating to `reddit` has been removed.

general election `Facebook` and `Youtube` collaborated to conduct this type of video interview with the leaders of the UK's 3 main political parties (Facebook and Youtube (2010); Guardian (2010a); Telegraph (2010b)). [5]

Some of the principles of Distributed Moderation are also now being used by political advocacy groups (Change.org, 2010; The Point, 2010) to facilitate the exchange of ideas and the organisation and prioritisation of campaigns. Likewise, a vote up/down mechanic has recently been used by a British Newspaper to facilitate political debate (Telegraph, 2010a).

In other instances Distributed Moderation systems have been applied directly to political policy ideas, either by political parties or governments. `AmericaSpeakingOut.com` is a website created by the American Republican Party which solicits political policy ideas from its users. These policy ideas are subjected to 'approve/disapprove' voting by other users, with the ideas that receive the greatest levels of approval being shown in the most prominent locations. The intention behind the website was that these most popular ideas would be put forward to Republican leaders and form the basis of policy; but the website appears to have failed in this regard (Benen, 2010; Milbank, 2010). `AmericaSpeakingOut` has since been discontinued, but it operated for a number of years with software that was quite similar to that employed by `reddit` - albeit lacking several of `reddit`'s features (e.g. comment voting and sorting).

Your Freedom (`www.YourFreedom.hmg.gov.uk`) was a website created by the current British coalition government in 2010. `Your Freedom` also solicited political policy ideas from the general public and allowed users to rate each other's ideas. In this case the intention was that users would use the system to collectively highlight pieces of legislation which should be repealed or submit ideas for the enhancement of civil liberty. `Your Freedom` opted for five-star ratings of ideas, as opposed to up/down voting, and an argument is put forward below that this was a sub-optimal choice. `Your Freedom` also had limited facility for commenting and no method of effectively rating and ranking comments - this limited any capacity `Your Freedom` users may have had to discuss and develop their ideas as a community.

Up/down voting vs. Five-star rating

Up/down voting systems accord each item a score, a single number which tells us something about the popularity of the item, large scores indicating that a lot of people endorse the item. If one wishes to know what proportion of people endorse an item one can look at its ratio of upvotes to downvotes. The scores produced by up/down voting are also easily converted into ranks, and a time penalty can be applied if required. Up/down voting also offers a neater conceptual fit with

---

[5]More recently this trend has 'come full circle' when Barrack Obama participated in an Ask Me Anything interview on `reddit` as part of his presidential campaign in 2012 (redd.it/z1c9z).

political policy ideas - when politicians vote on legislation they can either vote for or against it, or abstain.

Five-star ratings do not offer the same conceptual fit with voting in the domain of politics. We might imagine that a rating of 5 equates to full support and 1 equates to full opposition but what is the meaning of a 2, 3 or 4 star rating? In aggregating peoples' ratings the tendency is to calculate a mean rating for each item, but this is not as easily interpreted or used as an up/down voting score. For instance, an item with a full 5-star rating based on 5 votes probably doesn't warrant as much attention as an item with a rating of 4.9 that has received 500 votes. How would one interpret a mean rating of 3, which could be produced by many people with neutral feelings on the matter or two sets of individuals with strong feelings either way?

Section 5.3 presents some analysis of data from `Your Freedom` which was released after the website was discontinued.

3.3   Social News-makers

User submissions and votes are the raw materials of Social News, they are the engine that drives Distributed Moderation. As such, it is important to develop an understanding of how and why these users participate. Do Social News users all participate in the same manner (i.e. voting, submitting and commenting)? Do they participate for the same reasons? Or do they participate in different ways to different ends?

3.3.1   Motivation to participate

Websites which employ Distributed Moderation can be treated as an example of online Peer Production. The term Peer Production encompasses any endeavour which allows individuals to participate autonomously and provides an infrastructure whereby these individual contributions are combined to produce an information resource. `Wikipedia` and open source software are well known examples of Peer Production and the motivations of contributors to these endeavours have been studied by a number of researchers. These motivations are multidimensional in nature (Schroer and Hertel, 2009) - including the intrinsic love of writing/programming and also a set of motivations associated with participation in conventional social movements (i.e. the desire to be involved in an endeavour which is perceived as good or worthwhile). In the domain of open source software development these motivations are augmented by a set of personal benefits which may accrue to the individual contributor - namely the availability of better software, or career advancement resulting from the recognition of one's work on an open source project (Ghosh, 2005; Hertel et al., 2003).

Lampe et al. (2010) conducted a survey concerning the motivations of users on `Everything2.com`, an encyclopedia type website which receives a large volume of user contributions. This study adopted the 'Uses and Gratifications' theoretical framework (Ruggiero, 2000), which considers how users behave and how this is related to the gratification they attempt to elicit. Lampe et al. (2010) found that users tended to come to the website initially to gain information, but that they tended to obtain additional benefits (related to receiving entertainment and providing information) while on the site and these additional benefits are implicated in their subsequent return to the website. Lampe et al. (2010) report evidence that users create an account and participate due to 'social enhancement motivations' - they feel that they are important to the online community and that their participation will benefit this community. A major difference observed between anonymous users and those with an account is that the former are only interested in receiving information, while the latter are most interested in providing information.

Recognition of or attention for one's work has also been shown to motivate contributors to the video sharing website `Youtube` (Huberman et al., 2009; Wu et al., 2009) - with the number of views of an individual's videos being strongly related to the individual's propensity to continue contributing. Similar effects have been observed on `Slashdot` (Lampe and Johnston, 2005) and `Digg` (Halavais, 2009; Wu et al., 2009); with feedback received through the voting system increasing the chances that an individual will continue to submit content.

Lampe and Johnston (2005) studied `Slashdot` from the perspective of new users, analysing server logs and also conducting a survey of 233 `Slashdot` users. 11,079 new `Slashdot` users who created accounts within a two month period were analysed with respect to their commenting activity and the feedback received on their comments. The distribution of comments between these new users was found to be highly skewed - a group of 1763 users made 6467 comments between them whereas 55.1% of users who commented made only 1 comment. The survey conducted by Lampe and Johnston (2005) as part of this study attempted to measure the new users' previous experience of online communications platforms - to assess whether users who were more familiar with online communications would have an advantage when they began participating on `Slashdot`. No significant relationships were found between the survey measures and the outcomes related to the user's participation on `Slashdot` (i.e. more experienced users did not submit better comments or more comments).

A logistic regression on whether a user who submitted one comment would submit a second comment found no significant effect of that user's first comment having received moderation or replies (i.e. users whose first comment was rated were not more likely to comment again than users whose first comment was not rated). This model did however show significant effects related to the delay between account creation and first comment, and the number of `Slashdot` pages the user viewed (users who waited longer before making a comment, and who viewed more of `Slashdot`'s

pages, were more likely to submit a second comment). A regression model which applied the same explanatory variables to modeling the delay between a user's first and second comment found that having a positively rated first comment, and a longer delay between account creation and commenting, were both significant predictors that the user would make a second comment more quickly. Having a negatively rated first comment, number of `Slashdot` pages viewed, and receiving a reply to the first comment - were not significant predictors of the time until a second comment was made. (Lampe and Johnston, 2005)

Halavais (2009) studied a sample of 30,000 `Digg` users, and had access to the complete commenting history of these users. Of this sample 23,523 users had no comments and 2,728 had just one comment. However, the 6,468 users who had at least one comment were together responsible for 197,658 comments - again suggestive of a highly skewed distribution of commenting activity between users. Halavais (2009) looked at the delay between users' comments, reporting that when a comment received no moderation the user waited on average 19.8 days before commenting again, when a comment did receive moderation this figure was 6.1 days. Small negative correlations were reported between total `Digg`s (-0.1), buries (-0.06) and replies (-0.04) respectively, and the time until the user's next comment. Comments which were the last for a given user within the observation period (taken as implying drop-out) tended to receive fewer `Digg`s (3 vs 6.6), fewer buries (1.4 vs 2.7) and fewer responses (0.3 vs 0.5) than comments with a subsequent comment. These results were taken as support for the hypothesis that feedback received through `Digg`'s DM system encourages users to continue submitting comments.

In these cases positive feedback ('up-votes') is associated with increased attention but it also also represents an endorsement of the submitting user's contribution by the website's community of users. When a user of a Social News website submits an item of content which receives positive feedback this has a number of related effects - the submitting user is being told that other users have appreciated their contribution, with enough positive feedback their contribution could be seen by hundreds of thousands of other users, and all of the points received by their contribution are added to the submitting user's reputation score. It is likely this combination of possible rewards encourages users to keep submitting content to the Social News website. Many of the user's contributions will receive a modest number of up-votes which are added to their permanent reputation score, and there is always a chance that their post will 'hit the jackpot' and make it unto the Front page.

Motivation to Vote

Many of the motivations to contribute to these websites outlined here are not intuitively applicable to the issue of voting. Voting on `reddit` is a largely anonymous activity, eliminating motivations relating to personal benefit (such as the desire to have one's contributions recognised and appre-

ciated).

It has been shown that, across a number of Peer Production endeavours, participation begets further participation. The probability that an individual will cease contributing varies inversely with the individual's level of participation (Wilkinson, 2008). Wilkinson (2008) adopts the position that this effect produces a power law distribution (Clauset et al., 2009) of activity between participants in Peer Production - where many individuals seldom participate and the majority of contributions come from a small sub-set of very active users. This effect was observed for both submissions and votes on `Digg`, with the distribution of votes having a much lower exponent due to the relative ease of casting a vote as compared to making a submission.

This finding suggests that users of Social News websites may be motivated to vote through a feeling of involvement in the running of the website - in the same way that one motivation cited by `Wikipedia` users relates to their sense of involvement in the production effort (Schroer and Hertel, 2009). It would however be remiss to assume that users vote for such altruistic reasons exclusively. A user may vote because they understand that the website cannot function without this kind of input, or they may vote in a self-interested fashion for the purpose of influencing the collective decision about which items will be displayed most prominently.

A lot of attention accrues to content which reaches the Front page of a Social News website and this attention is itself a valuable resource. Huberman et al. (2004)'s study demonstrated that individuals would forsake financial gain for attention and status within their experimental group. On SN websites the attention of hundreds of thousands of individuals is focused on the Front page, and the social voting system determines what will be placed here for these individuals to see.

There are numerous stories of groups which have attempted to 'game' the Distributed Moderation systems employed by Social News websites - either for the purpose of filtering out content on ideological grounds (the 'Digg Patriots', (Halliday, 2010; Oleoleolson, 2010)), or for marketing purposes. Several companies have attempted to sell front-page placement of an item on `reddit`, which they would achieve through the control of many user accounts (redd.it/elqgs, (Woyke, 2010)), this kind of behaviour had also been observed previously on `Digg` (Warren and Jurgensen, 2007).

The power wielded by Social News users through the voting system was further demonstrated on `Digg` in the aftermath of the website's 'upgrade' to a new software version in August 2010 (Friedman, 2010a; Lardinois, 2010b; Wauters, 2010). Many `Digg` users did not approve of the changes made with 'version 4' of the website's software, which were seen as removing power from ordinary `Digg` users. Immediately following this upgrade `Digg`'s Front page was filled with items complaining about the upgrade and calling for a reversion to the previous system. In the days that followed, as it became clear that no reversion to the previous software was planned, `Digg` users began up-voting posts which linked to `reddit`. This produced an unusual spectacle whereby `Digg`'s

Front page was filled with posts linking to `reddit`, `Digg`'s main competitor, for a period of around one week. During this time it is estimated that `Digg` lost around 26% of its traffic (Lardinois, 2010a), while `reddit` saw a corresponding increase in its usage statistics (Reddit, 2010b). `Digg`, a company once valued at hundreds of millions of dollars, never recovered from this exodus - while `reddit`'s usage statistics continued to show strong growth `Digg`'s continued to decline.

These examples illustrate that Front page placement on a Social News website has value. As the decision about which items appear on the Front page is made collectively through the voting system, every user who votes is contributing to this collective decision, and votes from all users are weighted equally. This is a core aspect of the 'bargain' (Shirky, 2008) between a Social News website and its users. The opportunity to be involved in making a decision which is perceived as important may represent a motivation that is specific to voting in the domain of Social News. Further examples of the importance of this 'bargain' to the users of Social News websites were discussed in section 2.6 - when the owners of a Social News website are seen by its users to interfere unjustly with their decision as to what should appear on the Front page, this can provoke a vociferous response from users.

Distributed Moderation systems may themselves also hold some intrinsic appeal for would-be participants (or a sub-set thereof). Wise et al. (2006) conducted an experiment where participants viewed one page of a manufactured discussion forum in either of two conditions. In one condition the page contained some hallmarks of Distributed Moderation (e.g. a 5-point rating scale beside each post, a message informing the reader that posts would be rated by volunteer moderators and a 'comment not yet rated' message beside any submissions made by the participant). In the other condition these markers of DM were absent. Participants 'intent to participate' was then measured with a 10-item survey instrument. Participants who had completed the 'moderated' condition showed significantly greater intent to participate than those who completed the 'unmoderated' condition. This result suggests that the presence of Distributed Moderation may appeal to potential users of online communications platforms.

It is also possible to consider all of the potential motivations of `reddit` users as being either internal or external to the website. Internal motivations would stem from an affinity with the website and its community - a user might submit content for the benefit of other users, and be rewarded by positive feedback received through the voting system - or a user might vote to provide the kind of honest appraisal of content which the website needs to function. Further 'internal' motivations might relate to the accumulation of reputation points (i.e. one may comment on a post in the expectation that this comment will receive positive votes and increase one's score). External motivations would stem from the attention Social News websites channel. A user might submit content because they want it to receive attention (irrespective of where this attention comes from), or they might vote because they want to promote a certain type of content (e.g. links to

blogs with political leanings similar to the user's own) irrespective of the venue.

Badges and Reputation points

One of the motivational tools at the disposal of any Social Media website is gamification - the practice of adding game logic or mechanics in non-game contexts (Jane, 2011). There are two varieties of gamification which are prevalent in Social Media - badges and reputation scores.

The awarding of badges in recognition of an achievement has a long history (most notably in military and 'scouting' organisations)(Halavais, 2012). The issuing of *digital* badges is a more recent phenomenon. In 2002 Microsoft launched an 'Achievements' system which worked with games for their Xbox 360 console, and since then 'Achievement' badges have become commonplace in video games. In the Xbox case each player has their own profile, when they are playing a game and meet a certain criterion a graphic appears overlaid on the screen which informs them that they have just been awarded an achievement (giving its title and the number of 'gamerscore' points it contributes). Within each player's profile there is a record of all the achievements they have earned across all Xbox games and also a 'gamerscore' (reputation score) - each achievement awards a set number of points and a player's 'gamerscore' is the sum of all the points they have earned through achievements. Achievement badges for games are awarded for a variety of criteria - from awarding a badge the first time the player performs an action which is central to gameplay (e.g. the first time the player scores a goal in a football game), to rewarding mastery of the game (e.g. completing the game on a specific difficulty setting) or experimentation within the game.

Xbox achievements have proven to be highly effective motivators for at least some of the people who play Xbox games. Whole communities of practice have sprung up around 'achievement hunting' (e.g. `www.xbox360achievements.org`, `www.trueachievements.com`) wherein users discuss strategies for earning achievements efficiently. Such strategies include the identification of games for which achievements can be easily and quickly unlocked - and individuals who participate in 'achievement hunting' will obtain and play (often terrible) games for the sole purpose of unlocking their achievements. Beyond these communities of dedicated achievement hunters, it is difficult to establish how effective achievements are as motivators. One suspects that the extent to which individuals are motivated by achievements shows considerable variation and that many users discount these entirely.

Despite a lack of understanding of the motivational aspect of badges and reputation scores they have become commonplace in a variety of Social Media (Antin and Churchill, 2011). One can understand the appeal of badges from a designer's perspective - it costs very little to implement a system of badging and there is a widespread assumption that these will motivate at least some individuals to participate more frequently or for a longer duration. Badges in Social Media tend to be composed

of a title and small graphic (and usually display the criteria for which they are awarded). Antin and Churchill (2011) identify five primary functions of badges from a social psychology perspective: goal setting, instruction, reputation, status/affirmation and group identification.

`Reddit` utilises both badges and reputation scores. Badges are awarded for meeting certain criteria (e.g. having the top scoring post for a day, being a member for one year) and these are displayed through a user's profile. Reddit also keeps a record of two reputation scores for each user (known as 'link karma' and 'comment karma'), these scores are calculated simply by summing all of the scores for a user's posts and comments respectively. The regularity with which `reddit` users refer to karma scores indicates that they are perceived as influencing users behaviour. In these discussions karma scores are usually depicted as causing problems - with the common assertion being that people engage in certain behaviours which cause problems for the website (e.g. re-posting something which has recently been on the Front page) primarily to accumulate karma points. Badges are mentioned much less frequently on `reddit`'s comments pages, indicating that they may be perceived as less important. However, it must also be noted that there are several badges awarded for different types of voting behaviour - these represent one of the only ways in which voting users might receive recognition for their efforts, and it is possible that voting users are to some degree motivated by these badges.

There are however factors which make the study of badges and karma on `reddit` difficult. Reddit does not provide a list of the badges which are offered or the criteria for their award, and collecting data on which users have which badges is difficult because it requires one to query the profiles of users individually. While we will not explore the karma/badging systems of `reddit` directly, it must be noted that these exist and likely play a role in motivating some users to participate.

For a time both `reddit` and `Digg` provided a page which ranked users based on the success of their karma/reputation scores. However both websites have ceased providing these pages, because they were seen as encouraging competitive behaviours among users and attempts to 'game' the system - they also acted as a 'target list' for unscrupulous Social Media marketers who wished to utilise the popularity of 'power users' to lend credibility to the materials they wished to promote.

Lerman (2007a) studied the evolution of the ranks of `Digg` users (while these were still provided) and found that a higher rank helped a user to gain more 'followers' (or 'reverse friends'), which in turn increased the success rate of their submissions. They also suggested that users would engage in activity on the website in order to increase their ranking. These findings are suggestive of a feedback loop on `Digg` whereby successful posts lead to a higher rank which leads to more followers - giving the user an even higher success rate when they submit content. It is not difficult to see how this system encouraged the emergence of 'power users' (Arrington, 2006) on `Digg`, and the removal of the user ranking page can be read as an attempt by `Digg` to break this pattern of reinforcement.

3.3.2 Learning to participate effectively

The bi-directional nature of feedback received through DM voting systems means that they are much more than a proxy for attention. A user whose contribution receives more down-votes than up-votes is being informed that people have seen their contribution but that the majority of these people have not appreciated it. There is evidence to suggest that this kind of directional feedback can influence contributing individuals - with many users who continue to contribute seeing a long-term increase in the scores their content receives

Halavais (2009) looked at the comments of `Digg` users in terms of the commenter's experience. Halavais reported that the comments of 'more experienced users' are positively correlated with an increased number of `Digg`s, Buries and replies, although they do not specify the criteria for a 'more experienced user'. Halavais (2009) also report the median score received by comments in terms of the order they were posted in - for a user's 1st to 3rd comment this is zero, for the 4th to 12th it is one, and for comments beyond the 12th it is two. These findings cannot however be taken as particularly strong evidence for a learning effect as user dropout could account for this trend (do users who made negatively-rated 1st, 2nd and 3rd comments tend to make a 4th and 5th comment?). Halavais (2009) goes on to report a set of results which are more robust to user dropout. Looking at the first 30 comments for the sub-set of users who had made at least 30 comments, they found much smaller correlations between comment order and success. This analysis also revealed the presence of users who seemed to be deliberately seeking out negative reactions (these users are commonly referred to as 'trolls') - of 812 users studied 170 had a negative average over their first 30 comments and would fall into this category. It would seem that if feedback received through DM voting systems motivates users to participate it does not do so unilaterally. While many users might strive for positive ratings there are also users who are seeking a negative reaction.

Lampe and Resnick (2004) also looked for such an effect of users learning to submit 'better' comments. They used linear regression models to predict the score of users' 2nd comments, and included the score obtained by the user's first comment as an explanatory variable. This produced an unusual effect whereby positively rated first comments were associated with lower-rated second comments, and negatively rated first comments were associated with higher-rated second comments. This finding offers no evidence of users learning to post better comments through voting feedback and was interpreted as a regression to the mean.

3.3.3 Superparticipants

Previous research has documented the presence of a highly active minority of posters in various online discussion spaces - e.g. Twitter (Anstead, 2011) and a range of discussion fora (Wright, 2006;

Awan, 2007; Graham, 2013). Such highly active minorities are likely indicative of an underlying highly skewed distribution of activity between participants whereby most users have relatively low levels of activity and a significant proportion of contributions originate with a small number of users (perhaps similar to the power law distribution outlined in section 2.5).

These highly active minorities are often seen as dominating the discourse in their respective fora - 'domination' bringing with it certain negative connotations. Research which studies these minorities and the role(s) that they play is rare, with the work of Graham (2013) being one notable exception. In this paper the authors integrate quantitative and qualitative analyses of the moneysavingexpert.com forum - the former to detect 'superparticipants' (or more accurately 'superposters', a sub-group of 'superparticipants') and the latter to assess the role they play on the forum. They define super-posters as "any user that has created more than 2% of all messages on a forum with between 20,000 and 99,9999 posts and any user who has made more than 2,000 posts on a forum with over 100,000 thousand messages". In their sample 0.4% of posters meet these criteria, and between them they account for 47% of all posts.

In the qualitative analysis of super-posters' posts Graham (2013) reports that 81% of super-posters' posts were replies to other users. 36% of their posts were coded as storytelling or providing personal information - which was in turn interpreted as fostering everyday conversation. The second most common type of post for these users involved (often 'expert') 'advice giving/helping', accounting for 19% of their posts. In contrast, super-posters rarely made posts which were coded as degrading other users (2%) or curbing their capacity to be involved in the discussion (1%). As such, the moneysavingexpert.com forum's super-posters were found to play a largely positive role in the functioning of the forum. The paper outlines two further types of superparticipant (agenda-setters and facilitators) but the prevalence of these types of superparticipants on moneysavingexpert.com are not reported.

There is every reason to believe that DM websites will play host to superparticipants of various types. The presence of a voting system (and other modes of interaction) on DM websites raises the possibility of superparticipation which is defined not by the frequency or nature of posting but mediated by some other mode of interaction. On Social News websites, where voting is of central importance, it is easy to envisage superparticipants who are characterised by their voting behaviour - the frequency and/or nature thereof.

This suggests initial questions in the quantitative analysis of DM user behaviour - is there a highly active minority of users who we might term superparticipants? If we consider `reddit` to have three modes of participation (voting, posting, commenting) - would superparticipants tend to display trends associated with each of these modes (e.g. a generally high level of activity) or specialise in a particular mode of interaction? In other words, is superparticipation on `reddit` mode-specific?

Of the post types Graham (2013) coded for, 'curbing' (seeking to exclude others from the discourse) may be of particular interest as it relates to `reddit`'s voting system. In the context of discourse on `reddit` the down-vote button is effectively a 'curbing' mechanism. When a user down-votes the post or comment of another they are effectively pushing it towards exclusion from the discourse. If the item's score does not rise it will be seen by very few users - if it drops below a threshold it will literally be removed from the discussion and (in the case of comments) replaced by a 'comment below score threshold' link.

We can also speculate about how other types of superparticipant in Graham (2013)'s typology might map unto reddit's modes of interaction. The structure of `reddit` suggests particular locations in which agenda-setters might operate. To re-visit the reddit/forum analogy, subreddits are like individual boards, `reddit` posts the equivalent to forum threads, and `reddit` comments are like forum posts (although the comments themselves are nested within a threading structure). An agenda-setter might operate by making posts (with popular posts defining what a large number of *reddit*'s users will be discussing) or by making top-level comments (there are any number of discussions one can instigate based on, for example, a photograph - top-level comments are where these discussions are instigated).

One difference between reddit and a discussion forum is that a user's success in attempting to set the agenda will be defined largely by how users vote on their contribution. This is particularly so with regard to posts - which casual observation suggests will receive a number of comments orders of magnitude greater when they appear on the Front page as compared to the 'New' page. A post's capacity to elicit responses from those who see it will be secondary to its capacity to elicit up-votes - in determining whether that post sets the agenda (where 'setting the agenda' is defined as receiving a large volume of comments). Of course, there may be a substantial overlap between these criteria. It may also be possible for a post to be up-voted to prominence, receive hundreds of comments, and fail to 'set the agenda' for discussions of any merit.

3.4 How do the Distributed Moderation systems of Social News websites work?

Considerable efforts have been expounded investigating how these voting systems operate with respect to the ranking and sorting of items (Hogg and Lerman, 2007; Lerman, 2007a; Lampe and Resnick, 2004). This section provides an overview of these efforts.

3.4.1 Social Networks and Social News

Social News websites tend to include a networking component whereby a user can 'follow' or 'friend' other users, and an interface whereby a user can browse the contributions of these 'friends'. On `Digg` and `Slashdot` these networking components are intimately tied to the websites' functionality, whereas on `reddit` the capacity to add another user as a 'friend' appears to have been added as an afterthought and this feature is isolated from the workings of the website. The effect of these social networking components on the operation of the Social News websites in which they are embedded have been addressed by a number of researchers.

Lerman (2006) studied the voting behaviour of 'friends' on `Digg`, and found that users tended to like (or 'digg') stories submitted by those users they had designated as friends. This effect was particularly strong for 'early voting' (defined here as the first 25 `Digg`s received by a story) - with a user's friends accounting for a large proportion of the early `Digg`s their submissions received. This is a particularly important effect when we come to consider the importance of 'voting momentum' (i.e. gaining a place on the 'upcoming' or 'rising' pages) on Social News websites - suggesting that having a larger number of 'friends' on the website might help a user's submissions to pass this first hurdle towards Front page success. A weaker effect was observed whereby users tended to `Digg` the stories their friends had viewed and dugg.

Lerman (2006) goes on to compare `Digg` to `reddit`, in this case `reddit` serves as an example of a comparable site which did not employ a friending or networking component. Lerman (2006) finds that `reddit`'s system of deciding which items should appear on the Front page is weaker than `Digg`'s because posts on `reddit`'s Front page on average do not achieve as much of a 'boost' from placement here compared to stories on `Digg`. Lerman (2006) also reports that posts can reach `reddit`'s Front page with a much lower score than required for stories on `Digg` to do likewise, and this is cited as a possible reason for `reddit` operating on a faster timescale. The portion of this report which relates to `reddit` is out-dated (Reddit has changed quite significantly in the intervening period, most notably in terms of its general level of activity). Lerman (2006) also state that `reddit`'s 'Front page' algorithm utilises the votes of other users (in comparison to the browsing user's) to generate customised Front pages for each user (so if users A and B tend to up-vote the same stories, user A would be presented with stories which had been up-voted by user B). This is inaccurate (see section 5.1.2 for a discussion of `reddit`'s ranking algorithm).

Lerman (2006) on the whole espouses a positive disposition towards the 'social filtering' which results from the integration of social networking and Social News on `Digg`. However, they also recognise that such filtering based on network status (number of friends) could lead to a 'tyranny of the minority', where those users who are well connected have a much better chance of submitting stories which reach the Front page. These well connected users came to be known among `Digg`'s

community as 'power users'. The converse of an advantage for these power users is that new users who have no friends are hampered when they submit content. This goes against the ideal that items of content submitted to Social News websites should be appraised on their own merits - and ultimately the dominance of 'power users' on `Digg` came to be seen as a negative aspect of the website.

In fact, (Lerman, 2007b) re-visited `Digg` after some stories about the dominance of these 'power users' had forced `Digg` to change its promotion algorithm to reduce the effectiveness of 'bloc voting'.

Kunegis et al. (2009) have conducted social network analysis on the 'Slashdot Zoo'. `Slashdot` allows its users to designate other users as friends or foes and this zoo contains details of these relationships. Kunegis et al. (2009) found that negative edges (foe relationships) could be used to infer the existence of 'trolls' (users who deliberately provoke or mislead others) - and suggest a measurement which could be employed by `Slashdot`'s software to automatically identify these troll users and sideline them from discussions. These negative edges were found to conform to the principle of 'the enemy of my enemy is my friend'. `Slashdot`'s social network was also found to be a 'small world' network.

Ganley and Lampe (2009) also studied `Slashdot`'s social networking component, looking specifically at the relationship between users' 'karma' (Slashdot's reputation system) scores and the network's structure. The focus of this research is on social capital, specifically the way in which `Slashdot`'s social networking component allows this to develop, and the motivating effects this has on `Slashdot`'s users. Ganley and Lampe (2009) look specifically at structural holes in the `Slashdot` network, and at users who provide brokerage or closure. New `Slashdot` users find more benefit in associating with a broad group of individuals from different circles whereas more established users benefit more from a tightly connected circle of 'friends'. Ganley and Lampe (2009) relate their findings on network effects to the economic success of `Slashdot` - finding benefits of this (in that new users are more likely to maintain their activity and become regular users) but also some potentially negative consequences (power may become overly centralised within tight groups of established users).

Much of the research which has addressed Social News websites does so from the perspective of their economic viability and success - emphasis is placed on the ability of these websites to attract new users and maintain their participation. Network effects which are seen to encourage participation are appraised as positive aspects of the website. When these effects promote the emergence of 'power users' this is discussed in quite neutral terms. Power users may benefit a website through their high level of activity - they provide more of the 'raw materials' (submissions and ratings) which the website needs to function than other users, and consequently their presence improves the overall quality of promoted items and therefore broadens the appeal of the website. When the negative aspects of a website driven by power users is discussed these negatives centre

on the perception of new users and visitors. Power users are only seen as a negative when they are perceived by ordinary users as holding too much power - with these ordinary users being discouraged from participating due to the perceived elitism of the power users.

The present research is concerned with the social significance of Social News rather than the economic success of individual Social News websites. From this perspective Social News is of interest because it allows a single 'voice' to emerge from a large group of independant users in an organic fashion. My perspective on power users is thus rather different. Where a small group of power users exerts disproportionate influence on the DM process this erodes the ultimate potential of the DM system - the resource produced by the Social News website is largely a product of its power users, with ordinary users occupying a peripheral role in the production process. In we consider the *potential* of Distributed Moderation systems as arising from their capacity to aggregate the opinions of many individuals - then the presence of power users must be perceived in a negative light because their disproportionate power 'distorts' the collective decisions arrived at through the voting system. The presence of power users on `reddit` is considered further in section 6.1.6.

3.4.2 Modeling the 'popularisation' process

Lerman (2006) looked at 2858 stories submitted to `Digg`'s technology section by 1570 users over the course of one day, and followed these stories for a period of six days. Of these followed stories 98 (submited by 60 different users) reached `Digg`'s Front page. Lerman (2006) found that these stories tended to accrue 'diggs' at a constant rate (specific to the story), but that once a story had been promoted to `Digg`'s Front page this rate increased substantially. They also found that, somewhat paradoxically, stories tracked during the observation period which were submitted by the top-ranked users did not receive the most Diggs - but that over a much longer time period these top-ranked users accounted for a disproportionate number of Front page stories (15,000 Front page stories were submitted by the top 1,020 `Digg` users, with 35% of these stories coming from the top 3% of these users). This suggests that power users regularly submitted 'good' stories, but that 'great' stories were more likely to come from less established users.

These two groups of users can be thought of as serving different purposes on `Digg`. When an established power user submits a post there is a high chance that this post is good. If one is looking for good posts to promote to the Front page there is a high probability that one will find them if one looks at these users' posts. On the other hand there are thousands of users without this kind of established record, these users' posts likely fall on a continuum of 'terrible' to 'great' but if we start browsing through these the chances of a specific post being good or great are quite small, there will be many poor posts to sift through to find these rare gems. From the perspective of `Digg`, allowing power users to emerge is a somewhat lazy approach to guaranteeing the supply

of good content for the Front page. From the perspective of the ultimate potential of Distributed Moderation this seems like a wasted opportunity. If we have thousands of people browsing and rating items of content their efforts would be more useful if they were put towards sifting through the mass of new submissions to find the rare gems, as opposed to checking out the latest post from 'Power User X', and if it's good putting it on the Front page as quickly as possible.

Hogg and Lerman (2007) produced a stochastic model of user behaviour on Social News websites, and used this to generate predictions about the success of individual posts on `Digg`. This model included rates at which users came to `Digg` and viewed content on the 'hot', 'upcoming' and 'friends' pages. For each post a prediction of the number of votes which would be received was generated based on the post's 'interestingness' (measured by the post's voting record at the first four observation points, the last of these generally around four hours after post submission) and the number of friends the submitting user had on `Digg`. The number of votes a post would receive, as predicted by the stochastic model, showed an 87% correlation with the actual number of votes received. A simpler model which did not account for the number of friends the submitting user had and relied only on the post's early voting rate showed a lower correlation of 75% with the actual number of votes received. This suggests that consideration of a user's social network on `Digg` adds to the model, but that most of its predictive power comes from the inclusion of early voting. One of Hogg and Lerman (2007)'s conclusions is that network effects on `Digg` interact with the inherent 'interestingness' of posts. Well-connected users find it easier to submit posts which are promoted to `Digg`'s Front page, votes cast by their 'friends' help with this, particularly in the early stages of a post's lifespan. However, it is still possible for poorly-connected users to submit posts which have a high 'interestingness' value and reach the Front page, and also for well-connected users to submit uninteresting posts which do not reach the Front page.

While models which predict post popularity with 'early' voting are of interest one must be cautious not to over-state their utility. Four hours is a long time on a website like `Digg` or `reddit`. Within four hours of submission a post will have been seen by a number of users and it is possible for a post to reach the Front page within four hours. These results suggest that within four hours posts will have been seen by enough users to reliably extrapolate whether the post will reach the areas of highest visibility and appeal to individuals who see them there. Szabo and Huberman (2010) report on a number of similar models and find that they can make reliable predictions of a post's 'final' score (after 30 days) within two hours of submission (although they appear to have studied only posts which reached a certain level of success). Of interest here is the comparison with `Youtube`, where reliable predictions were generated based on a video's views over its first seven days. This is an indication of the rapid speed at which Social News websites operate, but considering that a post's active lifespan is rarely longer than 24 hours it is not surprising that its eventual score can be predicted based on a few hours of voting activity.

Lerman and Galstyan (2008) also considered the popularisation of posts on `Digg` in terms of network effects - but in this case the presence of early votes which did *not* come from the submitting user's network of friends was taken as a marker of a story's 'interestingness'. In this case 'interestingness' was classified as a binary variable, with posts which received a score of 520 or greater being classed as interesting and those which achieved a score of less than 520 being classed as uninteresting. Lerman and Galstyan (2008) looked in detail at the first 10 votes received by newly submitted posts and found that when many of these initial votes came from the submitting user's friends these posts would tend to ultimately be classified as uninteresting. Posts which received fewer votes from 'friends' within the first 10 votes were more likely to ultimately achieve a high score. This effect comes about because in the time it takes 10 of a user's 'friends' to see their new post and vote on it some number of 'non-friends' will also see the post in the upcoming queue; if these 'non-friends' Digg the post it is likely a truer reflection of that post's inherent 'interestingness' than the votes of friends. When all of a post's first 10 votes come from friends it can be inferred that other users have seen the post but not voted on it - and therefore that the post is probably of little interest generally.

Lerman and Galstyan (2008) adopt the position that this aspect of `Digg` voting could be used to improve the promotion algorithm. `Digg` requires around 40 votes to decide if an item should be on the Front page, but the method of looking at early 'non-friend' voting seems to make reliable decisions after just 10 votes (which seems to indicate that 'friend' voting adds little to the process of finding the 'best' items, an intuitive implication - if one wants impartial judgments of an item's quality one does not ask its producer's friends). However, an important aspect of these analyses is that only posts which were submitted by users in the 'top 100' and received at least 10 votes were considered. This paper should be thought of as dealing with the difference between posts which are successful and highly successful - it has little to say about the submissions of poorly-connected `Digg` users, and whether the many submissions made by this type of user will achieve Front page success or no-votes obscurity.

Hogg and Szabo (2009) studied `Essembly.com`, a website geared towards political discussion that incorporated elements of Distributed Moderation and social networking (`Essembly.com` has been offline since May 2010). Hogg and Szabo (2009) studied the distribution of activity between users and between items of content, in order to produce a model of voting activity on the website. This research observed high drop-out for new users of the website, and it was found that users who continued to participate did so at a set rate (whether a user continued to participate and their activity level are due to prior individual differences, and are not a result of their experience on the website). This research also found that recently submitted content was the focus of user voting activity.

Daily cycles on `Digg` and `reddit`

Two studies have documented daily cycles on `Digg` and `reddit`. Szabo and Huberman (2010) found that rates of `Digg`ing and submission activity were three times higher during the day as compared to at night (assuming that `Digg` users are largely North American). Weekdays were also found to be 50% more active than weekends. Virasoro et al. (2011) documented similar temporal cycles on `reddit`.

3.4.3 Do Distributed Moderation systems produce good collective decisions?

One of the primary functions of a DM voting system is to rank submitted content such that the best submissions receive the largest share of user attention. As such, assessing the performance of the DM system with regard to this task is of central concern to the present research. Every item submitted to a Social News website receives (or does not receive) a voting response from users. This response determines whether the item will be seen by many or few users, and can be treated as the voting system's 'judgment' of the item's quality (with items appearing on the Front page being judged the most worthy of attention). If an external measurement of an item's quality could be generated this could be compared with the judgment of users expressed through the voting system. Efforts to generate such an 'external' measurement are reported in section 7.3.

In addition to asking whether the collective decision reached through the voting system is good, it is also important to ask whether it is reliable. For example, will voting users on the 'pics' subreddit always make the same judgment about a particular item, or is this judgment susceptible to random or 'social influence' effects? If every item were to receive 20 votes at minimum, this would likely lead to judgments which are quite reliable as 20 users would represent a reasonably reliable sample of the voting user population. However, if an item which receives 2 initial down-votes is unlikely to receive any further votes then this is likely to produce judgments which are much less reliable and more susceptible to order effects (as a sample of 2 users will not reliably reflect the opinion of the voting population).

An excellent study conducted by Salganik and Watts (2009) should inform efforts to address this question. Salganik and Watts (2009) conducted a web experiment in which 2,930 participants listened to, rated, and downloaded 48 songs. These participants were allocated to independent 'worlds', and given feedback on the rating behaviour of other individuals in their own 'world' but not others. The rankings of these songs showed significant variation between different worlds. The judgments of participants were found to be highly susceptible to social influence effects - participants in a given world were influenced by the previous actions of participants within their world. This produced an effect whereby the distribution of ratings and downloads between songs

was itself stable between worlds, but the ranking of a particular song showed considerable variation between worlds.

Such an experimental design could be employed equally well in the context of Distributed Moderation. For example, independent subreddits could be created and populated with the same items of content, and participants recruited from the same population could then be assigned randomly to one of these independent subreddits. The outcome of the study would be the rankings assigned to each item of content through participants' up/down voting decisions, would these rankings show higher or lower levels of agreement between subreddits than the 'independent worlds' of Salganik and Watts (2009)?

3.4.4 How can Distributed Moderation platforms be improved?

A number of researchers have considered ways to improve the effectiveness of Distributed Moderation systems.

Most examples of Distributed Moderation elicit ratings from users on a single dimension, for example `reddit`'s up/down voting system produces a score on one dimension for each item. Furthermore, it is not always clear what the dimension items are rated on is intended to represent. In `reddit`'s case votes can be thought of as representing whether the voting user thinks the item of content is worthy of attention from other users, but there is rarely a clearly defined and commonly accepted understanding of what criteria these judgments should be made on. These criteria also vary substantially between subreddits (for example one can imagine the criteria being applied are quite different between the 'funny' and 'science' subreddits).

Lampe and Garrett (2007) considered the performance of ratings systems that utilised differing numbers and types of dimensions on `NewsTrust.net`, a social news-rating website. Lampe and Garrett (2007) conducted an experiment in which participants were recruited to rate one of two versions of a single news story. Participants were allotted to one of four conditions where they rated the story on either a single dimension (general quality), 6 dimensions, 8 dimensions or 13 dimensions. Additional dimensions beyond that of general quality reflected criteria like accuracy, credibility and fairness. The results of this experiment suggested that a single dimension was the best approach for discriminating between the high and low quality versions of the story, but that the condition with six normative dimensions was better at producing ratings which matched those of experts. There are however some disadvantages associated with increasing the number of dimensions which items are rated on, most notably the increased burden placed on the users who are to provide these ratings. There is also the question of how to use and present ratings on multiple dimensions, in the case of `NewsTrust.net` an average of the dimnsions is taken to use as a summary measure.

If Social News websites like `reddit` were to expand the number of dimensions on which items are rated, they would also need to calculate an average or summary measure to use in ranking the items, or offer users the option to see ranks based on their chosen criteria. However, if users were offered the option of seeing content which scored highly on a selected dimension this would have the effect of fragmenting the audience of Social News readers to some degree (i.e. people who want to look at posts ordered by 'fairness' would see different posts on their Front page as compared to people who wanted to order posts by 'originality'). The manner in which Social News websites focus the attention of their readers on the same items of content is likely one of the factors which gives these systems their appeal and strength (and this may allow them to work *against* fragmentation)(Sunstein, 2002). This, coupled with the increased burden on voting users, means that it is unlikely a Social News website like `reddit` will add multiple dimensions to the voting system as `NewsTrust` has done.

However, there may be some potential in adding a single secondary dimension for certain subreddits on `reddit`. Certain large subreddits that deal with content which requires more detailed appraisal appear to have a problem whereby poor quality posts are often promoted to the top of the subreddit. To take /r/science as an example, it is quite common for posts which sensationalise some research finding, or which concern poor-quality research, to be up-voted to the top of the subreddit. These posts will often have top comment(s) that criticise the post's content heavily, this is unusual as the same voting system is used to rank posts and comments and the votes are (presumably) coming from the same population of users. The problem here may be that users who vote on posts are applying the same criteria ('interestingness', for want of a better word) they apply in other subreddits (like /r/pics and /r/funny) when deciding whether to upvote or downvote posts on /r/science. This could result in posts about a cure for cancer (that are based on shoddy research) outperforming posts about solid research which doesn't grab users' attention in the same way. Perhaps there may be some virtue in adding a second dimension which explicitly addressed the accuracy of the post to a subreddit like /r/science.

These suggestions are about improving the utility of Social News sites, streamlining them to perform their tasks more effictively. We might also ask whether the 'public good' credentials of Social News, and DM more broadly, can be improved upon (Sunstein, 2002). The question of how DM could be improved or adapted is re-visited in Chapter 9.

3.5 The differences between Social News websites - structural or social?

In comparing two Social News websites differences will be observed - these may be due to differences in the structure of the website and its software, or the social norms and values of their users. Picking apart these two sources of variance is likely to be difficult, but in order to understand how Distributed Moderation works it is necessary to understand which aspects of its functioning are determined by the configuration of its software, and which are determined by the actions of its user-base.

Some important lessons can be learned here by making reference to research which has been conducted with the use of 'wikis' and `Wikipedia` in particular. The wiki is a particular type of software which allows users to add to and modify the information resource under construction - `Wikipedia` is but one (very well known) application of this standardised piece of software. Many wikis are ultimately unsuccessful in meeting the goal for which they were created, by looking at the behaviour of the groups of users who have participated in these endeavours with varying levels of success it should be possible to identify the aspects of group behaviour which are associated with a positive outcome. In `Wikipedia`'s case as the endeavour has grown in scale the community of Wikipedians has developed ever more nuanced approaches to organising the work-flow related to the project and making difficult decisions about individual articles (Stvilia et al., 2008). Many of these procedures and social practices have been explicitly defined by the `Wikipedia` community.

The same kind of approach can be applied to Social News websites (or any other class of Distributed Moderation website) - but this is complicated by the fact that Social News websites deploy differing software, and for many the specifications of this software are not publicly available. One way to circumvent these problems is to make comparisons between different Social News websites, and also between different parts within the same website. For example, `reddit` and `Digg` might be compared on measures such as the number of individual items which are featured on their Front pages, the rate of turnover thereof, the number of votes and comments these items receive, and the duration of their active lifespan on the website - differences observed on these measures are likely the result of differences in the software deployed by the two websites.

In order to look for differences with a social origin, individual sections of the same website could be analysed. For example, all of the subreddits on `reddit.com` are governed by the same software, but due to the subreddit subscription system these are likely to have distinct user-bases - although the degree to which this is the case is a separate question which should first be addressed. Provided that subreddits with largely distinct user-bases can be identified, it would then be possible to analyse the differences in the performance of these subreddits, with any observed differences being due to the distinct user groups who 'run' these subreddits. Section 5.1.12 looks for distinctive

patterns of behaviour on the 'IAmA' subreddit.

Reddit actually provides a set of voting guidelines (known as 'rediquette', see `http://www.reddit.com/help/reddiquette`) - the extent to which some of these are adhered to (e.g. 'Don't Plead for votes in the title of your submission.') could be subjected to empirical scrutiny. There are however many guidelines which relate to user voting behaviour (e.g. 'Don't Downvote opinions just because you disagree with them.') - and as voting behaviour is largely anonymous adherence to these guidelines would be much more difficult to investigate.

3.6   Q&A or Knowledge Exchange websites - `Stack Exchange`

Question and Answer websites utilise many of the same tools as Social News but to different ends. On Q&A websites the primary units of content are questions and these are subjected to a ranking-by-votes system. Secondary units of content are answers to these questions, and again these are subjected to ranking such that the best answers should be displayed most prominently. Therefore on Q&A websites we can think of questions as being analogous to posts on Social News websites, and answers as analogous to comments. In practice these websites share little similarity to Social News sites beyond the presence of a voting system. `Yahoo Answers` and the `Stack Exchange` network provide the most recognisable Q&A websites, the present research focuses on the `Stack Exchange` because cursory examination suggests that the `Stack Exchange` network provides a more valuable resource and the data which will allow us to probe the workings of the `Stack Exchange` is easily obtainable.

It is important at this stage to clarify the distinction between the `Stack Exchange` and `Stackoverflow`. The `Stack Exchange` is a network of websites all run by the same company using the same software. `Stackoverflow` is one of these websites which deals with programming questions - it was the first to be launched and remains the exchange network website with the highest usage statistics. Researchers have tended to focus on `Stackoverflow` because it is the most recognisable and active of the `Stack Exchange` websites.

3.6.1   History

Stack Exchange is a network of Question and Answer websites which all run the same software and are owned by the same company (Stack Exchange Inc.). The first website in what is now the `Stack Exchange` Network was launched in 2008. `Stackoverflow.com` is dedicated to Q&A concerning computer programming. `Stackoverflow` remains the largest of the `Stack Exchange` websites, with 2.2 million questions (80% of which have been answered), 827,000 users and receiving

2.7 million visits per day.[6] In 2009 two further `Stack Exchange` websites were launched, 'Super User' and 'Server Fault' - and together with `Stackoverflow` these remain the largest three `Stack Exchange` websites.

In February 2010 `Stack Exchange` Inc. released a product/service whereby third parties could pay to have their own Q&A website produced using the `Stack Exchange` software and hosted by `Stack Exchange` Inc. In April 2010 this approach was abandoned in favour of a system where all `Stack Exchange` websites were owned and maintained by `Stack Exchange` Inc. with user contributions (questions and answers) falling under a Creative Commons license. A new sub-site ('Area 51') was created where `Stack Exchange` users could collaborate to decide which subjects or topics should have their own `Stack Exchange` website.

3.6.2 The purpose of `Stack Exchange`

Stack Exchange is comparable to Social News in that it applies up/down voting to rank and sort units of content... however the intended outcomes of this voting activity, and the goals of the website as a whole, are quite different. Where `reddit`'s purpose is to rank and sort items according to some vaguely defined notion of 'interestingness', the goal of the `Stack Exchange` websites is to produce, and make easily accessible, good answers to questions. Long et al. (2011) identify three basic requirements of a Q&A website: that new questions recieve answers quickly, that provided answers are of high quality, and that questions and answers are organised so that they are easier to search. This difference in the aims of the website results in some easily observable differences in the website's layout and structure.

One of these goals is to produce answers to questions, and therefore the website is structured to focus user attention on new and unanswered questions. This is immediately apparant on navigating to the `Stackoverflow` Front page (`stackoverflow.com`) which showcases new questions that have been submitted within the last hour and do not have accepted answers. This is an immediate departure from `reddit`'s structure where the Front page showcases the posts which have been rated most positively. `Reddit`'s Front page displays the 'outputs' of its DM voting system, `Stackoverflow`'s Front page is geared towards ensuring that user 'inputs' are directed to the appropriate places (i.e. unansnwered questions). On `Stack Exchange` options exist to browse questions based on voting response or combined level of activity (views, answers and votes) for a specified interval but these play a much smaller role than the ranking by votes pages of Social News websites.

In order to facilitate the development of good question/answer diads both types of content can be edited by a user other than the user who origianlly submitted them, a feature more commonly

---

[6]These figures were observed in September 2011

associated with wikis. On Social News websites a post or comment is owned by its submitting user and can only be edited by this user - on `Stack Exchange` users with more than a certain number of 'reputation points' can edit the questions and answers of other users.

A second goal is to make the answers to questions accessible and to this end a tagging system is employed. Each question can be assigned a number of tags which identify its subject, and these tags can then be used to navigate to questions about a specific subject.

3.6.3 Stack Exchange's structure

The structure of `Stack Exchange` can be compared to a Social News website like `reddit`. In the same way that `reddit` is comprised of many subreddits, `Stack Exchange` is comprised of a number of exchanges. However, `Stack Exchanges` seem to operate more independently than subreddits. The Front page for `StackExchange.com` showcases popular questions from a range of the exchanges, in a similar manner to `reddit`'s Front page showing top posts from default subreddits. However, `StackExchange` does not offer the option to browse posts based on specified criteria in this aggregated form. For example, one cannot ask to see the top-scoring questions for the last week from all stack exchanges. Instead, the `StackExchange.com` Front page seems to operate as a portal which serves to direct a new user to the appropriate `Stack Exchange`(s) for their needs. If one sees an interesting question on the `Stack Exchange` Front page and clicks through to it, one is immediately transported to the exchange where that question is housed, and all of the options to continue browsing relate to browsing on that specific exchange.

If the `StackExchange.com` Front page serves as a portal where new users can find individual exchanges it would be expected that individuals who use `Stack Exchange` websites regularly would do so by browsing directly to the exchange they want. This is supported by traffic measures for the `StackExchange.com` and `Stackoverflow.com` addresses - with `Stackoverflow.com` (the largest exchange) having traffic statistics orders of magnitude greater than `StackExchange.com` (the network which `Stackoverflow` sits within)(Alexa, 2010).

The primary units of content on a `Stack Exchange` website are Questions. The secondary units of content are Answers. Answers are always tied to a specific Question, and a Question can have more than one Answer. Questions and Answers are both subjected to up/down voting. Up/down voting on questions determines whether a question will appear on pages which select items based on the voting score criterion. Up/down voting on answers determines the order in which answers will be displayed on the page for the question which they relate to. The user who submits a question also has the option to accept one of these answers as having answered their question, once an answer is accepted it becomes the most prominently displayed answer on the page.

3.6.4 Literature concerning Q & A or Knowledge Exchange Websites

`Yahoo Answers` (YA) is the largest Q & A type website and has been the subject of a number of studies (Adamic et al., 2008; Kim and Oh, 2009; Harper et al., 2009). Adamic et al. (2008) clustered YA forum categories according to content characteristics and the patterns of interaction between users. They found that YA forums supported a number of different forms of interaction - with some conforming to the expectations of an expertise sharing space while others incorporated discussion, advice and support. They also considered the participation of users and found that some focused narrowly on specific topics while others participated broadly accross categories. Users with a narrow focus (lower entropy) were found to submit answers which achieved higher ratings, but this effect was specific to categories dealing with factual expertise.

The `Stack Exchange` network has attracted the attention researchers more recently. Over the last few years the `Stack Exchange` network has grown to the point where it is a very widely known and heavily utilised resource among computer programmers. `Stack Exchange` websites also make large portions of the procedural data they generate every day accessible to the public - it is therefore likely that these websites will see increased attention from researchers in the coming years.

Mamykina et al. (2011) provided a range of impressive descriptive statistics for `Stackoverflow` as part of a thorough investigation of how the website works and why it works so well. For instance they report that questions tend to receive answers within a median 11 minutes, and tend to receive an answer which is ultimately accepted within a median of 21 minutes. These are impressive figures for a freely accessible resource which anyone can ask a question of and which deals with questions of a highly specialised nature. The scale at which stack overflow operates is also impressive, containing 800,000 questions and 2.2 million answers as of September 2010.

Mamykina et al. (2011) also report a number of findings relating to the users of `Stackoverflow`. As in many other examples of Peer Production the distribution of activity between users is highly skewed and approximates a power law. Users tend to specialise in asking or answering questions, with 21.4% of user accounts at the time having both asked and answered a question. Highly active users tend to answer rather than ask questions. Mamykina et al. (2011) identified four distinct groups of users based on their 'activity signiture' - a string of characters denoting whether the number of answers provided in each month was low or high (with < 20 answers being classed as low and >= 20 being classed as high). The group they termed 'community activists' ('users who are highly active on the site for multiple months') made up 1% of the population and supplied 27.8% of the answers. 'Shooting Stars' ('a single short period of high activity followed by low activity') made up 4.2% of the population and provided 21.9% of the answers. 'Low-Profile Users' ('who have intermittent activity, but who never become highly active') made up 94.4% of the population and provided 34.4% of the answers. The remaining 15.9% of answers came from users

who were not signed into an account. These figures illustrate the consequence of a power law type distribution of activity between users - in `Stackoverflow`'s case a small number of users are responsible for a considerable proportion of the answers submitted.

Oktay et al. (2010) produced an excellent paper on `Stackoverflow` which coupled an understanding of SO's history with analysis of data from the website to produce a series of quasi-experimental designs (QED). This kind of approach holds promise for gaining a causal understanding based on observational data, and may be applicable accross a range of Social Media. Oktay et al. (2010)'s first QED utilised a matched design whereby pairs of units with high similarity were identified where one had received a 'treatment' and the other had not. This design was applied to determine whether a pre-existing high quality answer for a question discourages other users from posting a subsequent answer. In this case the treatment was the presence of an 'accepted' answer and the dependant variable was the rate at which new answers were submitted. This analysis was repeated three times with varying levels of matching for the question pairs - under the most stringent question matching there was no significant effect on answer rate of an accepted answer being submitted.

A second QED used a change in `Stackoverflow`'s software as the treatment and compared units of content immediately before and after this change. This software change concerned the manner in which ties were broken in determining the display order of answers which had the same score. The design was used to ask a question about whether answer display order was related to the rate at which answers received up-votes, no evidence for this was found.

The QED which is of greatest interest for the present research concerned the motivational effects of the 'Epic' badge. The 'Epic' badge is awarded to users who hit the daily limit on the amount of reputation points they can earn 50 times - as such it is one of the more difficult badges to earn on `Stackoverflow`. The QED used to answer this question utilised an interrupted time-series design whereby the outcome variable (in this case daily number of posts) was observed for a certain time interval before and after the treatment (in this case being awarded the Epic badge) occurred. 54 users had earned the Epic badge when this study was conducted and these were all included in the QED, their number of posts per day was calculated for each of 30 days before and after they earned the Epic badge and these values were normalised by subtracting the user's average daily posting rate. Then two linear regression models were fitted, the first to daily answer rate before the badge was awarded and the second to the daily answer rate after the badge had been awarded. In this case users' daily answer rate was stable before being awarded the Epic badge but started to decline after the badge was awarded. The drop-off in rate of answer submission after the badge was earned suggests that the availability of the badge was motivating users to post - once the badge had been earned this motivation evaporated and the rate of post submission declined. This is an important finding because it represents one of the first pieces of empirical evidence that digital

badges affect user behaviour.

3.7   Guiding Questions

Before it is possible to address questions relating to the social implications of Distributed Moderation - it is first necessary to understand how websites which employ these systems function in practice beyond the basic principle of up and down voting. We outline here some questions which will guide the investigation of the DM systems under scrutiny. It is possible to approach these systems from the perspective of the website's content, and also from the perspective of its users.

The structure of a Social News website, and the importance of placement on their 'Front page', suggests that the distribution of user attention and voting activity will be highly skewed. It is expected that items appearing on the 'Front page' receive user attention and votes at a rate which is much higher than elsewhere on the website - therefore it is expected that these items will account for a disproportionately large share of votes cast by users. It is important firstly to look for such an effect, and then to attempt to quantify its magnitude. Given that a small number of posts receive a disproportionately large share user attention and voting activity is it possible to identify factors which can predict the ultimate success or failure of a post on `reddit`? The lifespan of `reddit` posts is also of interest here, how long do `reddit` posts remain active for? Are there a series of stages which a post must progress through before it reaches the website's Front page? A related issue concerns the pages `reddit` uses to display content - how is user attention and activity distributed between these pages? How much attention and voting activity can a new `reddit` post expect to receive relative to a post which has progressed through the voting system to reach the Front page, the 'Rising' page, or the 'Main' page for its subreddit? Chapter 5 deals with questions such as these.

How well do Distributed Moderation systems perform on their task of making the 'best' submitted content most visible? This is an important but difficult question to address, requiring an external measure of the item's 'quality' to compare with the item's voting performance. Do DM systems make reliable decisions or are they susceptible to social influence effects through voting order? Section 7.3 addresses these questions. The 'rich get richer' mechanism underpinning Social News, whereby positive votes lead to more attention and more votes, suggests that early voting on an item may hold much influence over its ultimate success or failure (see section 5.1.5).

By looking at the qualities of content and its `reddit` voting record it is also possible to address a range of questions about the effect of `reddit`'s voting system on the discourse which takes place on the website. If `reddit` users hold differing opinions on a topic (e.g. the `Wikileaks`' release of United States' diplomatic cables) will an opinion held by a minority of `reddit` users be expressed in high-visibility areas of the website? Does `reddit`'s voting system marginalise minority

perspectives on a subject, and if so to what degree? If, for example, 70 % of `reddit` users are pro-`Wikileaks`, does this mean that 70% of posts about the topic which reach the Front page will be pro-`Wikileaks`? Or is this percentage likely to be closer to 100, because a 70% majority of pro-`Wikileaks` users could down-vote (and filter out) content which is anti-`Wikileaks`? Section 7.1 considers these questions through a case study of `reddit`'s coverage of `Wikileaks`' release of United States' diplomatic cables. A series of more fundamental questions can also be asked using this data, for instance which types of source do `reddit` posts link to? Which stories are 'covered' on `reddit`? And how does this 'coverage' differ from coverage in the conventional news media? Does the 'news' processed by Social News websites fit with a more conventional understanding of the term 'news'? Or is the content promoted by these websites radically different to that which we associate with the concept of news?

Social News websites can also be approached from the perspective of their users. What motivates an individual to participate on a Social News website? Users of SN can participate in a number of ways: by voting, submitting content, or commenting on the content submitted by other users. Do users tend to participate in all of these ways or do individuals specialise in particular forms of participation? Are Social News websites host to a small minority of highly active users who dominate the discourse? If so, what kind of 'superparticipants' (Graham, 2013) might be operating on these sites?

`Digg.com`, a Social News website, has had well documented problems in the past with so-called 'power users' who had a much greater input into the collective decision about which stories appeared on the Front page than other users (Lerman, 2007b). `Reddit` has also seen a number of attempts to 'game' its voting system, although the general consensus is that `reddit`'s voting system is more robust to these attempts than `Digg`'s. Nevertheless, it may be possible for `reddit` users to adopt particular strategies or patterns of behaviour which result in them having greater input to the collective decision about `reddit`'s Front page. Do `reddit` users adopt such strategies or patterns of behaviour which aim to maximise their impact on the site? And if so, how effective are such strategies? Chapter 6 concerns DM users and addresses some of these questions.

How does participation on Social News websites affect those users who participate? Users who make successful submissions can quickly find themselves to be the centre of other users' attention for the brief interval in which their submission occupies a position on the website's Front page. How does such fleeting attention affect these users?

Users who vote on Social News websites are contributing to a collective decision about which items will receive a lot of attention, does this induce a positive affect in these voting users? Is there a directional effect determined by whether other users vote in agreement? Do users experience a negative affect when the collective decision made through the voting system is opposed to their own opinion on an item's merit?

Of the users and visitors who browse the Front (and other) pages of Social News websites, do the pre-existing judgments registered by other users affect the browsing users' perception of the items being voted on? For example, does seeing an article on `reddit`'s Front page with thousands of up-votes lead individuals to appraise it more positively than if they had arrived at that article through some other means (e.g. finding the article through a search engine)? How much 'faith' do Social News users and visitors place on the voting judgments registered by other users? Is 'faith' an appropriate concept through which to frame the question of social influence on Social News? Might 'peer pressure' or 'conformity' be better concepts to describe this influence?

This research considers Social News and Q&A applications of Distributed Moderation, are there other purposes or tasks for which Distributed Moderation could be utilised? Distributed Moderation has already been applied to political policy ideas, what can we learn from these failed attempts? Section 5.3 considers some data from `Your Freedom`.

Are there ways in which Distributed Moderation systems might be improved within the contexts where they already operate successfully? How might DM systems be adapted to provide a better fit for alternative applications? This issue is considered with respect to politics in section 9.2 towards the end of the thesis.

Chapter 4

Methods of research and data collection

The present research for the most part utilises naturally occurring data - records or observations of activity taking place on the websites of interest. The use of such naturally occurring data is supplemented with the collection of survey and experimental data. This chapter provides an overview of the methods used to collect the various forms of data which will be analysed in subsequent chapters.

4.1 Obtaining data directly from websites of interest by request

One of the advantages of studying social interactions which take place online is the availability of detailed records of this interaction. The human activity which powers Distributed Moderation systems is meticulously recorded on the servers of the websites which employ these systems - such records are required for these systems to function. A Social News website will contain on its servers a record of every user submission, vote and comment which has taken place. For a researcher who wishes to probe the workings of these websites access to the procedural data stored on their servers is invaluable.

As such, at the start of the present research three large Social News websites (`reddit`, `Digg` and `Slashdot`) were approached and procedural data were requested for research purposes. Of the three websites, `reddit` was the only one which consented to provide access to some of this procedural data. The data which was provided took the form of a list of 3.5 million votes cast in the month of March 2009. The record for each vote contained variables representing the direction of the vote

(up or down), a timestamp for when the vote was cast, and ID numbers representing the user who cast the vote and the post it was cast on. Further requests allowed us to supplement this data with a record of which user submitted each post, and the subreddit which the post was submitted to. It should however be noted that after this initial period of contact `reddit`'s administrators ceased responding to the researcher - and therefore could not be relied upon to provide further data-sets or information on any aspect of how `reddit` functions.

4.1.1 Handling and transforming back-end procedural data

The first problem encountered with this data relates to its size. The file received from `reddit` is around 200 megabytes in size, too large for the statistical environment `R` to open at all, and too large for `SPSS` to handle with any degree of efficiency. Furthermore, in its initial form (a list of 3,446,522 votes) the data were not suitable for analysis - many transformations would be required to work this data into a usable format. For these reasons the use of databases was explored.

`MySQL` was selected as the preferred database type because it is open source, easy to use, and reasonably powerful. The availability of `PhpMyAdmin` was a further reason for the choice of `MySQL`. `PhpMyAdmin` provides a web-browser interface for `MySQL` databases, this makes it much easier to browse and edit the database than would be the case using the command line exclusively. A further reason for the choice of `MySQL` is that the statistical environment `R` has a library (`RMySQL`) which enables `R` to retrieve and update records on the database with ease.

The first step towards transforming this data into a usable form was to create tables representing each post and user in the data. A separate list of post IDs was supplied with the author ID and time of submission for each post; this formed the basis for the posts table. The procedure for adding variables to this table was to compile the list of unique post IDs, then loop through this list selecting all of the related votes, calculating aggregate measures for the post and adding these to the relevant row of the posts table. As there were 352,902 unique posts in the data, executing these scripts involved making 352,902 separate calls to the votes table, a lengthy process (taking 12-24 hours) which was repeated several times as new variables were conceived of and added. The first such script generated basic measures for each post: the number of votes the post received and the number/proportion of these that were upvotes and downvotes. Later in the research process further variables were added: for example representing the subreddit the post was submitted to, the post's order (where a user submitted more than one post in the month this represents the order of submission, of interest chiefly for new users), and the day and hour of submission.

Similar procedures were employed when generating the users data table, with many more useful variables being created for users. There were 102,232 active users in the data. Again the first step was to compile basic information about the number of votes a user cast and whether these were

upvotes or downvotes, and also the number of submissions a user made and the proportion of their activity this behaviour accounted for. User variables were also created which (by interfacing with the posts table) assessed the performance of the submissions a user made. It was also possible to estimate the age of a user's `reddit` account based on the ID numbers, and by close inspection of users data it was possible to identify new accounts - those created within the month covered by the data. Later in the research process, when some interesting patterns of user behaviour had been identified, this table was returned to and augmented further with counts of the number of times a user engaged in these interesting behaviours (e.g. the number of times a user cast an 'early' vote, the number of times a user voted or submitted on an 'obscure' subreddit).

The initial votes table also saw significant augmentations. For each vote its 'post order' (i.e. the Xth vote to be cast on a given post) and time after post submission was calculated and stored. Similar metrics were created on the basis of the user who cast each vote - the 'user order' (i.e. a user's Xth vote in the month) and the delay between the vote in question and the user's previous vote. The Users table was then supplemented with variables based on these measures - e.g. whether a user tended to vote early or late in the voting lifespan of a post, and whether the user tended to vote quickly or slowly (average delay between votes).

The final major data table to be created concerned subreddits, this was produced after a further request for information from the `reddit` administrator who provided the data. There were 2,184 subreddits which saw activity during the month covered by this data. This table included the name and ID for each subreddit and was augmented with the number of posts submitted to the subreddit and the number of votes these posts received. Measures of the proportion up/down votes cast on the subreddit, and the proportion of 'text' posts, were also added.

Several further data tables and variables were created later in the research process to address specific questions - these will be detailed when the analyses which called for their creation are discussed.

Problems with the back-end data

While this procedural data contains a complete and accurate record of voting activity for one month, there are important questions for which it has little utility. Firstly, user and post IDs were anonymised before the data were provided - this data can tell us nothing about what any of the posts being voted on actually represented, and so it is of little use when we consider questions about the nature of discourse on Social News websites. Secondly, no information on comments or comment voting was provided - this is a serious problem given that the 'democratically mediated discussions' which take place in the comments threads for posts are important to the functioning of the website, and also interesting in their own right. Thirdly, the procedural data holds no

information on which page(s) a post appeared on throughout its active lifespan on `reddit` - these pages are dynamically generated in real time using the procedural data. Once a moment has passed there is no way to wind the clock back and know which posts appeared on the Front (or any other) page at that moment. To gain access to this information the page(s) of interest must be observed and recorded in real time. Finally, after consenting to our first request for data and several supplementary data tables, Reddit's administrators unfortunately did not respond to subsequent requests. If the present research was to consider anything more than a one-month snapshot of anonymised `reddit` data it became clear that this back-end data would have to be supplemented with data from another source.

4.1.2 Collecting front-end data through APIs

To circumvent these problems with the back-end data it was decided that data should also be collected directly through the website's front-end. The first approach to doing so involved scraping the `HTML` files which `reddit` serves when the website is visited. This is akin to visiting the website in a web browser, navigating to the page of interest, and saving the `HTML` source code. Extracting data from these `HTML` files was very difficult and time-consuming - the values of interest (e.g. a post's title, url, rank, number of up and down-votes) were scattered among thousands of lines of `HTML` and CSS formatting. The Python library 'BeautifulSoup' (Richardson, 2007) was utilised to automate this process, and after some time acquiring familiarity with the package it was possible (although still difficult) to extract data from these `HTML` files.

At this point `reddit`'s Application Programming Interface (API) was discovered by the researcher. These APIs are provided by many Social Media applications, they allow independent programmers to write software (e.g. mobile phone applications) which interfaces with the website's database. Any given page on `reddit` can be requested through this API rather than through a browser; when requested through the API the file is served in Javascript Object Notation (`JSON`) rather than `HTML`. These `JSON` files are much neater and easier to navigate, as they do not contain the mass of formatting tags associated with `HTML`. `JSON` has a further advantage in that the statistical environment `R` has a library which can read `JSON` files directly into data frames (`rjson`). This has allowed front-end data to be collected in a much more dynamic and complex fashion than would have been possible with `HTML` files.

The procedure for collecting front-end data involved the use of a dedicated web server, several `R` scripts, the '`rjson`' library (to read in data from `reddit.com`) and the '`RMySQL`' library (to store the data on the web server's `MySQL` database). This method of data collection was first employed in April 2010, but it was not stable until July 2010. From July 2010 until the observation period was concluded in March 2013, front-end data has been collected more or less continuously (with

the exceptions being a number of web server crashes). There follows details of the scripts used to collect this data.

The first of these scripts monitors `reddit`'s Front page and the Main pages of various subreddits at 30-minute intervals; creating a record of all of the posts appearing there, their title, url being linked to, number of votes and comments. A second script monitors the 'New' page of selected subreddits at 30-minute intervals; adding all of the new submissions to a watch list. A third script cycles through this watch list every 30 minutes, recording the voting performance of the post and, importantly, the text of all of the post's comments and the up and down vote tallies for each of these. This script also makes a record of whether a post has seen any new votes or comments since the last time it was observed. A fourth script monitors the watch list and removes posts which have seen no voting or commenting activity for four consecutive observations.

The effect of combining these scripts is that records are created of various pages of interest, and a selection of posts to subreddits of interest are tracked throughout their active lifespan on `reddit`. These records can then be combined to add information on the pages a given post appeared on at each time of observation. Indeed the first step towards analysing front-end data is to create a new table that contains all of this information for tracked posts, along with the relative increase in votes and comments as compared to the previous observation of the post. This allows for questions about the amount of attention and voting activity a post can expect, by virtue of appearing on a certain page, to be addressed.

The number of subreddits that can be monitored simultaneously in this fashion is limited by two factors. Firstly, certain aspects of these scripts are computationally intensive - in particular recording the text and performance of a post's comments. A popular post can have hundreds of comments, and the script must cycle through all of these creating or updating database records. As such, if the script is attempting to track too many posts the first iteration will still be running when the second is due to begin, either overloading the computer's resources or pushing subsequent iterations back such that the script is increasingly behind schedule. A second limiting factor is related to `reddit`'s servers - to avoid overloading these servers `reddit`'s administrators ask that no more than 1 request is made every 2 seconds; exceeding this limit leads to future requests being blocked.

Trial and error revealed that, staying within these two limiting factors, only two or three 'busy' subreddits could be monitored at the same time[1]. The choice of which subreddits to monitor is a difficult one, and in practice this changed from month to month as data from particular subreddits was required to address certain research questions. One of the limiting factors of this approach is that data must be collected in real time - if some event unfolds through a subreddit which was not

---

[1]By late 2012 the activity levels on 'busy' subreddits had increased to the point where it was no longer possible to comprehensively monitor more than a single subreddit

being tracked it is not possible to retrospectively collect the data which would tell us where posts appeared at a specific time or what their top comments were at this time.

This front-end data has several strengths and weaknesses as compared to the back-end data described above. Most importantly, it provides full access to the 'text' of the posts and comments passing through the voting system. It also allows one to produce, with some degree of accuracy, measures representing the page(s) a post appeared on at a given moment in time. The major disadvantage of front-end data is that it contains no information about which users voted on which posts or comments - as voting on `reddit` is an anonymous activity. A second disadvantage relates to a lack of temporal resolution. Where back-end data is accurate to the second front-end data is only accurate at each observation point, and these are generally spaced 30 minutes apart. Consequently, throughout this research there will be questions and analyses which are more suited to either back-end or front-end data - both will be utilised in an effort to develop a fuller understanding of how Social News operates. It is unfortunate that these data sources relate to different periods in `reddit`'s history, the website saw considerable growth between the month (March 2009) when back-end data were recorded, and the beginning of front-end data collection (July 2010).

Several further attempts to harvest front-end data related to Social News have been implemented with limited success. `Digg` also provides an API, and a Python wrapper (PyDigg) exists to facilitate making requests through this API. In May 2010 Python scripts were created to harvest data from `Digg` in a similar manner to the procedure implemented on `reddit` - with the intention being that the behaviour of `reddit` and `Digg` could be compared using this data. Differences in the operation of `Digg`'s voting system and API made it clear that a direct comparison of these websites would be difficult. Unfortunately, after just one month of stable data collection from `Digg` the website's administrators upgraded to 'version 4' of their software; a fundamental re-write which broke the PyDigg package and all scripts written to collect data using this package.

`Imgur.com` is an image hosting website created by a `reddit` user (redd.it/7zlyd) in response to an effect (also known as the `Slashdot` or `Digg` effect) whereby popular `reddit` posts that linked to an image on a web server with limited resources would bring so much traffic to this website that the server would be overloaded and crash. Many `reddit` users now upload an image to `Imgur` and submit the `Imgur` url to `reddit`, rather than linking directly to the image where it was found. `Imgur` provides a hit counter for each of these images; and this offered an opportunity to monitor the attention received by images posted to `reddit` as they progressed through `reddit`'s voting system. A script was written which would load the `Imgur` page associated with posts on the `reddit` watch list, and scrape the `HTML` of this page using BeautifulSoup (Richardson, 2007) to extract and store the number of hits the image had received. This allowed for the attention received by images posted on `reddit` to be quantified and analysed in relation to the post's voting rate and the `reddit` page(s) it appeared on. Unfortunately `Imgur` also upgraded their software

within about one month of this data collection regime being implemented, in the process breaking the `Imgur HTML` scraping part of the script.

These examples illustrate one of the pitfalls of collecting front-end data directly from a website. The process of creating these scripts and ensuring that they function correctly and reliably is very time-consuming. When the script can function for a period of months or years and collect a large volume of useful data this investment of time is worthwhile. However, a data collection regime like this is quite delicate and can easily be broken by changes to the structure of the website being monitored. When this happens shortly after a data collection regime is deployed it results in a disappointing data return for the time that was invested.

There are further problems with this method of data collection. Reddit's servers are frequently overloaded, and during most weeks there is at least one period where the website is unavailable for a number of hours[2]. It also seems that during times of heavy load API requests receive a lower priority than browser requests; in any case there are numerous gaps in collected front-end data. When transforming and analysing this data such gaps and other quirks must be sought out and worked around to produce a reliable data-set that can be analysed with confidence.

There are further 'data quality' issues relating to `reddit` which did not become apparant until later in the research process when data augmentation and analysis began. Information on the score of a post is not reliable between pages - when a post appeared on several pages for which data was being recorded it was common for each page to report differing numbers of up and down-votes, even though the observations of these pages were recorded within a few seconds of each other. Studying the recorded scores for posts over time revealed another problem - these values will occasionally 'reset' to an earlier value. For example, one might observe a post with 100 up-votes and on the next observation that post now reports a value of 50 up-votes - this lower figure persists for several observations and then the figure jumps to 150 up-votes. This causes problems when one wishes to model the rate at which posts accrued votes between observation points, introducing a large negative value at one observation point and then an artificially high positive value whenever the problem subsides. This phenomenon is possibly the result of `reddit` serving cached score information during times of heavy load, the problem is most pronounced for information recorded from the comments pages of posts - and keeping this information updated accurately is likely a low priority. This problem has been addressed firstly by preferring measures of post activity which are recorded from the New, Rising and Main pages (the problem occurs less frequently with this data). Secondly, steps were taken to assure data quality - where the records for a post indicate that the problem has occurred the data for this post is excluded from analyses.

It should be noted that at the outset of this research the researcher had no coding skills or knowledge

---

[2]This was true in 2010 but more recently `reddit` has invested in more staff and greater server capacity, with the result that the website very rarely goes down unexpectedly now.

of how websites function - instead operating on the assumption that data which would allow for an understanding of `reddit` existed on their servers and was in principle retrievable. As such, the data collection and augmentation tasks have probably not been performed in the most efficient manner possible. The situation was further complicated by a total lack of any official documentation for `reddit`'s API at the time - this was instead learned of by observing discussions between users on the site's comments pages. Similarly, there was at the time little published research on how social scientists could use techniques like data collection through scraping or APIs - aside from the fleeting mention that this had taken place in articles that used such data - e.g. (Halavais, 2009).

These three factors have exerted a strong influence on the manner in which data collection was approached - most significantly the lack of information on which types of data could be extracted from `reddit`, and a lack of confidence in the researcher's ability to create a data collection regime which could reliably collect such information. In hindsight, there is one aspect of `reddit` data which could have been collected but was not - that pertaining to individual users. There are a number of reasons why the collection of this data would have been more difficult and consumed more computational resources than the approaches which were adopted - but there are also additional research questions which could have been addressed through analysis of this data. This could also potentially have formed a link between the analysis of users (which is based almost entirely on the single month of back-end data from 2009) and other forms of analysis which utilised much more recent data.

There is also now an emerging literature on this methodology (Marres and Weltevrede, 2013) - which would have been a great benefit both in terms of learning the practical side of online data collection and ensuring that data collection proceeded in a manner which was theoretically sound. Throughout the thesis each source of data and the manner in which it was collected is described in some detail - it will be left alrgely to the reader to determine whether data collection was executed appropriately, and the issue will be re-visited in the concluding section of the thesis.

The way in which `reddit`'s API has been used to collect data differs to the way in which APIs are sometimes used to collect this kind of data for the purposes of social science research - e.g. Halavais (2009); Lerman (2006). This is partially a result of `reddit`'s API being slightly unusual, and partially related to the type of analysis which the data is being used for here. All of the data collected through `reddit`'s API relates to pages of the website directly, the API calls made by the present research all request the contents of a specific page.

In contrast, `Digg`'s API is in many ways more typical and allows one to request information along specific criteria, and has been used in this manner by researchers Halavais (2009); Lerman (2006). For example, **?** used `Digg`'s API to collect the complete posting and commenting history of a random sample of 30,000 users. This kind of approach is not possible on `reddit` because the API cannot be used to fetch a complete list of users - it is possible to request an individual's user profile

page (if one already knows their user-name) but this only contains their most recent post and comment contributions, and would need to be iterated upon to fetch older records (thus increasing computational load and the amount of server time required substantially).

The presence of an API which is distinct from the website itself raises the possibility that the structure of `Digg`'s API subtly shaped the approach and findings of researchers who used it, and this issue is relevant across a broad range of APIs (Marres and Weltevrede, 2013).

This potential problem is avoided by the present research because the website is itself the subject of the research - the research attempts to build an understanding of how `reddit` operates largely by considering the function and changing content of specific pages, and the API merely provides easier access to the contents of these pages (albeit with some data quality issues which have been discussed). Where empirical data collected through `reddit`'s API is discussed it is always in the context of the *pages* where these observations were conducted.

Of course `reddit`'s relatively limited API and its impoverished documentation[3] (along with the cessation of co-operation from its Administrators) have shaped the research at a more fundamental level - restricting the available approaches to understanding 'how `reddit` works' (Chapter 5) and how users behave (Chapter 6). While the available approaches have been constrained, there should not be any issues with the validity of their findings which stem from the nature of `reddit`'s API. The strengths and weaknesses of the present research's approach to data collection are re-visited in section 9.3.

4.2 Surveys and experiments conducted with the users of Social News

There are questions related to Social News and Distributed Moderation which cannot be adequately addressed through observation of the system in question. The present research makes use of both surveys and experimentation with the users of Social News as the subject population. The specifics of the surveys and experiments which are utilised by the present research will be described in the relevant chapters (6 and 7 respectively). There follows here a short discussion about a means of recruiting participants which Social News websites offer.

Anything with a URL can be submitted to a Social News website, this includes surveys and experiments which are conducted online. If the object of these surveys/experiments is to study the population of Social News users, then recruiting respondents/participants through the Social News website itself seems like a sensible way to proceed. An obvious advantage of doing so is that submitting to a Social News website costs nothing, and if the submission fares well through the

---

[3]It may be possible to use `Reddit`'s API in ways which I am not aware of

voting system thousands of participants can be recruited quickly. A further advantage is that the websites provide direct access to their user populations (a respondent recruited through a Social News website is almost by definition a Social News user or at least visitor). Finally, by monitoring the placement of the survey/experiment on the Social News website it is possible to identify sub-groups of participants based on the location of the post at the time they begin participation. For example, if one is interested in looking for a difference between users who browse the New pages and users who only peruse the Front page, these sub-populations can be identified based on the location of the post (e.g. users who begin participation within 10 minutes of the post's submission can only have seen it on the New page).

The major disadvantage of recruiting participants in this manner is that one is relying on some degree of success through the voting system in order to reach a substantial number of prospective participants, or any participants who do not browse the New pages. Reddit users have, in the website's history, up-voted several surveys to the Front page - and these surveys have gone on to recruit thousands of participants. However, as the site's popularity has grown users seem to have become more wary of attempts by 'outsiders' to study them. A further problem with recruiting participants in this way relates to the method of sampling - this is certainly not random, and is probably best described as an unusual form of convenience sampling. If the post containing a link to a survey or experiment reaches the website's Front page then it will be seen by a large number of users (who subscribe to the subreddit where the post was submitted) while it appears there. Participants would therefore be a self-selecting sample drawn from the population of users who visit `reddit` during the time when the post appears on the Front page.

Chapter 5

How does Distributed Moderation work?

This chapter aims to investigate the workings of Distributed Moderation websites - concentrating in the first instance on the Social News website `reddit.com` and subsequently considering the `Stack Exchange` network. The Chapter concludes with analysis of the `Your Freedom` website, an attempt to apply Distributed Moderation to political policy ideas. All of these websites were introduced and described in Chapter 3.

5.1 Social News - `reddit.com`

For much of the time that `reddit` has been under observation for this research the website's information pages contained only basic notes on how it operated - simply saying that up and down-vote arrows were related to an item's score which in turn determined where it would be displayed. Beyond this, users would often discuss the workings of the voting system in the comments pages for posts and so a user might learn about the website from other users.

One example of this occurred on the post redd.it/eaqnf [1]. In this post one user is querying why a story about North and South Korea exchanging artillery fire, which was submitted to the worldnews subreddit, has almost 5,000 down-votes - their reasoning is that this is a major world event and perfectly matched with the purpose of the /r/worldnews subreddit. In this instance one of `reddit`'s main employees of the time ('Jedberg') provided an answer which is of some importance here - namely that `reddit` 'fuzzes' up and down-vote numbers in order to hinder the development of voting-bots.

---

[1] this is a short-form url which links directly to a post's comments page on `reddit`, it will be used whenever reference is made to specific `reddit` posts

Placement on `reddit`'s Front page brings with it a significant volume of attention and attention is a valuable resource on the web. It is therefore unsurprising that there have been many attempts to cheat `reddit`'s voting system with 'vote-bots' - computer programs which mimic the voting behaviour of `reddit` users. The rationale behind the fuzzing of vote numbers is that when a voting bot casts an up or down-vote `reddit` has a secret algorithm which attempts to identify and nullify such votes - if the owner of the vote-bot observed that their votes were no longer increasing the total number of up/down votes then they could surmise that their bot had been identified, figure out how it had been identified, and improve the bot accordingly. Therefore `reddit` 'fuzzes' up and down-vote totals to hinder the development of voting bots. The score of a post (up-votes less down-votes) is however an accurate reflection of the votes which have been cast.

As `reddit` has grown, both in terms of users and more recently staff, the pages which provide information about the website (what it's for and how it works) have been expanded considerably. For example, the information about vote fuzzing is now displayed on the website's FAQ page (`http://www.reddit/help/faq`) - whereas earlier in `reddit`'s history this kind of information could only be obtained through discussions with other users. Over the full term of this research much more information about `reddit` has become available, and the website itself has undergone some major changes (most notably a huge expansion in its number of users and visitors).

Figure 5.1: Excerpt from the `reddit` FAQ and Reddiquette pages concerning the voting system - 1st November 2012

Figure 5.1 contains excerpts from the `reddit` FAQ and 'Reddiquette' pages which concern how the voting system works and how it should be used. The criteria for up-voting ('good' content) and down-voting ('junk' content) are vague, as is the description of how these votes are used ('Links which receive community approval bubble up towards #1'). This chapter's primary aim is to shed light on how the voting system works in practice, and to this end a number of resources are available. The software `reddit` uses is open source, and therefore the algorithms which operate on post and comment votes are known (or at least knowable) - sections 5.1.2 and 5.1.7 respectively

consider the nature of these algorithms.

`Reddit` is also a fundamentally public space - only a few areas of the website have controls on who can access them (e.g. a 'Reddit Gold' subreddit for users who have paid for the premium 'Reddit Gold' service). Through `reddit`'s API it is possible to request and record information from any of `reddit`'s public pages, it is data of this nature which the present chapter relies upon primarily. Additionally, Reddit's administrators provided voting data for one month in 2009 for research purposes and the chapter begins with some analysis of this data.

In summary, the specifics of how `reddit`'s voting system functions in practice are largely unknown but the resources through which these might be understood are readily avalable. This chapter's purpose is to determine what can be understood from the available data. There are a number of research questions which serve as a roadmap to this exploration, these are as follows:

- `reddit`'s voting system appears to have a built-in 'rich get richer' mechanism - positive votes for an item or comment lead to placement in a higher-visibility location - where the item will presumably receive further votes at an increased rate. It is therefore expected that the distribution of voting activity between items of content will be highly skewed, such that a small number of items (in particular those which reach the Front page) will receive a disproportionate share of the users' votes and attention. The first task is to investigate whether this is the case and to describe the nature of this distribution. Section 5.1.1.

- Social News websites utilise a variety of page types to display content. While it can be assumed with some confidence that the Front page will play host to the highest rate of voting activity; the questions of how much activity centres here, and how the remaining activity is distributed between other page types, are of considerable interest. Section 5.1.3.

- Much of the potential broader significance of Social News websites rests on the way in which they channel and focus the attention of their users. While voting activity can, to some degree, be used as a proxy for user attention - it is also important to try and quantify the amount of attention posts receive in various locations. Knowing how `reddit` channels attention, and how much attention is being channelled in this manner, is key to understanidng `reddit`'s place and significance in the 'hybrid media system' (Chadwick, 2013). Section 5.1.4.

- Given the focus these websites place on their Front page, it is of paramount importance to identify and describe the process through which a post comes to appear on this Front page. Section 5.1.5.

- `reddit` also applies up/down voting to comments on posts - and the comments pages for active posts showcase an unusual form of 'democratically mediated discussion', where thousands of individuals contribute with comments or votes and, through the voting system, the most popular responses are determined and positioned in the most prominent locations. It is

of considerable interest to understand how the ranks of comments for a post are determined in practice. Section 5.1.7.

- `reddit` can be thought of as having two 'components' - the software through which content is submitted, votes are cast and ranks are calculated - and the people who perform these actions through the software's interface. For `reddit` to 'work' the software must be designed and configured appropriately, *and people must use it appropriately.* If `reddit`'s users suddenly decided en masse to cast their votes along different criteria the nature of the website would change immediately - the norms around voting contribute to defining what `reddit` is, as do the algorithms which operate on these votes. Reddit's software is to a large degree standardised across the whole website, but there are certain subreddits which *rely on people adopting subreddit specific voting norms* if they are to function. 'Ask Me Anything' iterviews (on the /r/IAmA subreddit) require users to vote on comments along criteria which are distinct from the other subreddits. It is possible to test whether users' collective behaviour on this subreddit meets the specific requirements for an 'Ask Me Anything' interview to function. Section 5.1.12.

Two data types from `reddit` are of utility here - back-end procedural data supplied from `reddit`'s servers by an administrator, and front-end observational data collected through `reddit`'s API. Back-end data will be used to address questions related to the distribution of votes as it represents a complete and accurate record of one month's voting activity (there is no 'fuzzing' of this data). To address questions about the breakdown of user voting activity and attention between `reddit`'s pages front-end data will be utilised - as back-end data contains no information on which pages a post appeared on.

5.1.1 A highly skewed distribution of votes between posts

The month of March 2009 saw 352,902 post submissions and 3,446,522 votes cast. Initial inspection of back end data revealed a highly skewed distribution of voting activity between items of content (see figure 5.2). 167,688 posts (47.5%) only received one vote (attributed automatically upon submission) whereas the post with the most votes received 5,997 votes. Closer inspection of the back-end data revealed that 80% of the votes cast in the month accrued to 7.8% of the submitted posts.

Inspection of figure 5.2 gives an initial impression that the distribution of votes between posts may follow a power law. The power law was introduced in section 2.5.1.

The guidelines set out by Clauset et al. (2009) were adopted to check whether the power-law offered a plausible fit for this data. The first step in doing so was to generate plots of the data on logarithmic axes (see figure 5.3). The top-left pane of this graph shows raw data over the full
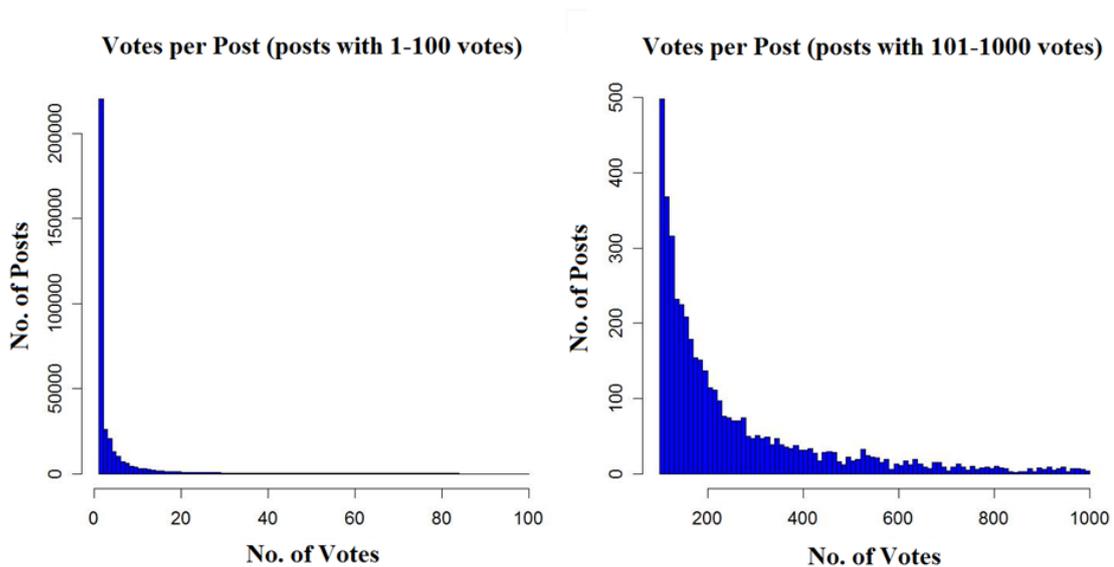
Figure 5.2: The raw distribution of votes between posts for two sub-sets of posts

range, the top-right pane shows the same data represented on logarithmic axes, while the bottom-left pane shows the inverse cumulative distribution on logarithmic axes. If the data follow a power law Clauset et al. (2009) report that plotting it on logarithmic axes will produce a straight line. Such a straight line is apparent for a certain range of voting activity levels. The line representing the frequency of posts with between 1 and 1,500 votes has a stable slope, but above 1,500 votes the slope of the line changes considerably. The top-right pane shows this most clearly - there are more posts with very large vote totals (> 1500), larger than would be expected based on the rest of the distribution.

Separate plots similar to 5.3 were produced for the 40 most active subreddits, with the majority of these exhibiting the same type of 'two-stage' distribution. The method of Clauset et al. (2009) was then employed to fit the power law to these data and estimate its parameters and goodness-of-fit. The decision was taken to fit the power law to subreddits separately because in many respects they behave as separate entities. At the time when the procedural data were collected there were twelve default subreddits and 25 posts on `reddit`'s Front page. A visitor who was not signed into a user account would therefore see the 25 highest-ranking posts from the default subreddits on their `reddit` Front page. All of the page types employed by `reddit` (e.g. New, Rising, Hot) exist both for individual subreddits and in aggregated form (i.e. compiled from all of the subreddits the user subscribes to) - the Front page is the aggregated form of the 'Hot' page type.

In order for a `reddit` user to vote they must be signed into a user account, and users who are signed into accounts can easily manage their subreddit subscriptions. For this reason it is prudent to treat subreddits as seperate entities (they have independant sets of and differing numbers of subscribers). The power law was therefore fitted separately to voting distributions from individual
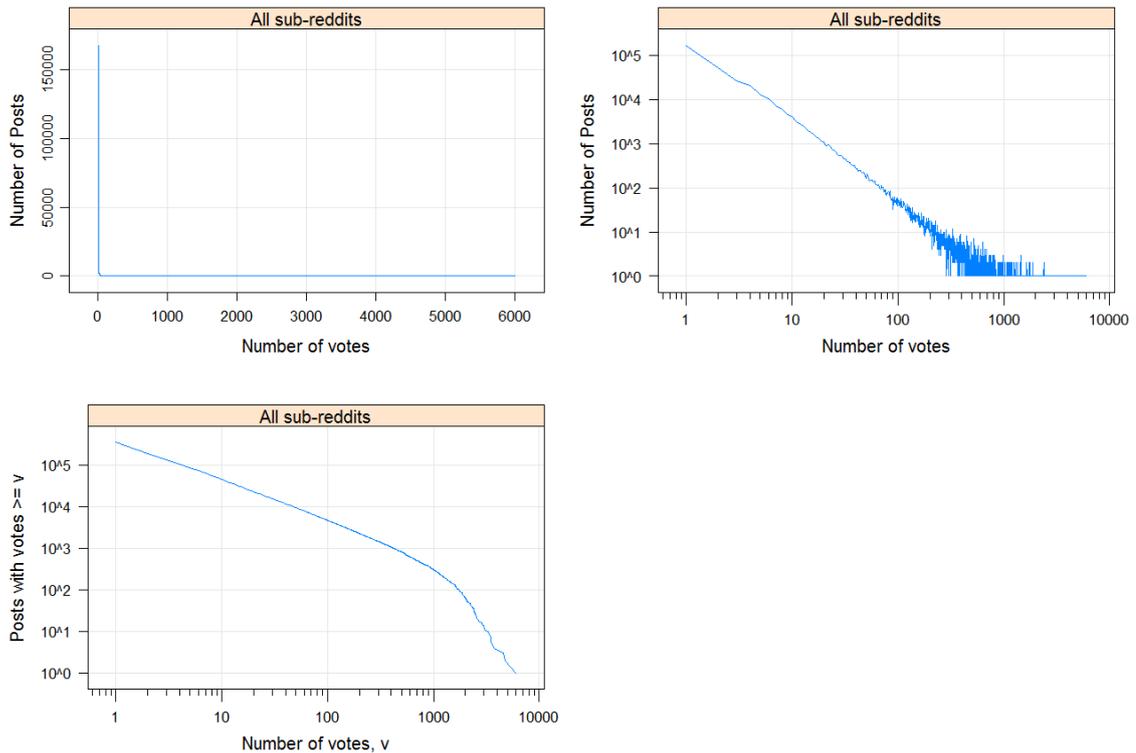
Figure 5.3: Plots of the frequency distribution of votes across posts. The top-left pane shows the data on untransformed axes, the top-right pane shows the data on logarithmic axes, the bottom-left pane shows the inverse cumulative distribution on logarithmic axes.

subreddits, as well as to the full data-set incorporating all subreddits.

The power law's goodness-of-fit for this data was assessed using the method of Clauset et al. (2009), parameters were estimated for individual subreddits and the power law's fit was compared to other distributions (poisson and exponential). These analyses are not reported because, on reflection, it was decided that estimating the parameters of the power law did little to inform our understanding of how it affected `reddit`. The highly skewed distribution of votes between posts is undoubtedly key to understanding how `reddit` 'works', but whether the power law (or some other distribution) is a plausible fit for this data is largely irrelevant for our purpose. The important finding from this work was that the plausibility of a power law distribution varied between subreddits.

This highlights the importance of `reddit`'s layered and multi-threaded structure. Despite the website's over-arching visual theme the subreddits it is comprised of are far from homogeneous. Rather, each subreddit is better thought of as its own distinct entity or process. There are thousands of these separate processes operating in tandem, with several hundred [2] hosting a considerable level of user activity. Each of these processes outputs a constantly updated ranked list of the top posts that have been submitted there. Through the subreddit subscription system individual users can

---

[2]This figure was given in 2009, it would be much larger in 2013

choose which of these 'output streams' they find useful or interesting, allowing these to feed into their Front page. When an individual subreddit suffers a degradation in the quality of content being up-voted to the top, or some other problem, this does not necessarily impact on the rest of the website - users who don't like the direction the subreddit is going in can unsubscribe, or the administrators can choose to remove it from the list of default subreddits.

This kind of compartmentalisation may have a further benefit in limiting the volume of submitting and voting activity which goes into the collective decision-making that powers the website. It has not yet been established whether there is an upper limit to the number of individuals who can participate in making one of these collective decisions and in so doing benefit the process. It may however be the case that the '/r/reddit.com' subreddit - a legacy from a time when subreddits did not exist, and by far the most active subreddit - actually surpasses this limit and has too many participants. 150,042 posts were submitted to this subreddit in March 2009, 104,503 of these did not receive a single vote. It may be the case that many of these posts were seen but not rated due to some particular quality which discouraged users from voting on them (e.g. being completely unremarkable). However, it seems more likely that many of these posts slipped through `reddit`'s voting system without a vote because they were not seen by any users. This would represent a scaling problem in using `reddit`'s Distributed Moderation system in situations where there are a large number of user actions - if a substantial percentage of submissions are being ignored at random it seems unlikely that the system will reliably allow the *best* submissions to rise to the most prominent locations.

This is not intended as a criticism of `reddit`'s Distributed Moderation system. It is in fact remarkable that user voting behaviour can impose any order whatsoever on this deluge of submissions. If we consider another communications medium, one can barely imagine the results of a single discussion board seeing 150,042 new threads created in a given month - complete failure of the board to facilitate any kind of discussion seems like the most likely outcome. We might also consider a newspaper which receives a large number of 'letters to the editor', or a politician who receives a large number of letters from their constituents. How is an abundance of correspondence dealt with in these instances? Might some form of Distributed Moderation improve the situation?

`Reddit`'s voting system seems to perform well at selecting 'good' submissions from the '/r/reddit.com' subreddit to place on the Front page - but with so many posts not receiving any votes it seems unlikely that these are 'the best' posts which were submitted. For `reddit`, this represents a minor flaw or an area that could be improved upon, but if we consider the application of Distributed Moderation in the arena of national politics this minor flaw takes on a new aspect. Where the posts are political policy ideas, and some action occurs as a consequence of these ideas reaching a high rank, knowing whether these ideas are 'the best' or a random sample of those which are 'good', becomes much more important. Furthermore, if such an application in the domain of na-

tional politics were to achieve widespread popularity it would likely be dealing with this volume of activity or greater. This is therefore an important consideration in assessing the broader utility and/or potential of Distributed Moderation systems, and one which will be re-visited throughout the present research.

Chapter 8 will consider the effect of a subreddit being added to the list of default subreddits; where the level of attention and voting activity accruing to submitted items shows an instant and substantial increase.

5.1.2 Reddit's 'hot' post ranking algorithm

As `reddit`'s software is open source it is possible to scrutinise the algorithms used to rank posts directly and a number of blog posts have done so (Dover, 2008; Salihefendic, 2010). The algorithm used to create the 'hot' ranking (used to determine which posts appear on the Front page) is displayed in mathematical notation in figure 5.4. The algorithm produces a 'hot' ranking score, and the Front page will display the 25 posts which have the highest ranking score at a given moment in time. If a user is signed into an account their Front page will be filled with the 25 posts with the highest ranking score from the subreddits they subscribe to. For users not signed into an account these posts are drawn from the subreddits which are currently defaults.

Given the time the entry was posted A and the time of 7:46:43 a.m. December 8, 2005 B, we have t_s as their difference in seconds

$$t_s = A - B$$

and x as the difference between the number of up votes U and the number of down votes D

$$x = U - D$$

where $y \in \{-1, 0, 1\}$

$$y = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}$$

and z as the maximal value, of the absolute value of x and 1

$$z = \begin{cases} |x| & \text{if } |x| \geq 1 \\ 1 & \text{if } |x| < 1 \end{cases}$$

we have the rating as a function $f(t_s, y, z)$

$$f(t_s, y, z) = \log_{10} z + \frac{y t_s}{45000}$$

Figure 5.4: Reddit's 'hot' ranking algorithm - from Dover (2008)

There are several aspects of this algorithm which are important in understanding how the content of

reddit's Front page is determined. Firstly, this algorithm makes use of post score and submission time but no other criteria. The fact that posts have up-votes and down-votes is not made use of by this algorithm, down-votes are first subtracted from up-votes to produce a voting score, and it is this voting score which the algorithm makes use of to generate the ranking score for each post. This means that, for example, a post with 100 up-votes and 90 down-votes will have the same ranking score as a post with 10 up-votes and no down-votes - if they were submitted at the same time.

These voting scores undergo a $\log^{10}$ transformation within the algorithm. This transformation serves to reduce the impact of very large voting scores, and accentuates the importance of early voting. A voting score of 100 bears only twice as much weight as a voting score of 10 and a voting score of 1000 bears only three times the weight of a voting score of 10.

The algorithm uses a post's time of submission in Unix epoch time format (number of seconds since 00:00:00 on 1st January 1970) and subtracts 1134028003 (the Unix epoch time for 07:46:43 on December 8th 2005). As a result the time variable for a post (ts) is actually the number of seconds between 07:46:43 on December 8th 2005 (thought to be the time reddit launched) and the post's submission. This means that the ranking scores for posts on reddit are themselves stable if they stop receiving votes (because their submission time does not change). Consequently, reddit's Front page tends to host content submitted within the last 24 hours because the ranking scores for new posts get higher every day (i.e. a post submitted today with a voting score of 100 has, and always will have, a higher ranking score than a post submitted yesterday which also has a voting score of 100).

The way this submission time variable is used, coupled with the logarithmic transformation of voting score, ensures that reddit's Front page will only display posts submitted within the last 24 hours at most. The logarithmic transformation of voting score ensures that no matter how well a post scores its ranking score will inevitably be overtaken by newer posts. As time passes the weighting of new votes for a popular post decreases while the advantage fresher posts have because of their later submission time increases.

Reddit's 'hot' post ranking algorithm is remarkably simple when we consider that it is fundamental to how reddit.com, the only product of a company now valued at hundreds of millions of dollars, works. The simplicity of this algorithm and the fact that it is publicly displayed is important when we come to consider the factors behind reddit's success.

5.1.3 Voting activity by page type

The analyses conducted on back-end procedural data suggest a focusing of user attention and voting activity on reddit's Front page - but it is not possible to verify this with the available back-end

data. This section utilises data harvested through `reddit`'s API to investigate the distribution of `reddit` voting activity between some of the pages of which the website is comprised.

`Reddit` provides four tabs which select posts to display according to different criteria. All of these tabs (and sub-tabs) operate in both aggregated form, and for individual subreddits. In aggregated form these tabs display content from all of the subreddits a user has subscribed to, users who are not signed into an account are shown content from the default subreddits.

The first of these tabs is titled 'hot' or 'what's hot', and it selects popular posts based on their score, with a time penalty for older votes. `Reddit`'s Front page is merely the aggregated form of the 'hot' tab. The 'hot' tab for a particular subreddit is referred to here as the subreddit's 'Main page' (because it is displayed upon navigating to the subreddit). `Reddit` also offers a 'New' tab which has two sub-pages, the 'New' page shows the latest submissions ordered by their submission time, the 'Rising' page shows new submissions which have attracted early positive votes. The remaining two tabs offer a drop-down menu which allows the user to specify which time interval they wish to look at (e.g. posts from today, this week, this month, this year). The 'controversial' tab displays posts which have attracted a similar number of positive and negative votes, while the 'top' tab shows the posts which have the highest score for the selected time period.

In order to estimate activity levels on the various pages one needs records of a posts' number of up-votes, down-votes and comments at regular intervals throughout its active lifespan and also to know which page(s) the post appeared on at each observation point. With this data it is then possible to cycle through the observations for an individual post in temporal order and at each observation point calculate how many *new* up-votes, down-votes or comments the post had received as compared to the previous observation point. The resulting data set is then comprised of values indicating how much a post's activity metrics had increased at each observation point, paired with information on which page(s) the post appeared on at each observation point. For example, the first time a hypothetical post is observed it is on the New page and it has 3 up-votes, 0 down-votes and 1 comment. The second time the post is observed it appears on both the New and Rising pages and it has 5 up-votes, 1 down-vote and 2 comments - the augmented data table will show that it had 2 new up-votes, 1 new down-vote and 1 new comment and that it appeared on the New and Rising pages.

The 'fuzzing' of up and down-vote totals means that the values which these analyses are based on are not entirely accurate but instead are artificially inflated. This puts some limits on the depth of analysis which the data can sustain and must be taken into account when interpreting results. For the present purposes, where the objective is to understand relative activity levels on `reddit`'s pages, the fuzzing of vote figures does not result in any insurmountable problems. The 30-minute gap between observations is another source of noise in this data as it relates to pages - when a post is observed on a page for the first time there is no way to know how long it had been there (other

than this value lying between 0 and 30 minutes). Similarly, a post may have appeared on a page briefly between observation points and this would not be reflected in the data. These problems would be easily circumvented with more privileged access to the data on `reddit`'s servers - for the present research they remain as caveats to the weight which can be placed on specific parameter estimates.

This section considers data collected between August 20th 2012 and September 27th 2012. Previously, this type of data set was constructed by 'tracking' posts individually. When a post was observed on the New page for the first time it would be entered on a 'watch list', then every thirty minutes a second script would retrieve data from the comments pages of all the posts currently on the 'watch list', with posts being removed from the watch list once they had shown no signs of activity for a number of sequential observations. Unfortunately it transpired that the comments pages for individual posts were not a reliable source of information on their number of votes and comments. A peculiar trend was observed in the data whereby numbers of votes and comments for a post would sometimes freeze or even 'reset' to an earlier value. If one was looking at the number of new up-votes per observation one might find that at a certain observation point this value was -100, suggesting that the post lost 100 up-votes - the reported number of up-votes would remain at this lower value for a spell and would then jump back up to a higher value. It is suspected that this is a result of `reddit` caching data during times of heavy load - when the servers are under stress they prioritise updating certain types of page, for pages that are of lower priority (like the comments pages for individual posts) this seems to result in the serving of cached information until the server finds the time and resources to update the values. This proved highly problematic as it is a source of considerable noise in the data.

To circumvent this problem I instead rely on observations of post metrics which are recorded on the `reddit` Front page and the Main, New and Rising pages of a selection of subreddits. The data set was constructed as follows: first a list of unique post IDs was taken from observations of the New page, then for each one of these posts every record was retrieved and sorted by observation time. Observations which occurred at the same time (within one minute) on different pages were grouped together to produce one row of the new data set - in this new data set values such as the number of new up-votes since the previous observation, the time which had elapsed since the previous observation, and which page(s) the post was observed on within that one-minute group were generated and stored. This allows for the fitting of models which use the page(s) a post appeared on as explanatory variables to model the frequency of an action over the preceding gap (usually 30 minutes).

This approach means relying on the Front, Main, New and Rising pages for all of the information on a post's progress - therefore the script which collects this data was set to retrieve 100 records for each page. By default `reddit` shows 25 posts per page and to see further posts a user must click

on a 'Next' button - here the script has effectively collected data from the first four 'leafs' of each page. However, in considering the page(s) a post appeared on only the first 'leaf' is considered initially - for a post to be counted as appearing on a page it must appear there with a rank in the top 25.

Modelling the data - up-votes

This section describes the process of finding a suitable model for voting activity. In this section we will consider only the 'Worldnews' subreddit and model it's number of new up-votes per observation. Each row of this data-set relates to an observation, a given post will generally appear in more than one row because it will have been observed multiple times. For the observation period there are 247,509 observations in total and these concern 9,598 different posts.

Due to the above-noted problem of frequency measures sometimes 'resetting' to earlier values the data-set needed to be cleaned. Every row of data on which the number of new up-votes, down-votes or comments had a negative value was identified as a problem. Where a post had a false value of -100 on one observation it would on a subsequent observation (after the problem had subsided) have a value which was 100 greater than it should be. Therefore for each instance of a problematic row all subsequent observations of the post it related to were also marked as problematic. This cautious approach resulted in 49,237 rows of data being classed as problematic - leaving 198,272 observations for analysis.

As with most frequencies in `reddit` data the number of new up-votes per observation was highly skewed with a large number of very small values and a small number of very large values. In this data the largest number of new up-votes for a single observation was 4,007 - this was recorded for a post about the death of Neil Armstrong immediately after it had hit the `reddit` Front page.

The dependent variable (number of new up-votes) is a count variable so the first model to be fitted was a Poisson Generalised Linear Regression with four binary explanatory variables indicating whether the post appeared on the New, Rising, Main or Front page(s), and with an offset of log(exposure time). The highly skewed nature of the data meant that a Poisson distribution was likely te be a poor fit as the data would be more dispersed than the Poisson distribution could accomodate (the variance exceeds the mean and in the Poisson distribution these should be the same).

Therefore a Negative Binomial Generalised Linear Regression model was also fitted with the same parameters and the fit of these two models was compared with a likelihood ratio test. The Poisson model had a log-likelihood of -974961 on 5 degrees of freedom whereas the Negative Binomial model had a log-likelihood of -323998 on 6 degrees of freedom - the likelihood-ratio test confirmed that the Negative Binomial model offered a significantly better fit to the data ($\chi^2 = 1520734$, p < 0.001).

The parameter coefficient estimates for the Poisson and Negative Binomial regression models were very similar - the main difference between the models being that the Negative Binomial model coefficients had larger standard errors.

As the data were collected at intervals a variable measuring the 'exposure time' (number of minutes since the previous observation) was included in all models. For most observations this gap is 30 minutes but when posts were observed for the first time on the New page the gap is shorter than 30 minutes. By including this variable as an offset parameter the model produces coefficients which relate to one minute of time.

This data-set has another peculiar aspect in that it contains a large number of zero-counts. In total 101,597 (51%) of the observations showed no new up-votes for their post. For this reason some more elaborate models which are geared towards count data with a high number of zero-counts were fitted on a random sub-set of 10,000 cases from the data. Specifically, the 'hurdle' and 'zero-inflated' models detailed by Zeileis et al. (2008) were fitted. Neither of these models resulted in a significant improvement in fit over the Negative Binomial model without a hurdle or zero-inflated component. The number of zero-counts predicted by each of these models were also compared to the number of observed zeroes in the sub-sample (5216) - and while the Poisson model predicted too few zeros (4469) the Negative Binomial model produced a closer estimate (5416). The hurdle model under-predicted zeroes (5019) to roughly the same degree as the Negative Binomial model over-predicted. A Negative Binomial model with a zero-inflated component produced the most accurate prediction (5231) but the model itself did not offer a significantly better fit given the number of additional parameters to be estimated.

Given that this data-set contains a lot of zero-counts, why did the modelling components which are designed to handle situations with large numbers of zero-counts add so little explanatory power to the model? The answer most likely lies with the fact that 84,143 (83%) of the observed zero-counts occurred when the post in question did not appear on (the first leaf of) any of the pages of interest. The hurdle and zero-inflated components are useful where the instances of zero-counts are in some way distinct from the non-zero cases. In this context having a 'new up-votes' value of zero is a common outcome when the post does not appear on any of the pages where it is likely to be seen and voted on. Therefore the model handles the high number of zero-counts quite neatly with a very low intercept term. The Negative Binomial regression model without a hurdle or zero-inflated component was therefore selected as the modelling tool of choice for this data. The model is given as:

$$ln(\hat{V}_i) = \eta_i = \beta_0 + \beta_{\text{new}} + \beta_{\text{rising}} + \beta_{\text{main}} + \beta_{\text{front}} + ln(E_i)$$

with $V_i \sim$ Negative Binomial$(\hat{V}_i, \theta)$

where V_i represents the number of new up-votes, E_i is the exposure variable for up-vote count i, β_0 is the model's intercept and the other βs are coefficients for the binary explanatory variables (pages the post appeared on when V_i was observed). θ is the overdispersion parameter for the negative binomial distribution, with dispersion set to $1 + \theta(\eta_i)$. A dispersion parameter of zero therefore indicated no overdispersion.

Table 5.1 contains details of this model fit. This model accounted for 71% of the deviance in the number of new up-votes. The coefficients in the model can be used to produce estimated 'up-votes per minute' rates by summing the page type coefficient(s) and intercept term and exponentiating the result. These estimated up-vote rates are also included in the table. In order to calculate the estimated rate for a post which appeared on the Front page - the Front page coefficient was added to the Main page coefficient because it is not possible for a post to appear on the default Front page without also appearing at the top of the Main page for the subreddit it was submitted to.

	Coefficient	Std Error	p value	Est. Up-votes per minute
Intercept	-4.0036	0.0049	< 0.001	0.0182
On New page	0.3828	0.0093	< 0.001	0.0267
On Rising page	1.3174	0.0169	< 0.001	0.0681
On Main page	3.3857	0.0078	< 0.001	0.539
On Front page	3.3183	0.0281	< 0.001	14.89

Table 5.1: Showing model parameters for a negative binomial regression model of number of new up-votes by page location with an offset for the time since previous observation. The dispersion parameter θ was estimated to be 0.7489 with a standard error of 0.004. This model accounted for 71% of the deviance in number of new up-votes.

This model confirms the expected relationship between page type and up-votes and provides some estimates of the difference in levels of up-voting between these pages. If we consider the order in which a post that ultimately reached the Front page would appear on each of these pages it would be New -> Rising -> Main -> Front. There is a clear effect here whereby each page in this sequence is associated with a large increase in voting rate for posts which appear there. As a post moves through this sequence it is at each stage presented to a larger audience and the votes of users who see it on a given page will determine whether it progresses to the next page in the sequence.

Thus, the ubiquitous voting system and 'Hot' algorithm, which in principle operates on two criteria (score and time), in practice results in a three-stage process because of the way content is displayed on the website. While in principle it is possible for a post to acquire a high enough score to reach the Front page while it appears on the New page - in reality this is highly unlikely to occur because the voting rate for posts on the New page is much too low. Instead the outcome of voting which

occurs on the New page is likely to be limited to determining whether the post will be displayed on the Rising page. If the post appears on the Rising page the number of up-votes it accumulates by virtue of its placement there will to some degree overshadow the votes it is receiving while it remains also on the New page. The same kind of stepping-up of voting rate occurs when a post reaches the Main page for it's subreddit - the model suggests that the up-shift in voting rate between the Rising and Main pages is much larger than between the New and Rising pages. In fact this model suggests that if there is one critical determinant of whether a post will appear on the Front page it may be whether the post can make it to the Main page for its subreddit and how it is received when it appears there. This is a question which will be addressed further in section 5.1.5.

Modelling other types of activity

We have seen that there is a sharp rise in number of up-votes for posts as they move through a sequence of pages - does this relationship hold true for other forms of activity? The same Negative Binomial regression models applied to new up-votes per observation in the previous section are here applied to down-votes (table 5.2) and comments (table 5.3).

	Coefficient	Std Error	p value	Est. Down-votes per minute
Intercept	-4.2552	0.0057	< 0.001	0.0142
On New page	-0.0298	0.0115	< 0.01	0.0138
On Rising page	1.2617	0.0211	< 0.001	0.0501
On Main page	2.8718	0.0092	< 0.001	0.25
On Front page	3.3183	0.0281	< 0.001	11.048

Table 5.2: Showing model parameters for a negative binomial regression model of number of new down-votes by page location with an offset for the time since previous observation. The dispersion parameter θ was estimated to be 0.5452 with a standard error of 0.003. This model accounted for 67% of the deviance in number of new down-votes.

These models suggest a strong relationship between the rate of new up-votes, down-votes and comments. The rate at which new down-votes are acquired is in general slightly lower than for up-votes, as evidenced by the smaller Intercept and similar coefficients. The pattern for down-votes differs to that for up-votes in that the New page is associated with a decreased rate while the Front page is associated with an increased rate relative to other pages.

For comments the pattern is very similar to that observed for up-votes but the rate of new comments is in general much lower than for up-votes. The coefficients for new comments per page type are very similar to those for up-votes aside from a slightly stronger effect of being on the Main page and a slightly weaker effect of being on the Front page. This is not surprising and is likely the

	Coefficient	Std Error	p value	Est. Comments per minute
Intercept	-5.7235	0.0096	< 0.001	0.0033
On New page	0.3139	0.0166	< 0.01	0.0045
On Rising page	1.2108	0.0297	< 0.001	0.011
On Main page	3.6267	0.0126	< 0.001	0.1228
On Front page	2.9853	0.0384	< 0.001	2.431

Table 5.3: Showing model parameters for a negative binomial regression model of number of new comments by page location with an offset for the time since previous observation. The dispersion parameter θ was estimated to be 0.4031 with a sandard error of 0.003. This model accounted for 66% of the deviance in number of new comments.

result of the comments pages for posts appearing 'saturated' when these have been on the Front page for some time (i.e. a user may be discouraged from commenting on a post which already has 1,000 comments because the chance of their comment being seen is perceived as small).

These models suggest that there is enough similarity between the distribution of up-voting, down-voting and commenting across different page types to consider any of these in isolation as a reflection of a more general activity level. The underlying basis of these trends is the number of users who will encounter a post when it appears in each location - there are only slight variations in how likely a user is to up-vote, down-vote or comment on the post as a function of it's location. It should be re-iterated at this point however that the analyses reported in this and previous sections only concern posts to the Worldnews subreddit.

In the analyses which follow this relationship between each of these measures and 'general activity level' will be tested but not reported unless it is found to be invalid. It can therefore be assumed that the relationship between an individual activity measure and general activity holds true unless stated otherwise.

'Next' pages

Thus far we have considered page types in their 'default' configuration - i.e. a given page holds 25 posts. In practice `reddit`'s pagination system is a little more nuanced. Firstly, while a page (e.g. the Front page) holds 25 ranked posts, at the bottom of the page is a 'Next' button which when clicked loads a second page displaying the posts ranked 26-50 (this will be referred to here as the 2nd 'Leaf' of the Front page). A user can keep clicking the 'Next' button and keep viewing subsequent Leafs of the page. Secondly, Users with accounts can alter the number of posts which are displayed on each Leaf of a page - this can be set to 10, 25, 50 or 100.

Reddit's API allows one to specify how many posts should be retrieved from a page with a maximum

value of 100. In the previous section we considered a post to be on a certain page if it appeared there within the top 25 ranks (and therefore on the first Leaf of the page for users who had not customised this parameter). In this section we will consider the Leaf of a page which the post appeared on. Posts which appeared on a page with a rank of between 1-25 are classed as being on Leaf 1 of that page, posts which appeared with a rank of 26-50 are classed as being on Leaf 2, etc.

It should however be noted that the Rising page is an exception to this rule in that only a certain number of posts for a given subreddit are classed as 'Rising' at any given time - and the Rising page will only show as many posts as currently meet that criteria. In the case of the Worldnews subreddit during this window it appears that there were never more than 25 posts classified as Rising at any observation point.

	Coefficient	Std Error	p value	Est. Up-votes per minute
Intercept	-4.6517	0.0117	< 0.001	0.010
New - Leaf 1	-0.002	0.01	0.841	0.010
New - Leaf 2	-0.366	0.01	< 0.01	0.007
New - Leaf 3	-0.3745	0.01	< 0.001	0.007
New - Leaf 4	-0.3531	0.01	< 0.001	0.007
Rising	1.4518	0.0149	< 0.001	0.040
Main - Leaf 1	3.8467	0.0109	< 0.001	0.447
Main - Leaf 2	1.7474	0.0115	< 0.001	0.055
Main - Leaf 3	1.0098	0.0124	< 0.001	0.026
Main - Leaf 4	0.573	0.0134	< 0.001	0.017
Front - Leaf 1	3.8122	0.022	< 0.001	20.23
Front - Leaf 2	2.6264	0.0342	< 0.001	6.18
Front - Leaf 3	2.1378	0.0346	< 0.001	3.79
Front - Leaf 4	1.8153	0.0347	< 0.001	2.75

Table 5.4: Showing model parameters for a negative binomial regression model of number of new up-votes by page (and leaf thereof) with an offset for the time since previous observation. The dispersion parameter θ was estimated to be 1.2356 with a standard error of 0.007. This model accounted for 80% of the deviance in number of new up-votes. In generating estimates for the number of up-votes per location posts appearing on the Front page are assumed to appear simultaneously on the Main page - Leaf 1

A negative binomial regression model was fitted to the data with presence on the leaf of each page type included as a factor with five levels (0 meaning not present on any leaf of that page). Details of the model are included in table 5.4. For the New page there is a contrast between leaf 1 as compared to the other 3 leafs which all exhibit the same (very low) voting rate. The subreddit's

Main page exhibits strong effects for the different leafs - leaf 1 has an estimated up-vote rate of 0.45 per minute and this drops to 0.05 per minute on leaf 2 and 0.02 per minute on leaf 4. If we compare these effects with the effect of being on the Rising page it appears that the Rising page is associated with a higher voting rate than the 3rd or 4th leaf of the Main page. The Front page exhibits strong voting rates for all of its leafs - especially when we consider that for a post to appear there even on leaf 4 it will also most likely appear prominently on the Main page. If we assume posts on the Front page always appear on leaf 1 of the Main page (always true with this /r/worldnews data-set) the model estimates a rate of 20 up-votes per minute for leaf 1 of the Front page, dropping to 6.2 up-votes per minute on leaf 2 and 2.75 up-votes per minute on Leaf 4.

The model suggests that users are most likely to browse past leaf 1 of the Front page. This is to be expected as for users with an appetite to be shown more than 25 posts the second leaf of the Front page is the most obvious place to look (this is assuming that users tend to start a browsing session on the Front page). On the other hand this model suggests that very few users browse past the first leaf of the New page - again this is not a surprising result, the New page has a very high rate of turnover so once a user had browsed through the contents of the first leaf they could refresh the page and be presented with more recent posts. Conversely, if a user spends a spell of time browsing the posts on leaf 1 of the New page and then clicks 'Next' to load the 2nd leaf they will probably encounter some of the posts they have just seen on leaf 1. This could be construed as a design flaw, but in any case the prognosis for a post which slips off the first leaf of the New page without making it on to the Rising page is bleak.

In all cases a regression model with 'page leaf' explanatory variables offer a significantly better fit than models which only consider placement on leaf 1 of pages.

Comparing other default subreddits

Negative Binomial regression models with 'page leaf' explanatory variables were fitted seperately to up-vote data covering the same time period from the Technology, Movies, Funny, BestOf, Atheism and AdviceAnimals subreddits. There is a high degree of consistency in most of the patterns in voting activity between these subreddits, suggesting that these patterns are universal on `reddit` amongst the default subreddits.

If we compare voting rates on 'leaf 1' of the New, Rising, Main and Front pages the rate is always hightest on the Front page, followed by the Main, Rising and then New pages. When the negative binomial regression model of up-vote rate per page leaf is fitted to other subreddits this pattern remains constant but the difference between up-vote rate on each page varies. This is best illustrated by contrasting /r/worldnews with a more active subreddit. Shortly after the data analysed here was collected a website was launched (`stattit.com`) which displays activity metrics

for subreddits - /r/funny was ranked number 1 in terms of average users online in the previous 24 hour period while /r/worldnews was ranked 26th by the same metric (so behind a considerable number of non-default subreddits). If we compare the data collected in this observation period from the /r/worldnews and /r/funny subreddit pages there is a clear difference in the number of posts being submitted - there are records of 9,598 individual posts to /r/worldnews and 124,367 individual posts to /r/funny.

	Coefficient	Std Error	p value	
Intercept	0.836	0.033	< 0.001	2.3
New - Leaf 1	-2.051	0.036	< 0.01	0.297
New - Leaf 2	-2.754	0.036	<0.001	0.015
New - Leaf 3	-2.699	0.035	< 0.001	0.015
New - Leaf 4	-2.689	0.034	< 0.001	0.015
Rising - Leaf 1	0.577	0.019	< 0.001	4.108
Rising - Leaf 2	0.255	0.018	< 0.001	2.977
Rising - Leaf 3	-0.43	0.019	< 0.001	1.500
Rising - Leaf 4	-0.853	0.033	< 0.001	0.983
Main - Leaf 1	1.669	0.034	< 0.001	12.244
Main - Leaf 2	1.049	0.034	< 0.001	6.586
Main - Leaf 3	0.316	0.034	< 0.001	3.165
Main - Leaf 4	-0.348	0.034	< 0.001	1.629
Front - Leaf 1	1.822	0.022	< 0.001	75.717
Front - Leaf 2	0.981	0.0252	< 0.001	32.655
Front - Leaf 3	0.633	0.024	< 0.001	23.057
Front - Leaf 4	0.395	0.022	< 0.001	18.174

Table 5.5: Showing model parameters for a negative binomial regression model of number of new up-votes by page (and leaf thereof) for the /r/funny subreddit - with an offset for the time since previous observation. The dispersion parameter (θ) was estimated to be 2.01 with a standard error of 0.006. This model accounted for 83% of the deviance in number of new up-votes. In generating estimates for the number of up-votes per location posts appearing on the Front page are assumed to appear simultaneously on the Main page - Leaf 1

The negative binomial regression model of up-vote rate for a random sample of 50,000 observations of posts submitted to the /r/funny subreddit is presented in table 5.5. The main difference between this model and the equivalent model for /r/worldnews is the presence of a much larger intercept term. This equates to a much higher predicted voting rate accross all of the pages for /r/funny. On leaf 1 of the New page of /r/funny the model predicts an up-vote rate of 0.296 per minute (as compared to 0.009 per minute for /r/worldnews). On leaf 1 of the Rising page the model

for /r/funny predicts an up-vote rate of 4.1 per minute (as compared to 0.04 per minute for /r/worldnews). On leaf 1 of the /r/funny Main page the model predicts an up-vote rate of 12.24 per minute (as compared to 0.45 for /r/worldnews). Posts to /r/funny which appear on the `reddit` Front page can expect to receive up-votes at a rate of 75.7 per minute (as compared to 20 per minute estimated by the 'page leaf' model for /r/worldnews posts). The /r/funny subreddit has much higher activity rates on pages which are specific to the subreddit, but when posts from these subreddits appear on the Front page the difference in voting rate is much smaller in magnitude.

The New page for /r/funny sees a much higher rate of post submission (13x higher than /r/worldnews) and voting (33x higher than /r/worldnews). The immediate effect of this is a much higher number of posts classified as 'rising' on /r/funny. At every observation point within the observation period there were at least 50 posts classified as 'rising' on /r/funny and at 15% of the observation points there were at least 100 posts classified as 'rising' (the maximum which could be retrieved through the API). In contrast the /r/worldnews subreddit never had more than 25 posts classified as 'rising' at any observation point. The Rising page is also the location at which /r/funny activity surpassed /r/worldnews activity by the greatest margin (102x higher voting rate for /r/funny).

The Main page for /r/funny had an up-vote rate 13x higher than the same page for /r/worldnews, and when posts from /r/funny appeared on the Front page they had an up-vote rate just 3.8 times higher than /r/worldnews. To summarise this comparison: there are still many users seeing and voting on /r/worldnews posts which reach the Front page - but the pages where users decide which posts from /r/worldnews will be displayed on the Front page have much lower voting rates. Based on these measures alone one would expect that the /r/funny subreddit is in a much better state of 'health' than the /r/worldnews subreddit. There is much more user activity feeding into the decisions about whether a post passes each of the 'hurdles' which stand between it and the Front page, and this is expected to result in better decisions about which posts pass each of the hurdles.

What is the locus of this trend? This is a difficult question to answer because each of these subreddits has been developing in an organic fashion for a number of years. Each has its own group of moderators and these moderators have made certain decisions about which types of content can be submitted to their subreddit, and have enforced these rules to varying degrees. These decisions and the makeup of the group of moderators themselves are not set in stone but rather subject to change. While this has been happening `reddit`'s active users have been continually making a decision about whether the subreddit is one they wish to remain subscribed to and whether they will participate on it's 'back-stage' (i.e. New, Rising) pages. Figuring out why these subreddits had the characteristics observed in September 2012 would in theory be possible with a full record of the posts and comments submitted to them - but would involve a historical research project of considerable scope and with ample resources.

As these resources are not available I will instead put forth some speculation based on the differing

nature of the content which /r/funny and /r/worldnews cater to (something which has remained broadly stable over the lifetime of the subreddits). Reddit's users are unpaid, they vote or submit content in their spare time. While a user might want `reddit` to provide them with news of current events (so they subscribe to the /r/worldnews subreddit) they might find reviewing the posts submitted there to be a little too much like work. Worldnews caters to content of a serious nature, and when some major world event takes place we might expect that it is flooded with many posts linking to different articles about the same event. One can imagine how reading through ten articles about the same event from different news organisations and trying to decide which are best could be perceived as work.

The Funny subreddit on the other hand deals with content which is intended to be humerous in nature and therefore the reviewing of such is likely to hold much more intrinsic entertainment value. Furthermore, the length of time it would take to 'read' and pass judgment on a Funny post is likely to be much shorter than for a Worldnews post. In this case even if the same number of man-hours were dedicated to reviewing content on the /r/funny and /r/worldnews New pages - this would be expected to result in a higher voting rate for /r/funny.

5.1.4 How much attention do posts receive on `reddit`'s pages?

One of the primary functions of a Social News website is to sort 'good' submissions from 'bad' and promote 'the best' submissions to the most visible locations - with these judgments being made using users' votes. As such they serve to focus the attention of their users and visitors on 'the best' content (as determined through the voting system and displayed on the Front page). It is therefore of interest to investigate the levels of attention received by content on `reddit` as a function of the page(s) it appeared on.

While voting activity can be used to infer relative levels of attention between different pages, it cannot provide reliable information about the number of people who see the items being voted on. `Reddit` does not provide information on how many times the resource linked to by a post is viewed. This information can however be obtained in some cases by consulting the web resource being linked to. `Imgur.com` is an image-sharing website created by a `reddit` user (redd.it/7zlyd) to circumvent a problem often referred to as 'the `Slashdot` effect' or 'the `Digg` effect' (and if it were not for a service like `Imgur` this would perhaps be known now as 'the `reddit` effect'). This problem occurs with popular Social News posts that link to smaller websites; the influx of traffic from the Social News website overloads the external website's server with the result that the content being linked to becomes unavailable. Since its creation `Imgur` has become popular among `reddit` users, and many of the image posts submitted to `reddit` now link to a copy of the image hosted on `Imgur`.

Imgur displays information on how many times the images hosted there have been viewed (hits). For a period of study in August 2010 the `Imgur` website was scraped to extract this information for image posts which were being tracked on `reddit`. Posts to the 'pics' subreddit were monitored, and over a three week period 95,702 observations of the hit rates for these posts on `Imgur` were recorded. A negative binomial generalised linear model was fitted to these hit rates, with the `reddit` page a Post appeared on at that time as the explanatory variable. The results of fitting this model are presented in table 5.6.

Page	Coefficient	Std Error	p value	Est. Hits per minute
Intercept	1.92	0.005	< 0.001	6.8
On Rising page	-0.18	0.01	< 0.001	5.7
On Main page	2.02	0.01	< 0.001	51.4
On Front page	5.55	0.03	< 0.001	137

Table 5.6: Showing model parameters for a negative binomial regression of `Imgur` hit rate by location.

This model suggests that an item submitted to the 'pics' subreddit which reaches the Front page could expect to receive around 137 hits per minute while it was located there. Items appearing on the Main page for the 'pics' subreddit could expect to receive around 51.4 hits per minute, while items on the Rising page received just 5.7 hits per minute on average. In contrast to the model of voting activity above, items which did not appear on any of the listed pages actually received more hits than those which appeared on the Rising page. The reason for this is suggested by figure 5.5, showing number of hits over the lifespan of a single `reddit` post.

Figure 5.5 depicts the lifespan of a popular 'pics' post; this post appeared on the Front page by its third observation, where it remained until the 42nd time it was observed (represented by the dashed line). The graph indicates that the level of voting activity on the post dropped off several hours before it slipped off the Front page; but that it was still receiving fresh hits at a steady rate several hours after slipping off the Front page. This is likely due to the fact that the post had just been seen by around 180,000 people, and some proportion of these individuals may have forwarded a link to the image to their friends/contacts through e-mail or some other on-line communications system.

It should be noted that the analysis of `Imgur` hits was conducted at a time when `reddit` had a much smaller user-base and were the same analyses conducted in 2012 all of these figures would likely be much larger. A model fitted to new total votes for a number subreddits using data from this period predicted 5.1 votes per minute for posts on the Front page, 0.4 votes per minute for posts on the Main page of its subreddit and 0.13 votes per minute for a post on the Rising page of its subreddit. Unfortunately shortly after the data on hit-rates were collected `Imgur` fundamentally
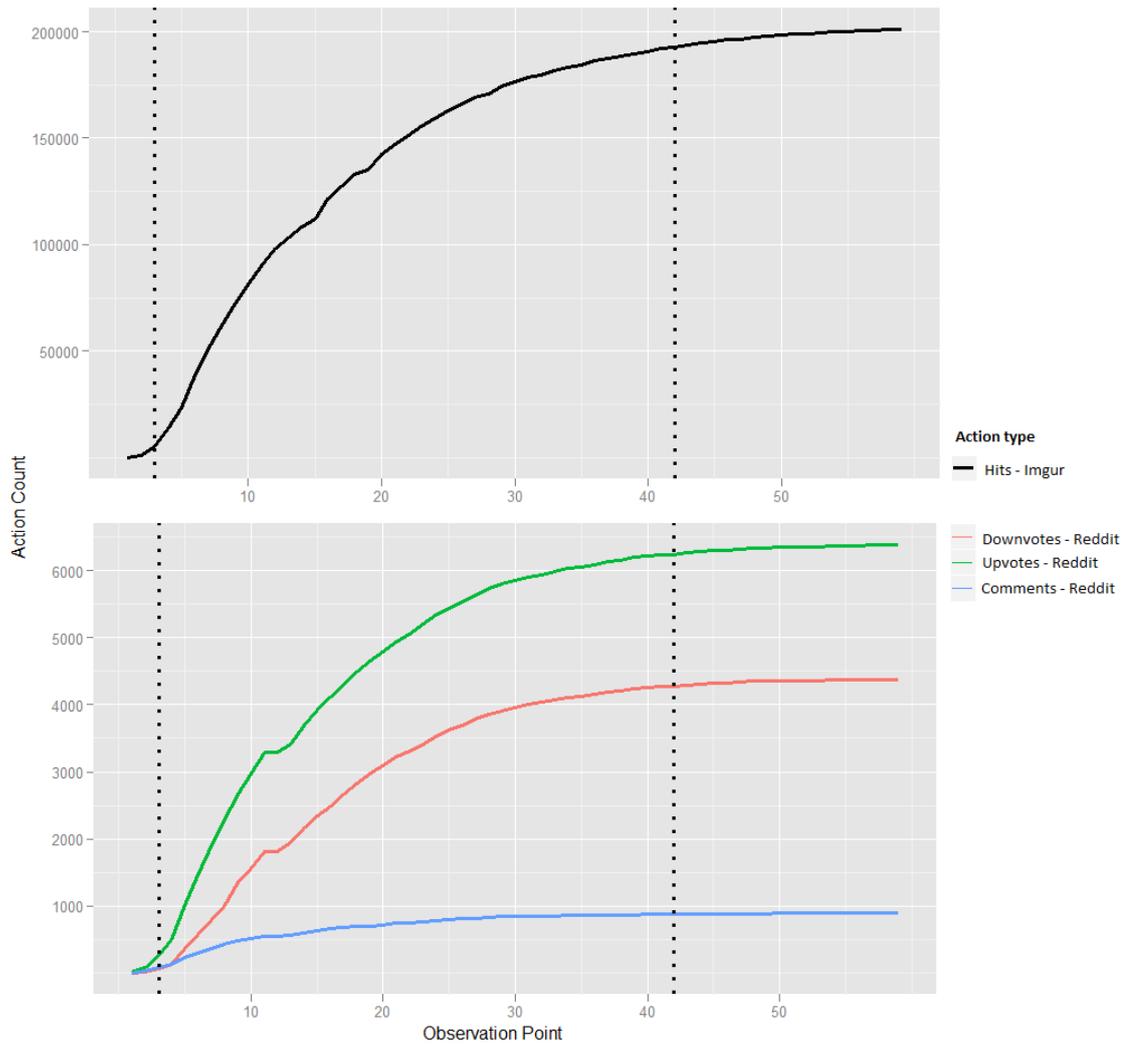
Figure 5.5: Showing cumulative Hits (upper pane), Votes and Comments (lower pane) for a single /r/pics Front page Post throughout its lifespan on `reddit`. This post appeared on the Front page first at observation point 3 and remained there until observation point 42 (represented by dashed vertical line).

changed the structure of the `HTML` which makes up their pages - and therefore collecting fresh data after this point would have required a total re-write of the scripts used to collect data from `Imgur`.

At the time when these analyses were conducted `Imgur` was very much a `reddit` 'sister site' - it was being used primarily by `reddit` (and later `Digg`) users. In the intervening period `Imgur` seems to have become more 'independent' - it now has its own voting and commenting facilities, both being quite similar to those employed by `reddit`. Were the same data collected and analyses conducted today the viewing habits of `Imgur` users who are not `reddit` users would likely add considerable noise. When the data was collected `Imgur` had very few facilities for browsing images on the website and therefore we can assume with confidence that images submitted to `reddit` were being seen primarily *through* Reddit, individuals either directly following links from `reddit`

or indirectly by receiving a hyper-link from a `reddit` user.

5.1.5 Reaching the Front page

`Reddit`'s Front page represents the pinnacle of visibility on the site and is referenced in `reddit`'s self-selected title, "The Front page of the Internet". This section considers the utility of the available explanatory variables in predicting whether a post will reach the `reddit` Front page - and by doing so seeks to expand our understanding of how this central aspect of `reddit` fuctions. It bears re-iterating at this stage that there is no universal `reddit` Front page and therefore this section refers to the 'default' Front page and consequently only to 'default' subreddits.

In section 5.1.2 we considered `reddit`'s 'Hot' ranking algorithm - the algorithm that generates the ranking scores used to determine which posts appear on the Front page (and also the Main pages for individual subreddits). This algorithm operates on just two parameters - a post's score and time of submission. One way of thinking about the passage of a post to the Front page involves voting 'momentum'. Every post submitted to `reddit` appears for some duration on the New page of the subreddit it is submitted to. Beyond the New page placement on any other page must be earned through the accumulation of score. A post must therefore acquire a high enough score to be positioned on the Rising page before being pushed from the New page - and maintain a high rate of score accumulation if it is to ultimately reach the Front page. This suggests that early voting performance for a post may be a good predictor of whether it reached the Front page.

In section 5.1.3 models of voting activity across `reddit`'s pages revealed a considerable stepping up of voting activity as a post appeared on the New, Rising, Main and Front pages. If we assume that a post passes through these pages in this order, a post's passage to the Front page can be conceived of as overcoming three obstacles or hurdles - appearing on the Rising page before being pushed off the New page, appearing on the Main page before being pushed off the Rising page and then outscoring other posts which also appear on the Main page (and to some degree the posts appearing simultaneously on the Main pages of other subreddits) to reach the Front page.

If this 'hurdling' scenario is applicable we would expect progressivly fewer posts to reach each page in the sequence - and posts which appeared on pages later in the sequence to have a progressivly higher chance of going on to appear on the Front page. In section 5.1.3 we saw a considerable difference in the 'depth' of voting activity on the /r/funny and /r/worldnews subreddits (/r/funny had a much greater level of activity on the New and Rising pages compared to /r/worldnews - relative to the gap in activity on the Front page). This should result in a stronger form of the 'hurdling' process on /r/funny than /r/worldnews.

The first step towards looking for evidence of this kind of process was to consider the number of posts to each subreddit which reached each page in the sequence. Figure 5.6 shows the number of

posts which appeared on the New page and how many of these went on to appear on the Rising, Main and Front pages. For this figure appearance on any of the four observed leafs of a page is counted as an appearance on that page. There are two dominant patterns on display in figure 5.6 - subreddits which show the expected drop in posts which reach each page in the sequence (e.g. /r/funny) and subreddits on which very few posts have been excluded from appearing on the Rising and Main pages (e.g. /r/worldnews).
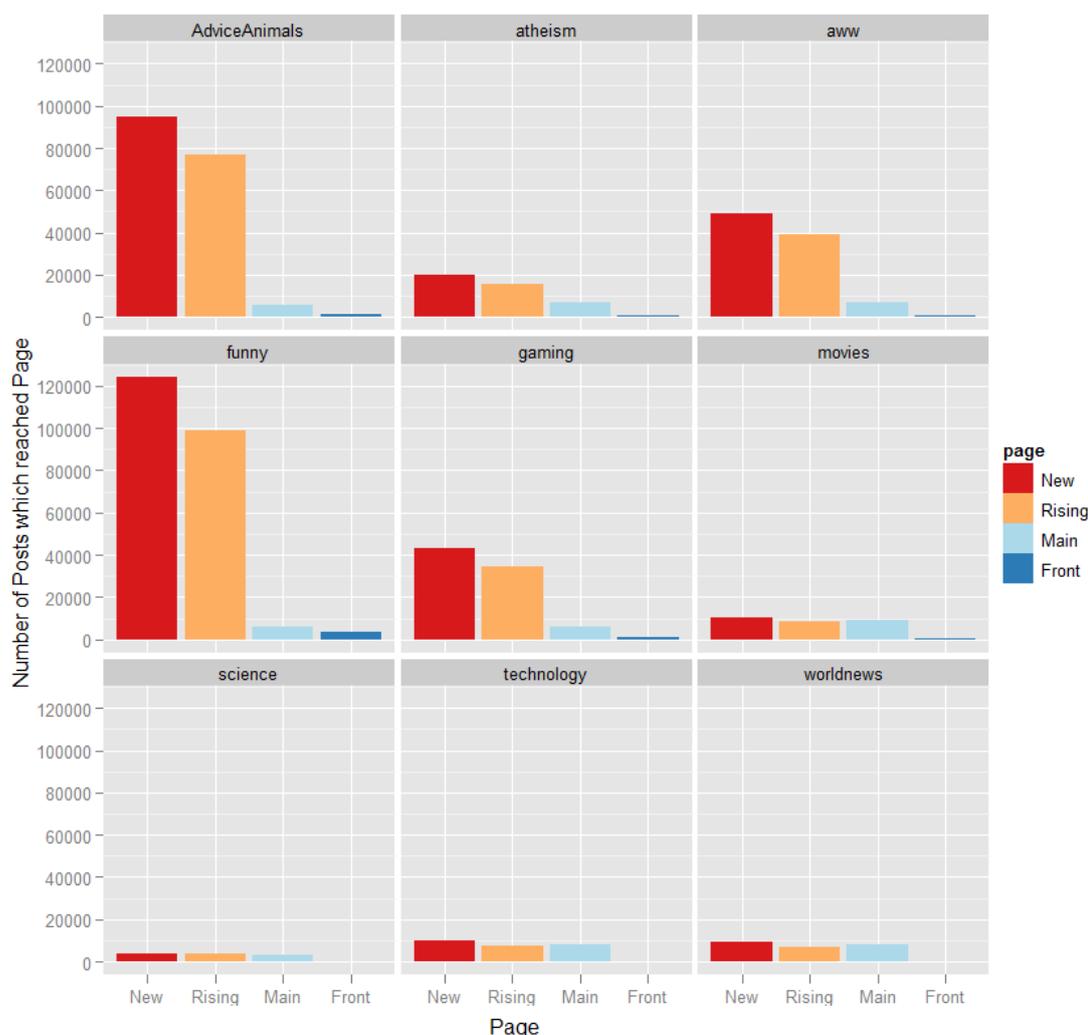
Figure 5.6: Showing the number of unique posts which appeared on the four observed page types for a number of subreddits - August 20th to September 27th 2012

For subreddits which fall into the former category the primary filtering point seems to be the transition from Rising to Main page. The number of posts which appeared on the Main page is similar for all of the subreddits, whereas the Rising page seems to display a certain percentage (70-80%) of posts which appeared on the New page. Therefore whether a subreddit exhibits the 'hurdling' effect appears to be governed by its rate of post submissions.

Looking at the subreddits which do not exhibit the 'hurdling' effect (movies, science, technology,

worldnews) reveals two things. First, these subreddits have a higher number of posts which appeared on their Main page than their Rising page. Secondly, the number of posts which appeared on their Main page tends to be higher than the number of posts which appeared on the Main pages of more active subreddits that exhibit the 'hurdling' effect. This second point in particular suggests that the New and Rising pages for these subreddits are not contributing much to the decision about which posts go on to the Front page - most of the posts submitted to these subreddits appear at some point on the first four leafs of the Main page and it may not be until this point that a judgment on these posts is reached through the voting system.

The subreddits which exhibit the strongest filter or hurdle between the Rising and Main page (in this set /r/funny and /r/AdviceAnimals) have the highest rates of post submission and also the highest rates of voting on the New and Rising pages. One can assume that the voting activity which takes place on the New and Rising pages of these subreddits leads to a better decision about which posts should progress in the sequence and appear on the Main page - as compared to subreddits with much lower activity on these pages where the majority of posts reach the Main page.

To test this assumption a measure of post 'quality' is required - where 'quality' is the attribute(s) users are seeking when they browse `reddit` and varies considerably between subreddits. . The obvious candidate is a post's score, with the caveat that 'quality' in this context is a reflection of how `reddit`'s users have voted rather than conforming to some concrete definition of the term. Furthermore, a post's final score will be heavily influenced by the pages it has appeared on - i.e. a post which appeared on the Front page will always tend to have a higher score than a post which almost reached the Front page, and this distinction could easily result from fortuitous timing rather than a difference in 'quality'.

Bearing these caveats in mind, consideration of the final scores achieved by posts is useful in this context. In particular, whether a post's final score was positive or negative conveys valuable information. If the final observed score for a post is negative this is a strong indication that the post did not meet the criteria which `reddit`'s users look for when they up-vote posts (the users who saw it were more likely to down-vote it than up-vote it). If the voting which takes place on the New and Rising pages of subreddits which exhibit a hurdling effect serves a purpose - this would be expected to result in higher scores for the posts which reach the Main pages for these subreddits as compared to posts which appear on the Main pages of subreddits which do not show evidence of filtering. In particular, we would expect fewer 'bad' posts (those which ultimately had a negative score) to appear on the Main pages of subreddits that exhibit hurdling.

Figure 5.7 reveals exactly such a trend. Subreddits which do not exhibit a hurdling effect allow more 'bad' posts (posts which had a negative score at their final observation) and 'unremarkable' posts (posts which finished with a small positive score) to reach their Main page. This is quite a strong effect, occasions where a post appeared on the Main page of a subreddit with the hurdling
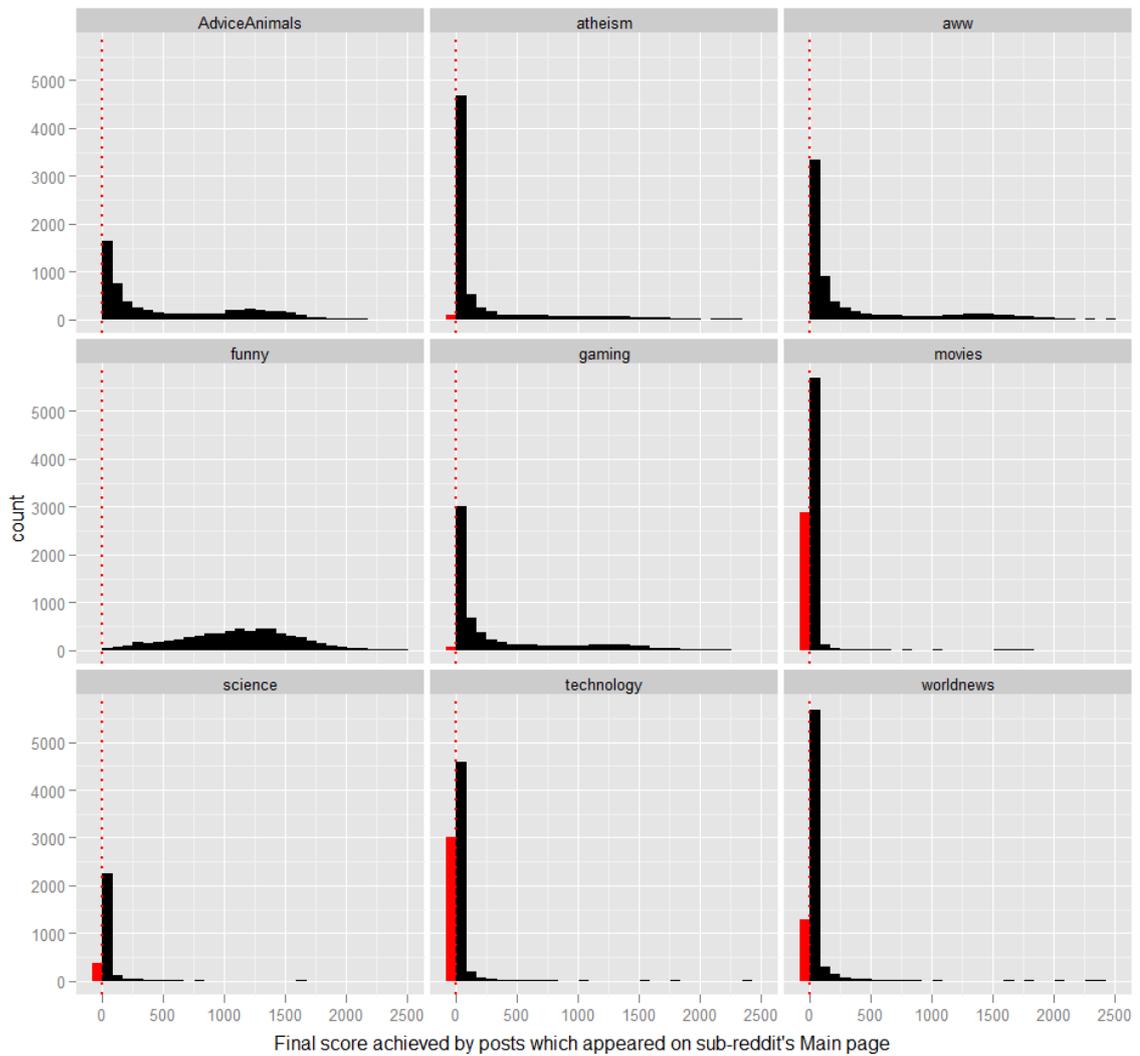
Figure 5.7: Showing the final scores achieved by posts which appeared on the Main page (leafs 1-4) of their subreddit. Excludes 132 posts with a final score of greater than 2500. Bars representing posts with a score of zero or less are coloured red. 'aww' is a subreddit for cute pictures of animals, 'AdviceAnimals' caters to image macro memes.

effect and then went on to finish with a negative score are very rare. In contrast this happened regularly on subreddits without the hurdling effect (7% of the posts which appeared on the Main page of /r/science ended with a negative score ranging to 23% for /r/technology). This could in principle be due to a difference in the number of posts to each subreddit that end with a negative score but table 5.7 rules out this possibility. The very low number of posts to /r/aww which ended with a negative score bears comment - it seems that users make much less frequent use of the down-vote button when the content being voted on concerns cute pictures of animals.

Returning to figure 5.7, the histogram for /r/funny is unusual in the context of **reddit** because it bears a strong similarity to the normal distribution. This is remarkable because every other activity distribution on **reddit** which has been hitherto considered (both those pertaining to posts

Subreddit	% All Posts with Final Score Negative	% of Main page Posts with Final Score Negative
AdviceAnimals	22%	< 0.01%
atheism	20%	< 0.01%
aww	4%	0%
funny	21%	< 0.01%
gaming	30%	< 0.01%
movies	22%	18%
science	14%	7%
technology	31%	23%
worldnews	17%	9%

Table 5.7: Showing the percentage of all posts, and of posts which appeared on the Main page for their subreddit, which had a negative score at final observation

and to users) is much closer in shape to the highly skewed power law distribution. This could be due to the filtering of posts or to the high activity levels on the Main page of /r/funny. Perhaps if posts are sufficiently filtered such that only 'good' posts remain their scores will tend to follow the normal distribution. It could also be the case that this is a result of having an abundance of voting users and activity - in a scenario where every post will be rated by enough users to produce a 'fair' score these scores will be more likely to follow the normal distribution. These factors are likely to work against the presence of a peak at the low end of the distribution. To put this another way: subreddits tend to have a power law type distribution of post scores because there are either not enough voting users to rate all of the posts enough times to produce a 'fair' score, or because there are a high number of unremarkable posts (which would naturally have a low score) in the data-set.

A third factor concerns the activity differential between the /r/funny Main page and the `reddit` Front page. To re-visit section 5.1.3, models estimated a voting rate for posts submitted to /r/funny which was 6 times higher on the Front page than on the subreddit's Main page. This is in contrast to a subreddit like /r/Worldnews where posts on the Front page could expect to receive up-votes at a rate which is 27 times higher than on the subreddit's Main page. The relatively small differential for /r/funny as compared to other subreddits likely results in a distribution of scores which is less 'stretched' towards the high end of the distribution.

Predictors of Front page success

It is not possible to model the impact of placement on each page in the New -> Rising -> Main -> Front sequence because Posts which did not appear on these pages never appeared on the Front page. This section instead considers the utility of very early voting activity in predicting whether a post would reach the Front page. The first models to be fitted were binary logistic regressions with

whether a post appeared on the Front page as the response variable - and scorechange-per-minute (calculated by dividing its score at first observation by the time delay between submission and observation in minutes) as the only indicator variable. These models revealed a significant effect of scorechange-per-minute for eight of the nine subreddits under consideration - the effect was not significant for the /r/worldnews subreddit.

The effect was strongest for the /r/funny subreddit, where the model's Intercept (in this case representing posts with a scorechange of zero) suggested that posts had around a 2.6% chance of reaching the Front page and that this would increase by 0.4% for a post which achieved a scorechange of +1 per minute until its first observation (with the opposite being true for posts whose score decreased). The effect was strongest for subreddits with the 'hurdling' effect and a relatively high voting rate on New and Rising, and weaker for subreddits which have lower voting rates on these early-stage pages. However, these models are quite weak - for /r/funny the model only accounted for 1% of the deviance in whether a post reached the Front page. The relationship between very early voting response and Front page success is a weak one.

It was decided that treating score at first observation as a categorical variable may yield more success - if the New page's primary purpose is to determine what appears on the Rising page then merely knowing whether a post's score was positive or negative at first observation might prove as useful as knowing how quickly it's score was rising or falling. Post score at first observation was converted into a 3-level factor (negative score, neutral score, positive score) and a binary logistic regression model with this variable as the indicator and Front page placement as the response was fitted for the nine subreddits under consideration. These models are better than the previous model with a 'scorechange-per-minute' indicator variable both in terms of the proportion of deviance they account for and their AIC values. This improvement is most pronounced for the most active subreddits (e.g. this model for /r/funny accounts for 14% of the deviance in whether a post reached the Front page).

Table 5.8 shows the model coefficients for these logistic regression models. There is quite a strong degree of similarity between the subreddits in terms of the effect of a post's score at first observation being negative, neutral or positive. For all subreddits posts which were first observed having a positive score were the most likely to ultimately reach the Front page. The strength of these models as opposed to a model based on scorechange-per-minute until first observation supports the idea that the activity at this stage is best thought of as determining whether the post will progress to the next step. Voting rate per minute at this stage is not a strong predictor of Front page success - this could be because the measures are being collected over too short a period of time (up to 30 minutes) and based on too few votes to make a strong prediction. An alternative interpretation would be that the votes at this stage serve primarily to determine whether the post will progress in the sequence and are not indicative of how the post will be received at subsequent stages. Perhaps

Subreddit	Parameter	Coefficient	Std Error	p value	Chance of Front page
AdviceAnimals	Intercept (Negative)	-6.009	0.12	< 0.001	0.25%
	Neutral Score	1.113	0.13	< 0.001	0.75%
	Positive Score	2.846	0.12	< 0.001	4.2%
atheism	Intercept (Negative)	-5.987	0.3	< 0.001	0.25%
	Neutral Score	1.885	0.31	< 0.001	1.7%
	Positive Score	3.653	0.3	< 0.001	9.7%
aww	Intercept (Negative)	-8.483	1	< 0.001	0.02%
	Neutral Score	1.113	1	< 0.001	0.7%
	Positive Score	2.846	1	< 0.001	3.9%
funny	Intercept (Negative)	-6.209	0.11	< 0.001	0.2%
	Neutral Score	1.697	0.12	< 0.001	1.1%
	Positive Score	3.646	0.11	< 0.001	7.7%
gaming	Intercept (Negative)	-5.695	0.15	< 0.001	0.33%
	Neutral Score	1.216	0.16	< 0.001	1.1%
	Positive Score	3.272	0.15	< 0.001	8.9%
movies	Intercept (Negative)	-6.939	0.71	< 0.001	0.09%
	Neutral Score	2.378	0.72	< 0.001	1%
	Positive Score	4.479	0.72	< 0.001	8.5%
science	Intercept (Negative)	-4.849	0.41	< 0.001	0.8%
	Neutral Score	1.918	0.42	< 0.001	5.3%
	Positive Score	3.037	0.42	< 0.001	16.3%
technology	Intercept (Negative)	-5.593	0.33	< 0.001	0.4%
	Neutral Score	1.535	0.35	< 0.001	1.7%
	Positive Score	3.181	0.34	< 0.001	9%
worldnews	Intercept (Negative)	-5.229	0.28	< 0.001	0.5%
	Neutral Score	1.035	0.30	< 0.001	1.5%
	Positive Score	2.71	0.29	< 0.001	8%

Table 5.8: Showing parameters for nine logistic regression models of whether a Post appeared on Front page (leaf 1) - Indicator variable is score at first observation on 3 levels (< 1, 1, >1), a score of zero is counted as negative because posts begin with a score of 1.

this is a hint that the users who vote on very fresh posts are in some manner distinct from those who vote on posts which appear on later-stage pages like Rising and Main.

5.1.6 Summary - Post voting on `reddit`

The previous sections have explored the nature of post voting on `reddit`. Several hypotheses have been confirmed - the distribution of votes between posts is highly skewed such that it approximates a power law, and the voting activity of `reddit`'s users is found to be concentrated on posts which appear on the Front page (although the strength of this effect varies between subreddits). The algorithm which `reddit` uses to turn votes into ranks was dissected and found to be rather simple - only operating on two parameters, post score and time of submission. However, the various pages `reddit` uses to display posts complicates the way which this algorithm plays out. Subreddits with a high level of post submissions and voting on the New/Rising pages exhibit a 'hurdling' effect whereby a post must pass a number of obstacles to have a chance of reaching the Front page. On these subreddits 'bad' posts are filtered out before they appear on the Main page of the subreddit - and on the Main page posts of reasonably high quality compete to reach the Front page. For a number of the less active default subreddits this filtering or hurdling process is absent - a high proportion of posts which are submitted appear on the Main page and many of these end with a negative score.

The analyses reported in this section were conducted initially in 2010 and have been re-visited in 2012 and performed again with better, cleaner data. In the intervening period one of `reddit`'s administrators wrote a blog post which considers some of the same issues (ketralnis, 2011) - and being an administrator, they have access to clean data without 'fuzzing' or caching-related issues. Firstly, they describe the process of a post reaching the Front page in a very similar way to the hurdling process outlined here - but do not state that this applies to varying degrees on different subreddits. Secondly, they report on voting activity - just over 6 million post votes were cast in a 3-day period presumably not long before the blog was published on July 17th 2011. This suggests a huge increase in voting activity as the website has grown - for the whole month of March 2009 (the only other time for which we have accurate voting activity information) there were just 3.5 million votes, in just over two years the voting rate increased around 18-fold. There is also a shift in the nature of voting, with this blog post reporting that 82.9% of votes are up-votes as compared to March 2009 when 76.5% of votes were up-votes. The blog post also notes a dip in activity levels when American users would be sleeping, a highly skewed distribution of votes between posts, and that posts which are ultimately successful tend to 'take off' very quickly in terms of their score (with numerous graphs showing this).

During the time when this research has been underway several communities have also been grappling with similar questions to those considered here. These are however not communities of academic researchers communicating through conferences and peer-reviewed journals (although there is now some research on `reddit` appearing through these avenues also) - they are communities of `reddit` users communicating on their own subreddits and using the language of `reddit`. One

such post is particularly pertinent here (redd.it/vqy9y) - the top comment, from user 'joke-away', gives an account of the voting system and 'hot' algorithm and explains why he thinks these are 'anti-content'. This critique centres on the speed at which people vote, stating that content which is quickly consumed and evaluated (e.g. pictures) does better because it achieves a higher voting rate per unit of time spent viewing - high-quality but lengthy posts fall by the wayside because they are competing with posts that turn views into votes at a much higher rate. This comment was originally submitted to a post on the /r/circlebroke subreddit (a subreddit dedicated to discussing negative aspects of reddit), where it was very popular but had a maximum audience of about 13k subscribers. Another user submitted a link to this comment to the /r/bestof subreddit (a subreddit showcasing the best of reddit) and from here the comment went on to appear on reddit's Front page where it was widely seen. Here some anecdotal evidence for the broadcasting effect of reddit's Front page comes into play - a friend of the researcher, upon learning that the researcher studied reddit, began to voice many of the opinions outlined in this post about the nature of reddit's voting system.

There is nothing put forward by this comment which the present chapter presents evidence against, but in performing rigorous quantitative analyses there are certainly aspects of the situation which have come to light. Firstly, the federal system of subreddits means that posts from light-hearted or entertaining subreddits are not competing directly with those from more serious subreddits - the top-ranking post for /r/worldnews always has a spot on the Front page but there is some 'meta-competition' between subreddits whereby subreddits with higher voting rates have a greater number of slots on the Front page (see Chapter 8).

Looking at the data, the main problem which subreddits like /r/worldnews and /r/science have is that there aren't enough votes being cast on their early-stage pages (New and Rising). If we compare /r/funny and /r/worldnews - /r/funny has an up-vote rate which is 31x higher on the New page and 100x higher on the Rising page. Part of this will be due to the fact that it takes less time to consume and vote on an /r/funny post - but this seems unlikely to be the whole story. Part of the problem for these subreddits, at the time when this data was collected (mid-2012), is simply that they don't have enough users voting in the right places. It may be the case that this situation now is a result of 'valuable' /r/worldnews users migrating to alternative subreddits (e.g. /r/worldevents) over time in response to a perceived decline in the quality of the subreddit.

The comment from 'joke-away' is nonetheless insightful, and it is important to note that this comment is the tip of an iceberg - there are several areas of the reddit website which are reasonably active and where the subject for discussion is reddit itself - how it works, its shortcomings, and how it might be improved. At this point in time it is tempting to suggest that an individual wishing to learn about reddit would be better served perusing these areas of the website than reading the handful of peer-reviewed journal articles on the subject. Certain posts from these areas will be

discussed in the relevant sections from time to time, and this 'amateur research' aspect of `reddit` will be discussed in its own right in section 8.7.1.

5.1.7   Comment ranking

The comments pages for `reddit` posts are central to the appeal of the website. Comments are subjected to the same up/down voting system as posts and ranked according to these votes. Popular posts on `reddit` can receive hundreds or thousands of comments and these comments can in turn receive hundreds or thousands of votes. The result is that when one clicks through to the comments page of a popular post one is immediately presented with the most popular comments on that post as judged by the many users who have voted on them. The nature of these popular comments varies from post to post, in some cases they are the most insightful or persuasive arguments related to the post's content, in other cases they are simply the most popular joke or pun.

When a post appears on the `reddit` Front page `reddit`'s users have collectively decided through the voting system that it is worthy of your attention. The comments page for the post tells you what this same collective of users' think about the post or what their most common reactions to it are. That is the implicit logic behind the ranking of posts and comments, it is however very difficult to assess the validity of these statements. Casual observations suggest that comment voting places highly relevant, important or useful comments at the top of a post's comments page often enough to give the impression that the top-ranking comments really are 'the best'.

Reddit's 'Best' comment sorting algorithm

In October 2009 `reddit` introduced a new default method of sorting comments (Munroe, 2009). The first thing to note about this new ranking algorithm is that it was not produced by a `reddit` administrator or an employee of `reddit`'s parent company. Rather the 'Best' comment sorting algorithm was devised by Randall Munroe, a `reddit` user and the author of the *xkcd* web-comic.

The 'Best' comment sorting algorithm is based on the Wilson score confidence interval for a Bernoulli parameter (Wilson, 1927), specifically the lower bound of the 95% confidence interval. The 'Best' comment sorting algorithm therefore makes use of the fact that comments have both up-votes and down-votes, placing the comments with the highest 'true' ratio of up-votes to down-votes at the top after allowing for the uncertainty which comes with a low number of total votes. Reddit's previous default comment sorting algorithm had been heavily biased towards early comments - the early votes these comments accrued allowed them to become entrenched in the highly-visible top part of the post's comments page. Even when subsequent comments were highly relevant or important early comments had such a 'head-start' that they could not be out-scored. Under the new 'Best' ranking the only advantage early comments with an initially higher score have

is that the confidence interval around their current ratio of up-votes to down-votes is smaller. This algorithm appears to strike a good balance between sorting based on absolute score (biased towards early comments) and sorting based on absolute ratio (biased towards newer comments with no down-votes).

'Best' comment sorting in action

While the logic of the 'Best' sorting algorithm appears sound the results which it produces will be in large part determined by the comment voting behaviour of `reddit` users. A key question here concerns the mobility of comments. If the rankings attributed to comments by the sorting algorithm are similar to their order of submission this would suggest that the comment voting system is largely ineffective.

To investigate the mobility of comments the politics subreddit was monitored extensively for a period of four days between July 26th - July 30th 2012. During this time 1,886 new posts to the politics subreddit were observed at 30-minute intervals until they were no longer active, with the top 500 comments at each observation point being recorded along with their rank and score. This yielded 42,023 comments for study. However, many of these comments were submitted to posts which saw very little activity. The 'Best' comment sorting algorithm is designed to work with highly active comments pages - as much of the attention of `reddit`'s users is focused on the Front page, and the comments pages for posts which appear there always have a high level of activity. Therefore to consider comment mobility it was determined that only those comments submitted to posts which received at least 50 'top-level' comments would be considered. Top level comments are direct responses to their parent post, these comments can in turn be replied to by 2nd-level comments which can in turn be replied to by 3rd-level comments and so on. Second and lower level comments are displayed in a thread which is attached to their top-level parent comment and therefore the visibility of a 2nd-level comment will be largely determined by the visibility of its parent comment.

46 of the posts tracked during this four-day period had at least 50 top-level comments; in total these posts had 6,652 top-level comments for which records were available. Figure 5.8 displays a kernel density plot of comments' submission order against their final observed rank. Kernel density plots can be thought of as 'three-dimensional' scatterplots with a degree of smoothing - where many data points would appear in the same location on a scatterplot this location is coloured to reflect its higher density. For these plots white represents the highest density with brown, orange and yellow in that order representing areas of decreasing density. Green represents the lowest density. In some senses the kernel density plots are analogous to ordnance survey relief maps - where colour is used to represent *data density* in an area rather than its height.

Figure 5.8: A kernel density plot showing comments' submission order against final observed rank. The plot is similar to a scatterplot and shows 'data density' when comparing comments' order of submission to their final observed rank. Colour denotes density, from highest to lowest - white, brown, yellow, green. The high-density off-diagonal represents a relationship between order of submission and final observed rank.

It is immediately apparant that there is a strong relationship between submission order and final observed rank. There is a clear off-diagonal of high density accross almost the full range of the data. The area of highest density is in the bottom-left corner, and it is indicating a strong trend wherby comments submitted early tend to remain in a prominent location. There are however also many cases where comments achieved a final rank which was very different to their submission order. There are both comments which were submitted early but slipped to a low rank and comments which were submitted relatively late but which rose to a relatively high rank. The advantage for early comments seems most prominent for the first 20 top-level comments to be submitted for a post, with these comments often maintaining a high rank.

If instead the first observed rank for these comments is compared with their final observed rank a very different picture emerges. Figure 5.9 suggests much lower mobility for comments between their

Top Level Comments submitted to posts with 50+ comments
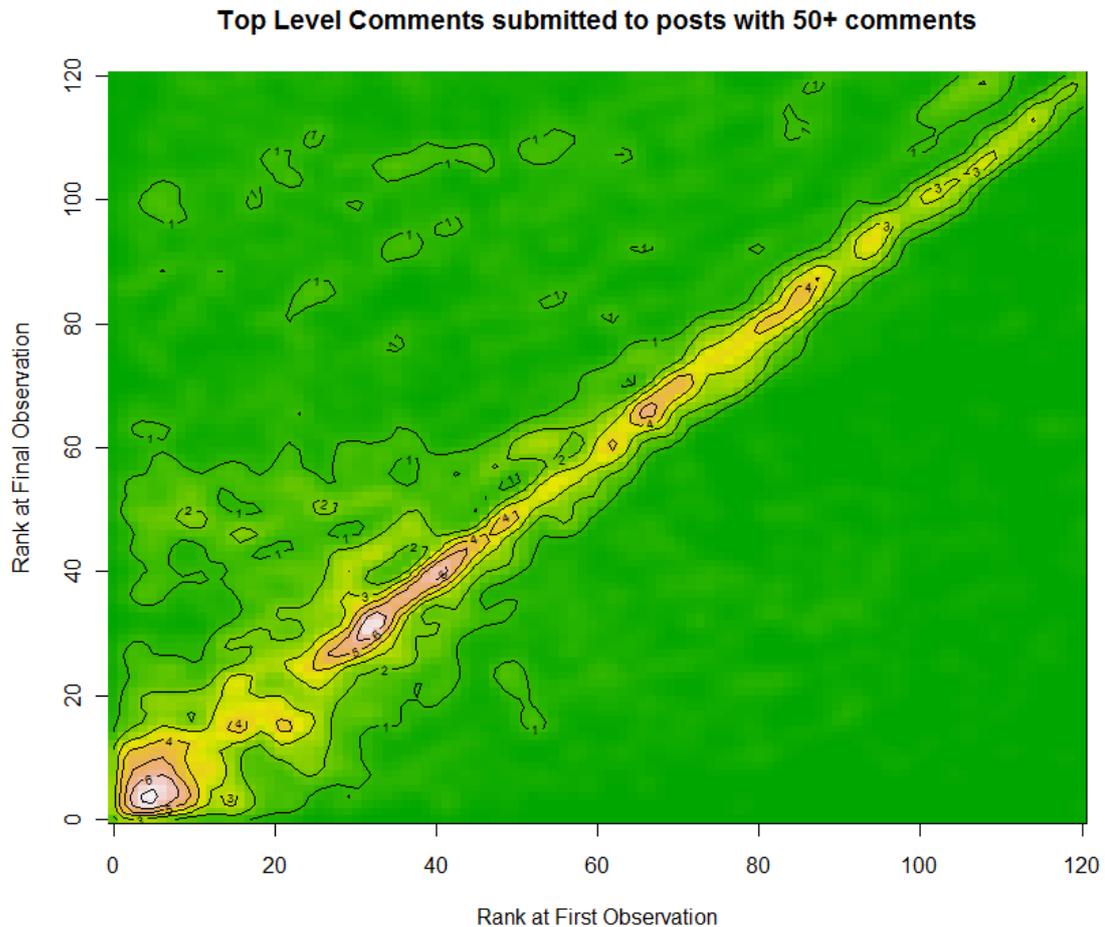
Figure 5.9: A kernel density plot showing comments' first observed rank against final observed rank. The plot is similar to a scatterplot and shows 'data density' when comparing comments' first observed rank to their final observed rank. Colour denotes density, from highest to lowest - white, brown, yellow, green. The high-density off-diagonal represents a relationship between first and last observed rank.

first observation and their final observation. Observations were recorded at 30-minute intervals and so a comment is likely to be observed for the first time within 30 minutes of its creation - for this sample there is a median gap between comment creation and first observation of 16 minutes. It appears that the votes a comment received in its first 30 minutes or so have a lot of influence on the comment's ultimate rank, with comments having a strong tendancy to maintain the rank they were first observed at or similar.

The nature of the mobility on display in figures 5.8 and 5.9 is also different. When comparing order of submission to final rank there are many comments whose ultimate rank is much higher or lower than their submission order would suggest. When comparing first observed ranks to last observed ranks big movements seem much more likely to occur in one direction - downwards. There are many comments which ended with a much lower rank than that which they held at their first

observation, while there are far fewer which made big gains in rank throughout their lifespan. In figure 5.8 there is a general trend whereby later comments end with a rank which is slightly better than their submission order; this is accounted for by the fact that these late comments will be displayed above earlier comments which were down-voted to negative scores, even if the late comments themselves receive no votes.

These findings have several implications for the utility of `reddit`'s 'Best' comment sorting algorithm. Firstly, comments submitted early have a strong advantage over those submitted later. Secondly, the first few users to see a comment and vote on it (or not) likely have a big impact on that comment's chances of achieving a high rank. Thirdly, once a comment has passed this initial 30-minute period its rank is unlikely to change substantially. Where a comment's rank does change substantially after its first observation its rank usually decreases.

This implies that most of the sorting of comments happens very quickly after their submission when the comments page is relatively fresh. As the comments page matures big changes in rank are less likely to occur, the action at this stage seems more akin to 'fine-tuning' the comments' ranks. Although these changes in rank are not very large they are likely important, it appears that any comment first observed with a rank in the top 20 had a reasonable chance of reaching the very highest ranks. There is undoubtedly a very large difference in the number of people who will see a top-ranked comment as compared to the number of people who will see a comment with rank 20 (in part due to the fact that each top-level comment is accompanied by it's top-scoring child comment thread(s)).

The exception to this is a relatively small number of comments that were ejected from the high ranks they were initially observed at and slipped to very low ranks. This suggests that it is possible for `reddit`'s comment voters to collectively 'change their mind' about a comment that was initially popular. However it is much less likely for a comment which appeared at an initially low rank to make its way into high-visibility slots - likely because there are simply not enough comment voting users seeing these comments for their votes to have a noticable impact on rank.

What can these kinds of analysis tell us about the functional utility of `reddit`'s 'Best' comment sorting algorithm? An idealistic outlook on `reddit`'s comment sorting algorithm would be that comments are sorted on quality with the top-ranked comments being the best of those which were submitted. The strong effects of submission order and rank at first observation argue against such a utopian interpretation. If we assume that comment quality and timing are independent, timing seems to be exerting a much stronger effect on final rank than any other variable (including quality).

However, a comparison with some utopian ideal is not necessarily fair and a better comparison might be with more traditional online discussion fora which do not utilise voting systems. In this

light `reddit`'s comment sorting system has a number of features which may be advantageous and which certainly give the comments pages on `reddit` a different feel to a conventional discussion board where contributions are sorted chronologically. It seems fair to assume that in both cases contributions which appear near the top of a discussion will be seen by more people and will serve to set the tone for the discussion. In `reddit`'s case the first 20 comments for a popular post all seem to have a reasonable chance of reaching the highest ranks and setting the tone for the discussion. As the discussion unfolds there is still considerable mobility among the top 10 comments, allowing the tone of the comments page as a whole to shift subtly as the page is exposed to a larger audience and receives an influx of votes. Reddit's voting system also appears to allow voting users to collectively eject comments from the high-visibility area of the page entirely, even when these comments start well and appear in high-visibility locations initially. This is in contrast to a more conventional bulletin board, where the earliest contributions will remain at the top of the thread regardless of their quality - unless they are deleted (with only the submitting user or a board moderator having the power to do so).

Also, Figures 5.8 and 5.9 are geared towards describing average behaviour and obscure the fact that among the 6,652 comments they describe there were 216 top-level comments whose rank decreased by more than 100 places between first and last observations, and 170 top-level comments whose rank increased by more than 100 places. This suggests that when a comment is especially bad or good in the estimation of `reddit`'s users it is possible for its rank to change considerably during its active life-span.

Perhaps the sorting of top-level comments by the 'Best' algorithm *can* be thought of as providing a summary of what `reddit`'s users think about the parent post or its topic. Reddit users who see the post early can submit their comment and have a reasonable chance that (if the post reaches the Front page) their comment will be held aloft in a high-ranking position as emblematic of `reddit`'s collective response to the post. Users who see the post's comments page a little later will be unlikely to see their own comment rise to a prominent location but might still be able to affect the final order in a significant way by up-voting good comments which are at that time further down the order because they do not have many votes. If the post appears on the Front page, comment voting activity will increase sharply, reducing the chance that an individual's vote will have much impact on the final order - but with these votes collectively fine-tuning the top 10 comments threads such that they are ordered to more accurately reflect the opinions of the thousands of users who read the comments and vote at this stage.

While comments become less mobile as a post ages they retain some degree of mobility throughout the active lifespan of their parent post. The 20th user to contribute to the page has a chance that their comment will reach the highest ranks and be seen by tens of thousands of people, the 100th user to contribute can still have a major impact on what the 'Front page' audience will see if the

post reaches that level of success, either by voting or submitting their own comment. The 1000th user to contribute can still vote to make their opinion known (e.g. if they think the 2nd-ranked comment is better they could up-vote it and/or down-vote the top-ranked comment) and if enough users agree the order will change to reflect this.

A discussion board without a voting system cannot accomodate this level of activity. If one is the 100th person to contribute to a discussion post one's contribution will only be seen by the people who are enthusiastic enough about that post to read through the contributions of the 99 previous discussants. Contributions which the community might consider good are mixed with those they would consider poor, and every user must decipher for themselves which is which. The result is that determining what a moderately sized community thinks based on one of these posts is a laborious process that many will not have the time or inclination to engage in. On `reddit`, getting an impression of 'what the community thinks' seems as easy as looking at the top-ranking comments - whether these comments reflect the feelings of `reddit` users *accurately* is however another issue which cannot be addressed with this data. For this, some external measure of 'what the community thinks' about each topic would be required.

5.1.8   Comment threading

The analyses of comments reported thus far have concerned top-level comments exclusively. `Reddit`'s commenting system is such that comments can be threaded, a comment can be a response to a previous comment and when this is the case it will appear indented and below its parent comment. At the highest level all top-level comments are ranked according to the 'Best' algorithm, this process also occurs seperately for all of the 2nd-level comments on each top-level comment, for all of the 3rd-level comments on each 2nd-level comment, etc. By default only child comments which reach a certain score threshold are displayed beheath their parent comment in an order determined by the 'Best' algorithm. Child comments which do not reach this threshold will be hidden unless the reader clicks a 'load more comments' link.

It has been established above that top-level comments for a post can quite quickly become 'saturated' to the point where a new top-level comment which is submitted too late has very little chance of being up-voted to a high-visbility rank. It is still possible for a user who makes a comment at this stage to have their comment displayed prominently if they submit it as a response to a top-level comment with a high rank (or which will subsequently rise to a high rank). Where a user does this with a comment that bears little relation to its parent comment it is known as 'hijacking' the top comment. Sometimes a user who has information relevant to the post, but who has seen the post too late for a new top-level comment to be seen widely, will submit their comment as a response to the comment which is presently ranked top, often beginning their comment with a

statement like "sorry for hijacking the top comment but...". This is one example of `reddit`'s users employing a work-around to circumvent a flaw with `reddit`'s setup (namely that if a top-level comment is submitted too late it will not be widely seen no matter how important it might be). One assumes that where a user attempts this but their contribution is deemed unimportant by other users their comment will quickly be down-voted to the point where it is no longer displayed.

The data on comments which has been collected allows us to address a number of questions about comment levels and threading. The first question we would ask concerns voting rates on the various comment levels through time. As top-level comments are submitted before lower-level comments we would expect that early in a post's active lifespan most comment votes are being cast on top-level comments. As time passes for a post the ranks of its top-level comments become more stable, it is possible that at this stage `reddit`'s comment voting users shift their attention unto lower-level comments whose ranks are more changeable.

5.1.9 The comments page for a high-activity post (redd.it/x8kra)
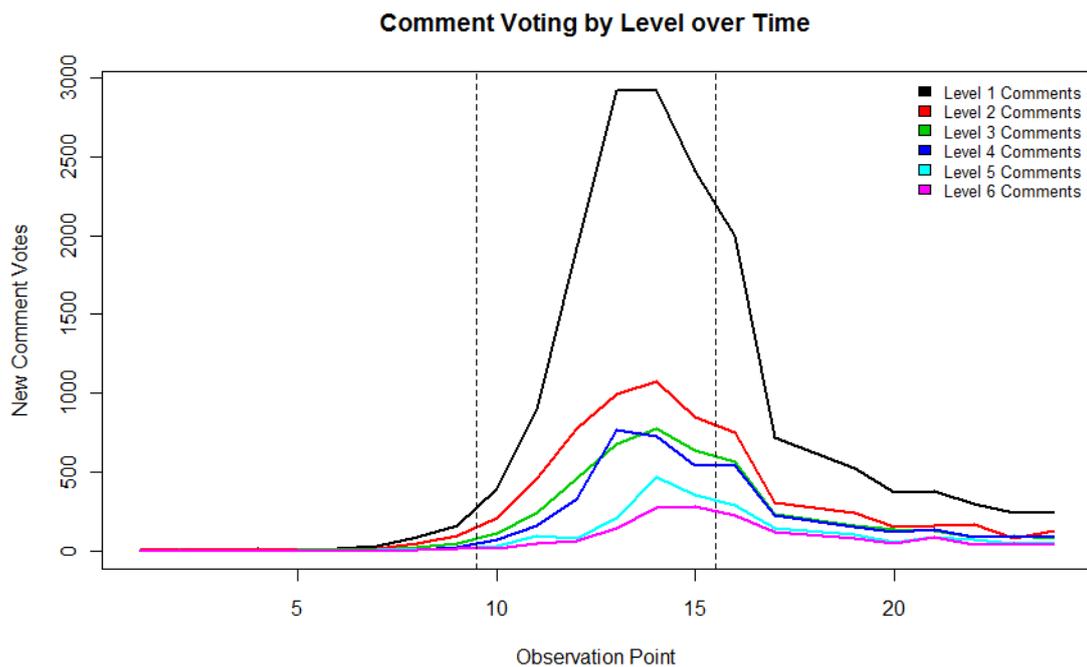
Figure 5.10: Showing new votes by comment level over time for post redd.it/x8kra - dashed vertical lines represent the approximate points at which the post appeared on the Front page for the first and last time. Observations were recorded at 30-minute intervals.

Figure 5.10 shows the number of new comment votes by comment level over time for a single post to the politics subreddit (redd.it/x8kra) concerning the 2012 US presidential election. This post appeared on the default `reddit` Front page and had one of the more active comments pages of all

the posts recorded during the observation time. All observations have been binned according to the length of time between the observation itself and the creation of the post. For each observation point the total number of new comment votes is calculated for comments with level 1-6. For example, time point 1 includes all observations which were recorded within 30 minutes of the post's submission, for this post there were 4 votes observed for top-level comments and 3 for 2nd-level comments (with no votes at any other level) observed within the post's first 30 minutes. The rate of comment voting is quite low for all levels at the start of the post's lifespan. At the 10th observation point (so after 5 hours) this post appeared on the `reddit` Front page (the dashed vertical line on the graph represents the fact that the post appeared on the Front page between observations 9 and 10). Shortly before this (while the post was making its way up the rankings on the 'Main page' for /r/politics) a surge in comment voting began.

The clearest trend in this graph is that higher-level comments tend to receive more votes throughout the lifespan of the post. Top-level comments in particular are voted on much more frequently than comments at any other level - while this is expected it is not inevitable, parent and child comment scores are independent and it is possible for a child comment's score to eclipse that of its parent. There is also a trend whereby lower comment levels begin to receive a surge in votes after their higher-level counterparts. Voting rates for comments at level 3 and 4 follow a similar pattern, and the same can be said for comments at levels 5 and 6. The trends on display in figure 5.10 are on the whole quite unremarkable, and can be accounted for by the fact that lower-level comments do not appear until later in the post's lifespan and when they appear they are presented as subordinate to their parent comment. There is no evidence here that `reddit` users' comment voting behaviour shifts towards lower-level comments as higher-level ranks become more stable. Between observations 15 and 16 the post stopped appearing on the Front page, and shortly afterwards there is a general decline in voting activity at all comment levels.

Figure 5.11 reveals another piece of this picture, showing the number of comments at each level which were displayed on the comments page. A maximum of 500 comments will be displayed for a post upon navigating to its comments page - further comments can be loaded if the user desires but for the purpose of these analyses we only consider comments being displayed on the first page. For this post it appears that all comments were being displayed until observation point 11, at which point the number of total comments exceeded 500 and only the top 500 were shown. There is an interesting trend here whereby once this maximum had been reached top-level comments quickly proceeded to push many lower-level comments out of the top 500 (effect strongest at oservation point 13 with 327 top-level comments being displayed). After this peak for top-level comments they begin to be replaced by lower-level comments until at observation point 16 (once the post stopped appearing on the Front page) the number of comments from each level being displayed becomes quite stable (with 154 top-level comments being displayed).

Number of Comments displayed by Level over Time
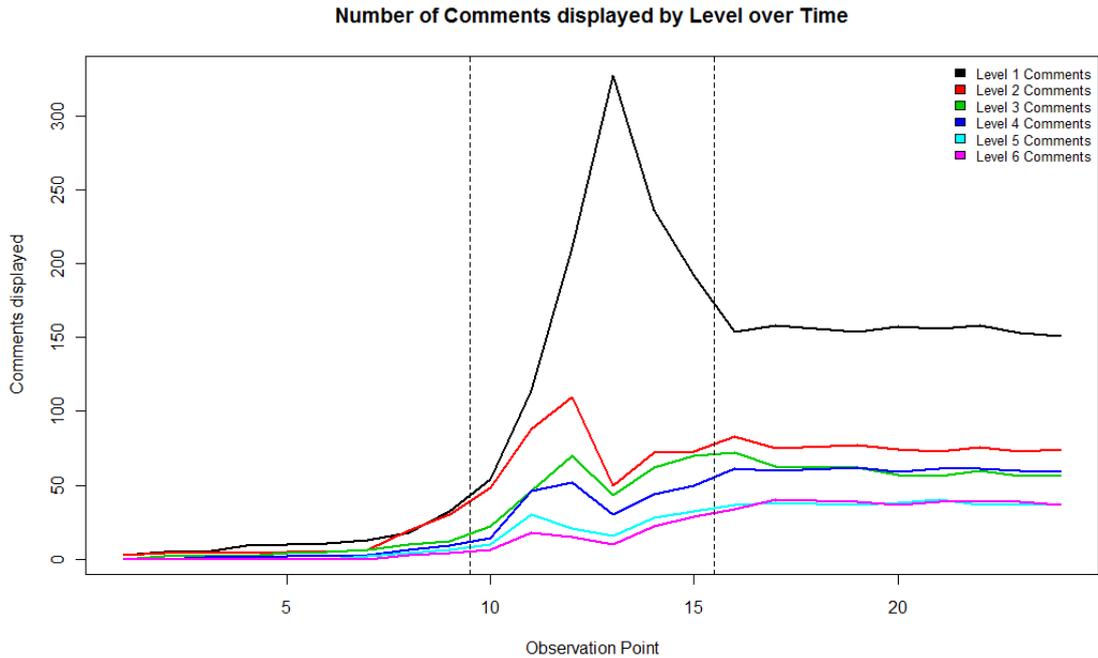
Figure 5.11: Showing number of comments displayed by level over time for post redd.it/x8kra - dashed vertical lines represent the approximate points at which the post appeared on the Front page for the first and last time. Observations were recorded at 30-minute intervals.

The Kernel density plot of first observed rank against last observed rank (Figure 5.9) suggested that comment ranks tend to be quite stable between first and last observation. This is somewhat at odds with figure 5.10 which showed that top-level comments are the most voted on throughout the lifetime of a post. If the ranks of top-level comments are indeed stable then why would thousands of users continue to vote at this level once it had reached stability? To address this issue we will look more closely at the 10 top-level comments for post redd.it/x8kra which finished with ranks 1-10 after 12 hours (24 observations).

Figure 5.12 shows the ranks of these posts at each observation point and it is immediately apparant that these comments did not maintain stable ranks over time. The comment which would ultimately be top-ranked was first observed with rank 9 after the post had been active for 4 hours - although by its second observation this comment had received 17 up-votes to 1 down-vote, propelling it immediately to the top rank where it remained. The comment which was top-ranked before this was one of the earliest comments, this comment slipped gradually into a final rank of 8th. Another early comment started with rank 4 and finished with rank 2. The comment which would finish with rank 3 was first observed after 6 hours of the post's lifespan with rank 17 - this comment was submitted after the post was already on **reddit**'s Front page but still managed to reach the top 3 ranks. These observations fit with the idea that once a post has hit **reddit**'s Front page much of the comment voting activity serves to 'fine-tune' the ranks of comments which are already in the

Final Top 10 Comments for Post redd.it/x8kra
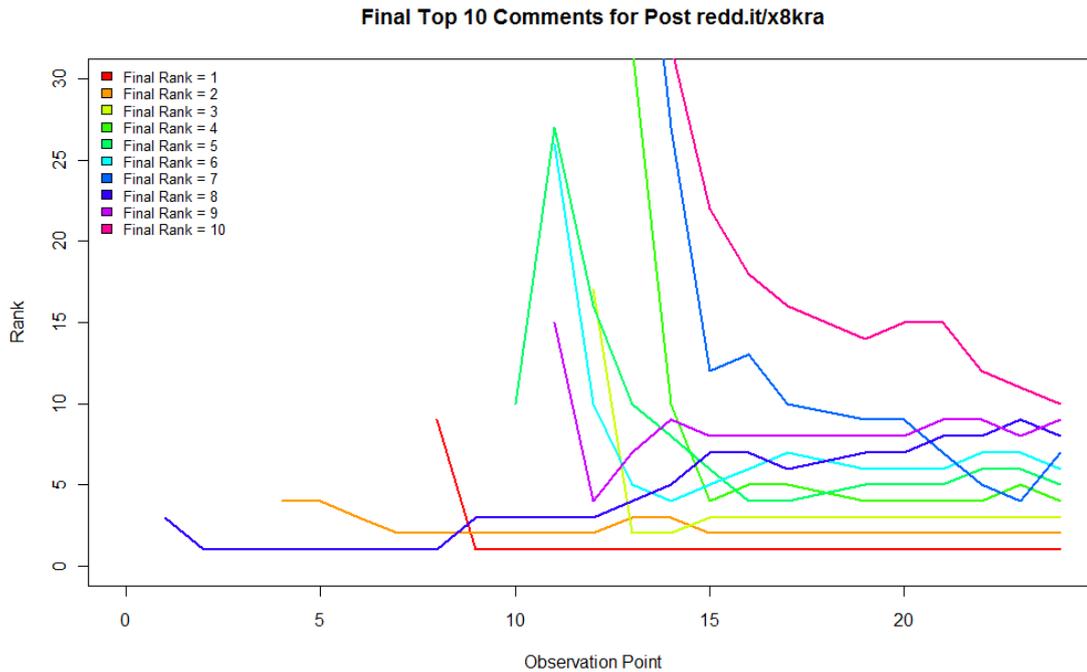
Figure 5.12: Showing the ranks over time for the 10 top-level comments which concluded with the highest ranks on post redd.it/x8kra. Illustrates that the ranks of these comments fluctuate initially but become more stable towards the end of the observation period for this post. Observations were recorded at 30-minute intervals.

top 20 positions.

However, Figure 5.12 suggests that 'fine-tuning' is not the only result of the collective comment voting behaviour of `reddit`'s users. There are three comments here which finished with a rank inside the top 10 but which were initially observed quite late in the post's lifespan and with ranks outside the top 50. The highest-placed of these comments finished with rank 4 after being initially observed at rank 49 and then slipping further to rank 72 - at this point the comment received a run of 8 up-votes without any down-votes, propelling it back up to rank 10, and from there it quickly improved its rank to 4th. Comparing figure 5.12 with figure 5.10 suggests that during the peak of top-level comment voting (observations 13-15 in particular) several low-ranking comments were propelled into the top 10.

5.1.10 Comments which criticise their parent Post

There is also some insight to be gained by looking at the content of these comments. The post itself (redd.it/x8kra) is titled "Does anyone else want to see Obama win for no other reason than to watch conservatives lose their shit?" and was submitted on July 27th 2012. Of the 10 top-level comments which finished with top ranks none echoed the sentiment of the post. The comment

which finished with the top rank expressed a more constructive opinion about wanting to see the Republican party "get the crazies back on the short leash". The post which finished 2nd was humerous in nature ("In Canada, we'll be entertained regardless of who wins."). Most of the remaining top 10 comments are highly critical of the post's subject. One of the most scathing criticisms comes from the comment which ended with rank 4, "This post is what is wrong with the current political system. This 'us vs. them' mentality does nothing progressive for our country. If you don't have any sound reasons why you would want one candidate to win over the other, then you have no business voting this November.". In fact all of the posts which were up-voted into the top 10 during the surge of activity between observations 13 and 16 are highly critical of the post's subject.

This is important because it may yield insight into an unusual trend on `reddit`. When looking at the comments page for a post which is displayed on `reddit`'s Front page it is not uncommon for the top comments to be highly critical of the post and/or the fact that it is positioned on `reddit`'s Front page. This has been observed most often on subreddits like politics, worldnews and science - where the most frequent criticism is that the post has a sensationalist title or links to an article which provides very little evidence for its claims. At first glance this is difficult to comprehend, why would the same group of users up-vote a post to the Front page and also up-vote comments which decry said post?

There are two possibilities relating to differentitation among `reddit`'s users which could account for this. Firstly, it is possible that users tend to 'specialise' in either voting on posts, commenting, or voting on comments - this would mean that the group of users who voted the post up to the Front page are to some degree distinct from the group of users who commented on it and/or voted on these comments. A second possibility relates to a differentiation among users based on *when* they vote - with the users who browse, vote and comment on the New and Rising pages being to some degree distinct from the users whose activity concerns posts which are already on the Front page. Chapter 6 will look for signs of differentiation among `reddit`'s users.

A third possibility concerns `reddit`'s structure and algorithms - we have seen in section 5.1.2 above that `reddit`'s 'Hot' ranking algorithm weights early votes much more heavily than later votes and that it uses the post's score as opposed to a ratio of up-votes to down-votes. This raises the possibility that the decision as to which posts appear on the Front page is actually quite a 'shallow' one; if a new post happens to receive some up-votes quickly it stands a good chance of reaching the Front page and can actually do so with a relatively low total score. It is also possible to vote on posts quickly based on their title and without looking at the resource the post links to. If there are a large number of users voting quickly on recent post submissions this could account for 'poor quality' posts appearing on the Front page - and because the voting rate on the Front page is so much higher than 'precursor' pages any post which appears there is likely to receive

enough votes to keep it there for some time.

The data displayed in figure 5.12 do not allow us to differentiate between these possibilities - but they do seem to fit quite neatly with the 2nd and 3rd possibilities. A simple way of describing this process would be that the decisions about which posts appear on the Front page are relatively shallow and susceptible to random timing-related effects - but that once a post appears on the Front page and it's comments page has received thousands of votes ('maturing' to a point of relative stability) the ranking of top comments can reflect the more considered opinions of `reddit`'s user-base as a whole (or at least the segment who are active on that subreddit). However, when this 'considered opinion' involves the post itself being poor it appears that the users who have decided this do not have the power to quickly down-vote the post off the Front page. This could be due to two factors - the post already acquired a high enough score to keep it on the Front page during the hours before the critical comments rose to the top of the page, and/or, there are still more 'shallow' voters on the Front page making their decisions to up or down-vote quickly based on the post's title, than there are users who will read the post and its comments before making their decision.
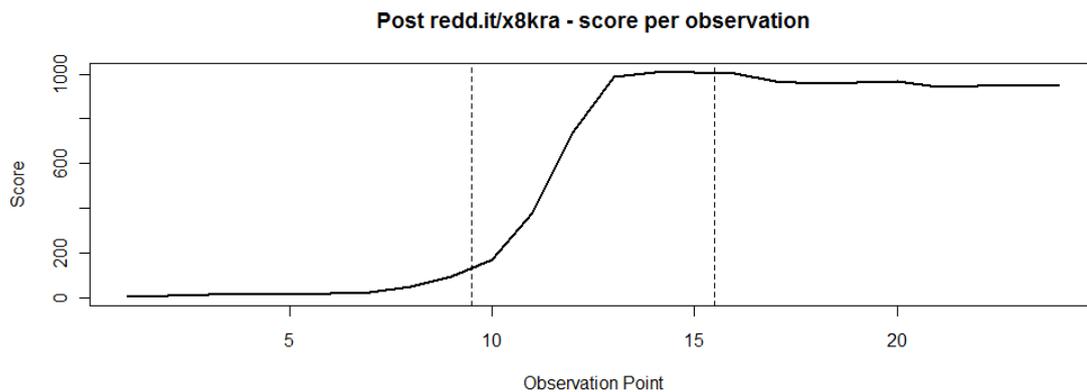
Figure 5.13: Showing the score observed on reddit for the post redd.it/x8kra at each observation point (30-minute intervals between these). Dashed vertical lines represent the points where the post appeared on the Front page for the first and last times.

Returning specifically to the post redd.it/x8kra we can consider the post's score and look for an effect at around the time comments which criticised the post began rising to the top of its comments page. Figure 5.13 shows the post's score over time, as expected there is a rapid increase in score at around the time when the post first appeared on the Front page. The post's score then appears to stagnate at around observation 13, so at the same time as the most critical comments began to appear in high-visibility locations. The post was still receiving votes at a high rate but from this point roughly 50% or slightly more were down-votes - this did not push the post from the Front page immediately but almost certainly curtailed the length of time it would be displayed there. It

is however not possible to determine whether the display of critical comments caused the change in voting response or whether these two effects share a common cause.

The issue of top-ranking comments which criticise their parent post will be considered further in Chapter 7.

5.1.11   The comments page for a moderately active post (redd.it/x6pb4)

The level of comment mobility on display for the above post (redd.it/x8kra) is higher than expected based on the kernel density plots of first against last observed rank shown in Figure 5.9. This is possibly due to the fact that this post appeared on the Front page and had one of the most active comments pages of all posts considered from this time period. It is therefore prudent to also look at a post with a lower level of activity. The threshold for a post to be included in the kernel density plots was that it should have at least 50 top-level comments, this section considers a post with 50 top-level comments (redd.it/x6pb4).
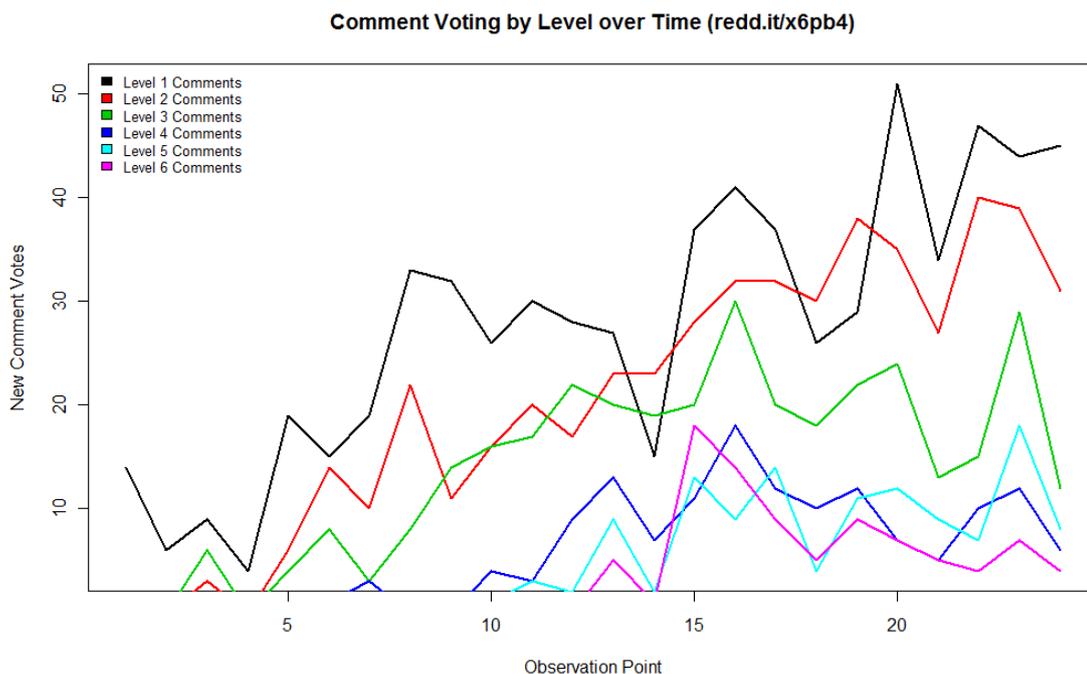
Figure 5.14: Showing new votes by comment level over time for post redd.it/x6pb4. Comment voting is more evenly distributed between comments of different levels in comparison to a highly active post. It should also be noted that the level of comment voting on this post is in general much lower. Observations were recorded at 30-minute intervals.

The most prominent difference between this post and the post which appeared on the Front page is that the number of votes cast on comments for this post is orders of magnitude lower than for the post which appeared on the Front page (Figure 5.15). This post did not experience the surge

Final Top 10 Comments for Post redd.it/x8kra
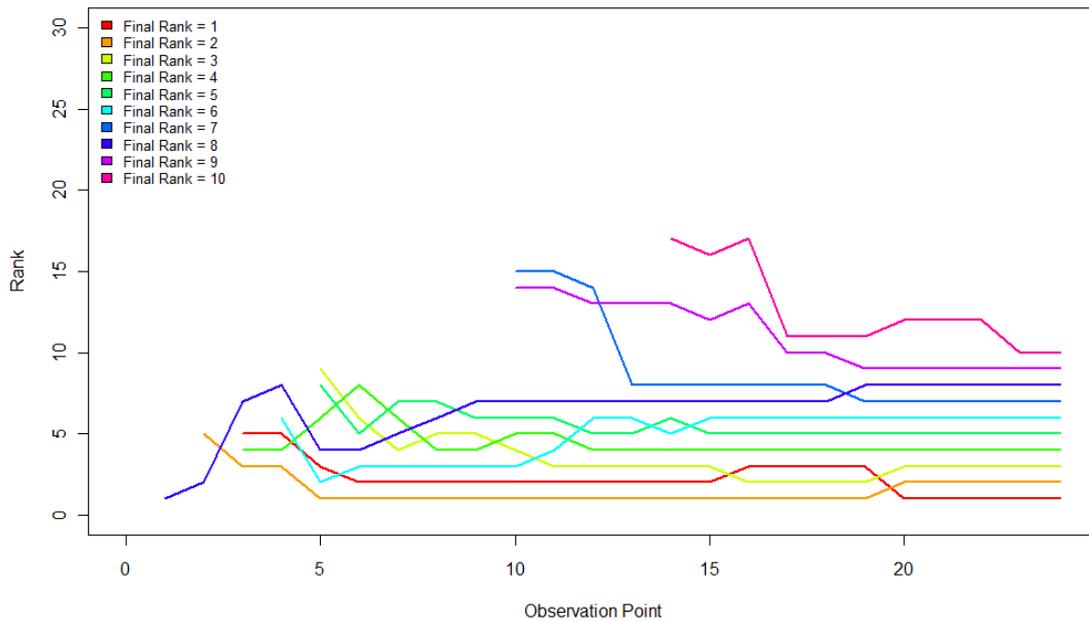
Figure 5.15: Showing the ranks over time for the 10 top-level comments which concluded with the highest ranks on post redd.it/x6pb4. The ranks of these comments show greater stability throughout the lifespan of this post, when compared to the ranks of comments on a highly active post. Observations were recorded at 30-minute intervals.

in comment voting that began to occur on the other post just before it reached the Front page - here a maximum of 50 votes were cast on top-level comments in a half-hour period, for the post which appeared on the Front page this maximum was approaching 3,000.

Figure 5.15 shows the ranks through time for the top-level comments which finished in the top 10. There is a much lower degree of comment mobility on show here than was shown for the post which appeared on the Front page. Only three comments which were first observed outside the top 10 made their way into the top 10 and these ultimately occupied ranks 7, 9 and 10 with none of these being observed with a rank outside the top 20 at any stage. This level of mobility is more in line with what would be expected based on the kernel density plots, and it seems likely that these kernel density plots are more representitive of the type of moderately active posts which made up the majority of the sample. There appears to be a fundamental difference between how the comments page for a moderately active post develops as compared to the comments page for a highly active (Front page) post.

5.1.12 Using the comment voting system to conduct 'Ask Me Anything' interviews

The /r/IAmA subreddit is host to an unusual form of interaction which relies on `reddit`'s voting system. The principle of an 'Ask Me Anything' (AMA) interview is that the individual to be interviewed (usually a celebrity or someone with specified life experience, on one occasion the President of the United States) creates a post in which they identify and sometimes describe themselves. Other `reddit` users then submit questions for this individual as comments on their post, users vote on these comments with the idea being that the interviewee answers the most popular questions (as rated through the voting system). Answers are submitted (generally) as 2nd-level comments which are replies to the top-level comments asking the question being answered - sometimes this expands to a back-and-forth discussion between the interviewee and `reddit` users through lower-level comments.

The data from /r/IAmA being analysed here was collected in October 2012 and by this stage there is some relevant background information and jargon on /r/IAmA which should be discussed first. As elsewhere on `reddit` (and some other websites) the user who created the post (in the case of IAmA the interviewee) is often referred to as the 'OP' (Original Poster) and this term will be used here. Posts to the /r/IAmA subreddit are subjected to a verification procedure, in cases where the OP is a celebrity this is usually handled by their announcing the IAmA post through a service where their identity has already been established (e.g. `Twitter`). In cases where the OP wishes to remain anonymous (where the subject of the AMA is a life experience, sometimes sensitive or embarassing) the subreddit's moderators will verify that the post is genuine privately and announce in a comment that the post has been verified. Verification procedures were adopted because previously the veracity of IAmA posts was often doubted and discussed by users.

Comments submitted by the OP have a special marker across all of `reddit` now (where their name appears above the comment it has a blue background), this is particularly useful for /r/IAmA posts. There is also at this stage a schedule of upcoming IAmA posts by recognisable individuals displayed in the subreddit's side-bar. This schedule was likely introduced because `reddit`'s voting system could not be relied upon to identify all of the AMA posts which users would want to see - without a schedule there would be a random element to whether the post progressed from the New to the Rising page as it may only be voted on by a small number of users on the New page. The presence of a schedule means that individuals who are interested in the AMA can make a point of checking the New page at the appropriate time and up-voting the relevant post.

IAmA posts represent an interesting form of interaction, but there is a further reason why they have been identified for study. There is nothing unusual about the structure of the /r/IAmA subreddit, the voting system and its algorithms are standard site-wide. For this form of interaction

to work *Reddit's voting users must collectively behave in certain ways.* The /r/IAmA subreddit therefore offers an opportunity to explore the interaction between `reddit`'s social and technical components. If the activity on the comments pages for /r/IAmA posts is different to the activity on other subreddits' comments pages - this difference can only be the result of a difference in the collective behaviour of `reddit`'s users. This would show that `reddit`'s population of users can use the same site-wide voting system to perform a different kind of task to that which the system was designed for - based only on a shared understanding that this particular subreddit serves a specific purpose.

There are a number of trends which would be expected because of the nature of IAmA posts, and for some of these it is reletaivly straightforward to check whether they are occuring.

Firstly, where the OP has replied to a question with a 2nd-level comment we would expect this comment to be the top-scoring 2nd-level comment for that parent comment. The purpose of an AMA post is to see the OP's answers to questions, where the OP answers a question this answer should be made as visible as possible and should take priority over the comments of other users. `Reddit`'s users must vote in a certain way for this to happen.

Secondly, the OP is supposed to answer top-scoring questions - it is the voting on questions which sets IAmA posts aside from other instances where a celebrity answers questions submitted by readers or viewers (in many cases one suspects these questions are screened and cherry-picked by an intermediary party). On /r/IAmA the principle is that users collectively dictate which questions are being asked as part of the 'interview'. This is something which can be investigated with the available data.

Thirdly, if there are cases where the OP answers a question which was at that stage *not* one of the top-scoring comments - then it is hypothesised that the top-level comment which was replied to will subsequently be up-voted to a more prominent rank - because the purpose of IAmA posts is to see the OP's responses to questions, if questions with low ranks are answered the answers will not be widely seen unless the questions themselves become more visible (i.e. more highly ranked).

Data analysed here was collected between 9th and 24th October 2012 on the /r/IAmA subreddit - 2,414 different posts were observed on the /r/IAmA New page in this time period and followed until they became inactive - with comments and their scores being recorded at 30-minute intervals. Of these, 1,146 posts were 'AMA Request' posts - through these posts users request that certain people or people with certain experience participate in an AMA - comments from these posts will not be considered.

Front page IAmA posts

In looking for the above-predicted trends we will first consider posts which appeared on the Front page - these posts will have the highest levels of comment voting activity so this seems the most likely place to observe the above-hypothesised trends related to voting. In the data there are a total of 43 IAmA posts which were not requests and which appeared on the Front page, one of these has been excluded because it was already underway when the data collection window opened. For the remaining 42 posts there are 32,034 individual comments represented in the data - there were likely more in total but at any given observation point the data collection script will obtain data for only the 500 comments which are displayed by default at that time, and the data collection script only collected comments at up to level 6. Of these, 3,688 were comments submitted by the OP for their post (i.e. the interviewee).
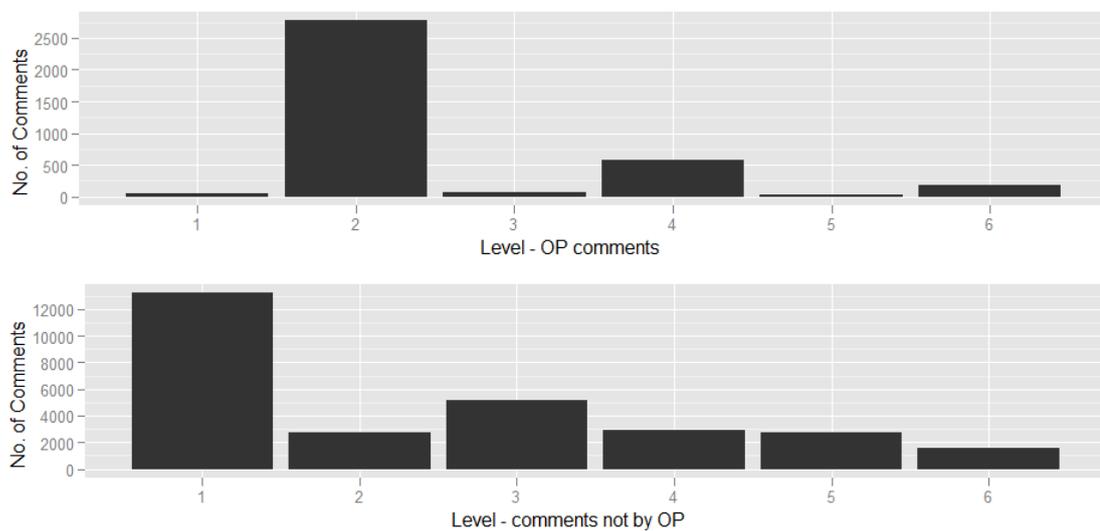
Figure 5.16: Showing the level of comments submitted to Front page IAmA posts - for comments by the OP and by other users.

Figure 5.16 shows that the majority of OP comments occurred at level 2, as expected. There are also a substantial number of OP comments at level 4 and to a lesser degree at level 6 - these are most likely instances where the OP has become involved in a back-and-forth discussion with another user or users. Looking at the comments submitted by other users, there is a clear drop at level 2 suggesting that users are less likely to post comments at this level than they would be on other subreddits - perhaps because they regard replying to level 1 comments as the OP's domain. There are however still roughly as many comments from non-OP users as there are from OPs at level 2 (around 2,500).

One of the hypotheses relates to this. Where there are comments from non-OP users at level 2 one would expect the OP comments to be ranked most highly. Figure 5.17 suggests that this is almost
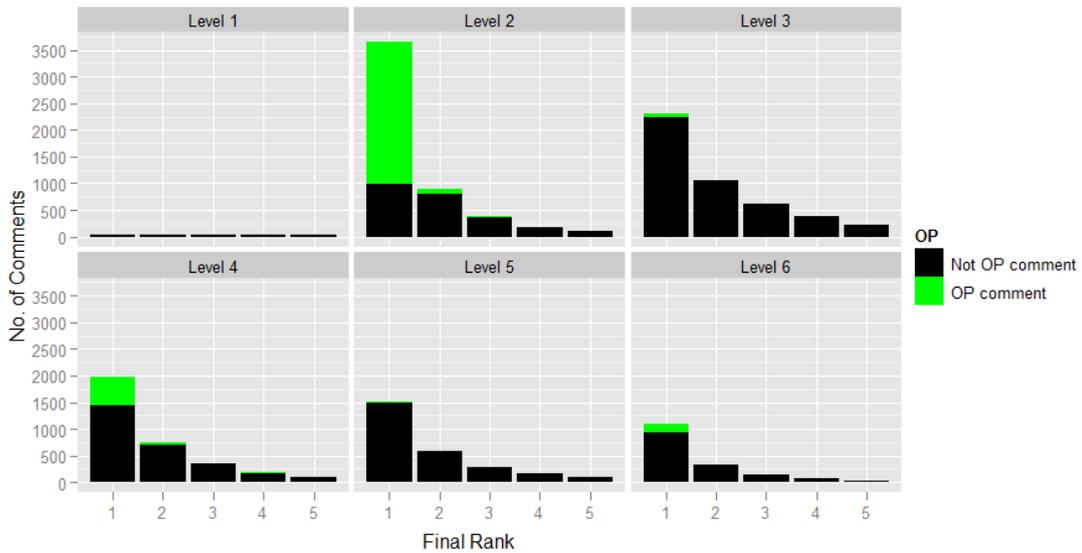
Figure 5.17: Showing the final ranks of comments submitted to Front page IAmA posts by level. Comments with final rank greater than 5 have been excluded, N = 17,736. The bars at level one are all equal to 42 because a given post can only have one comment at each rank, whereas at level 2 there are as many comments at rank 1 as there are level 1 comments which have been replied to.

always the case. There were 2,659 OP comments at level 2 and 96% of the time these were ranked number 1 at final observation (i.e. 2553 of these comments were the top-ranked reply to their parent comment). 90% of the 585 OP comments at level 4 and 93% of the 180 OP comments at level 6 finished with rank 1. In the great majority of cases `reddit`'s voting users have collectively behaved in the way which was expected (and required for IAmA to function) - up-voting OP comments so that they appear above the comments of others. An attempt was made to contrast these findings with observations on another subreddit (/r/politics) but there were too few comments from OPs to make a meaningful comparison.
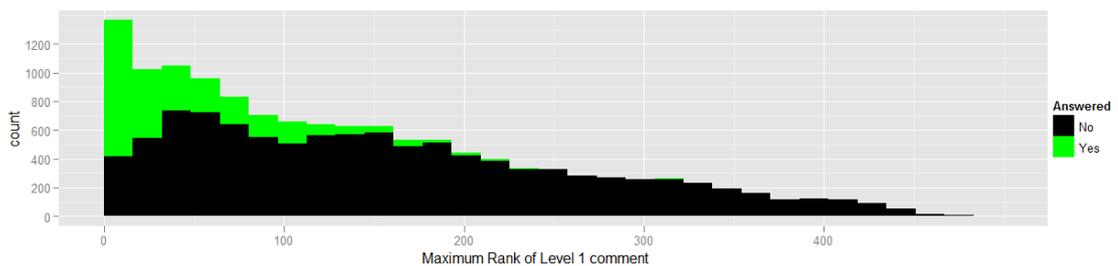
Figure 5.18: Showing the maximum ranks of level 1 comments and whether they received a response from the OP

Do the interviewees also play along with the rules of IAmA, answering those questions with the highest ranks? Judgments about whether a response to a question constitutes an answer (or a dodging of the question) would add significantly to the complexity of answering this question -
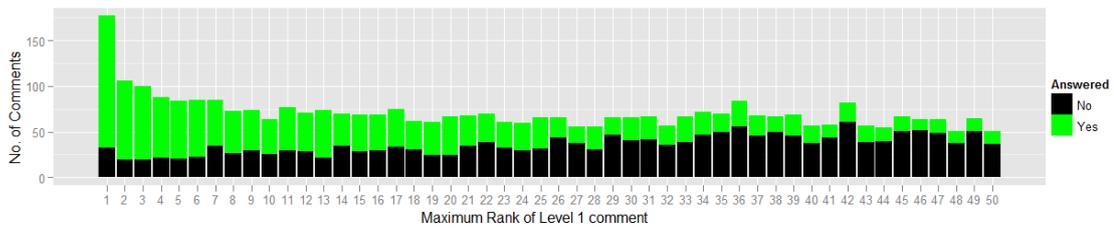
Figure 5.19: Showing the maximum ranks of level 1 comments which appeared at some stage in the top 50 comments - and whether they received a response from the OP

instead we will only consider the presence or absence of a response from the OP, where a response is present this will be assumed to be an 'answer'. There is however another unavoidable source of complexity - the ranks of level 1 comments can change considerably over time (e.g. if the OP doesn't want to answer the top-ranked question but answers many others - the question(s) that they skipped may suffer a decrease in rank). One approach is to consider this in terms of the maximum rank at which a level 1 comment (i.e. a question) appeared.

Figure 5.18 suggests a relationship between the maximum rank a level 1 comment appeared at and whether it would elicit a response from the OP (but does not rule out the possibility that comments rose to those ranks after they had received a response). The number of low-ranking comments which received a response is perhaps surprising, and does suggest a degree of cherry-picking of the questions they would answer by some OPs. Figure 5.19 shows the same information for comments which appeared within the top 50 ranks only, making the relationship between maximal rank and whether the comment received a response more clear. 82% of comments which were ranked 1st at any observation point have received a response from the OP, 74% of comments which appeared in the top 10 ranks have received a response from the OP dropping to 60% for comments which reached ranks 11-20 and 43% for comments which reached ranks 21-30. This suggests that the OP most often answers high-ranking questions but will sometimes ignore prominent questions.
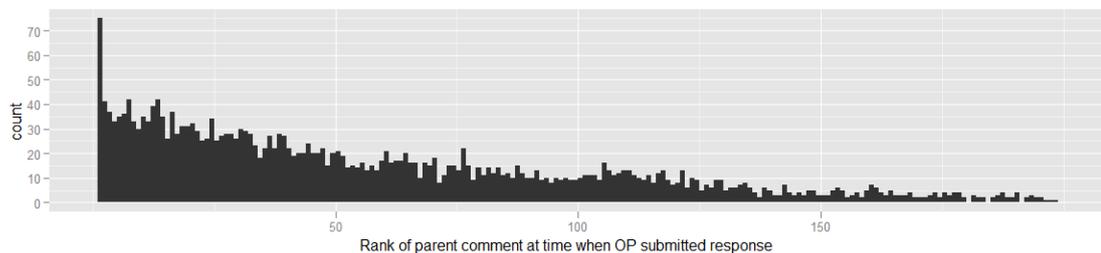
Figure 5.20: Showing the ranks of level 1 comments at the observation point before they had a response from the OP

It is possible that the ranks of top-level comments change dependent on whether the OP responds to them - in the above figures it may be the case that the OP actually answered whatever questions

166

they pleased and that those questions which they answered subsequently rose to high ranks. To investigate this we will consider the ranks of parent comments at the observation point before they received an answer (or at the same observation point if the question was not previously observed). Figure 5.20 suggests that top-level comments do tend to gain higher ranks once they have received a response from the OP - the distribution is not as highly skewed towards the top ranks as it was when we considered maximum rank in figure 5.19, suggesting that top-level comments do move up in rank when they receive a response from the OP.

The question of whether top-level comments receive a boost to their score after the OP has replied to them can be pursued through modelling of score change per observation - with a little augmentation of the data. First a score change per observation variable was created in a similar manner to the variables which were described for *posts* above. Then an 'observation number' variable was added - for each comment this number starts at 1 for their first observation and at each subsequent observation point is increased by 1. Also an 'observations since OP response' variable was created for level 1 comments - this was constructed in a similar manner to the observation number variable but the value remained at 0 until the OP responded to the comment in question, whence it began to increase by 1 at each subsequent observation point. The creation of this variable was computationally intensive (because for each comment several `MySQL` SELECT queries on very large data tables were required) and therefore it has only been created for comments on the first six IAmA posts in the set.

We will only consider the first five observation time points of level 1 comments on these posts here, leaving a total of 5,974 observations (which come from 1,198 comments) to analyse after outliers (a score change of lower than -50 or greater than +50) have been removed. The score change variable is approximately normal (with a mean of 0.9) when plotted on a histogram and so a linear model for the normal distribution was chosen. While we are interested here in an effect related to the receipt of a response from the OP, there will be other variables which likely have stronger relationship with score-change (the rank a comment appeared at and how long it had been displayed for), these will be controlled for by including them in the model.

A comment's rank at the observation point has been converted to a binned ordinal variable, chiefly because one suspects that below a certain rank (e.g. rank 40) there is very little activity but the rank variable can run into the hundreds (one suspects that a top-level comment with rank 100 sees a similar level of activity to one with rank 300). Ranks have been placed in bins of five (i.e. ranks 1-5, 6-10, etc.) up to rank 40 (beyond this rank the effect becomes non-significant if further categories are added). Observation number and the number of observations since a repsonse from the OP was received have been included as five-level categorical variables.

Table 5.9 shows details of this model. Unsurprisingly, the strongest effects are associated with a comment's rank, with the score of comments at ranks 1-5 increasing by around 4.5 more per

Variable	Coefficient	Std Error	p value
Intercept	0.684	0.092	< 0.001
Rank 1-5	4.529	0.296	< 0.001
Rank 6-10	2.680	0.277	< 0.001
Rank 11-15	1.400	0.267	< 0.001
Rank 16-20	2.965	0.263	< 0.001
Rank 21-25	2.350	0.258	< 0.001
Rank 26-30	1.577	0.269	< 0.001
Rank 31-35	1.731	0.275	< 0.001
Rank 36-40	0.901	0.271	< 0.001
Rank 40+	Intercept		
Observation 1	Intercept		
Observation 2	-0.948	0.130	< 0.001
Observation 3	-0.532	0.135	< 0.001
Observation 4	-0.415	0.139	< 0.01
Observation 5	-0.296	0.143	< 0.05
Before OP reply	Intercept		
Observations since OP reply - 1	1.094	0.208	< 0.001
Observations since OP reply - 2	1.393	0.221	< 0.001
Observations since OP reply - 3	1.489	0.236	< 0.001
Observations since OP reply - 4	0.248	0.272	0.362
Observations since OP reply - 5	-0.493	0.37	0.183

Table 5.9: Showing parameters for a linear model on score-change with categorical explanatory variables representing its rank, observation number, and number of observations since the OP has responded to the comment (0 when OP has not responded). The reference group are comments with rank greater than 40, observation 1 and no reply from OP. N = 5,974 observations of level 1 comments. Model accounts for 13% of variance in score change.

observation than the scores of comments ranked 41+ (the reference category). The effects of the observation number variables suggest that comments' scores tend to increase quickly at first but then this rate slows at observation 2 before increasing again at subsequent observations. The model shows a significant effect for some levels of the 'Observations since OP reply' variable, suggesting that once the OP has replied to a level one comment its score will increase by on average one point more at the next observation (as compared to comments which did not receive an OP response), rising to 1.5 points more at third observation after OP response. The effect is non-significant at the fourth and fifth observation after OP response (and at the fifth observation it actually seems to be reversing) but these coefficients are based on relatively few data points (i.e. there were only 95

comments whose fifth observation after OP response is included in the data, these are cases where the level 1 comment and OP response were first observed at the same time, by contrast there were 289 data points representing a first observation after OP response).

For the 42 IAmA posts which appeared on the `reddit` Front page there is evidence which supports our three hypotheses. When the OP replies to a question their comment is almost always the highest ranked of all replies to that question. The OPs seem to give priority to the highest ranked questions when they submit responses - but it seems that at least some of the OPs will choose to ignore specific questions even if they appear in a prominent location. There is also evidence that once the OP has responded to a level 1 comment that comment will see a boost to its score, although this effect may be short-lived (disappearing after 2 hours).

Main page IAmA posts

We have seen that IAmA posts which appeared on the `reddit` Front page conform to several expectations about how their comments should behave. In the above sections which dealt with comments on other subreddits there is a suggestion that comments pages benefit from a high level of activity - with a higher number of comment votes serving to fine-tune the ranks of comments. It is therefore of interest to ask whether IAmA posts which did not appear on the Front page (and which will therefore have a generally lower level of comment voting) still conform to these expectations. This section will quickly re-visit the analyses performed above on Front page IAmA posts, applying them to a sample of IAmA posts which appeared on the subreddit's Main page (on any of the four leafs) but not the `reddit` Front page.

The data considered here are 1,503,379 observations of 38,821 comments on 663 IAmA posts - observed between 9th October 2012 and 20th October 2012. Of the 38,821 comments, 12,875 were submitted by the various OPs - so it is immediately clear that OPs accounted for a much greater percentage of comments (33%) in this sample than for the Front page posts (11.5%). Figures 5.21, 5.22 and 5.23 suggest that the same trends are present and perhaps even stronger than in the Front page data.
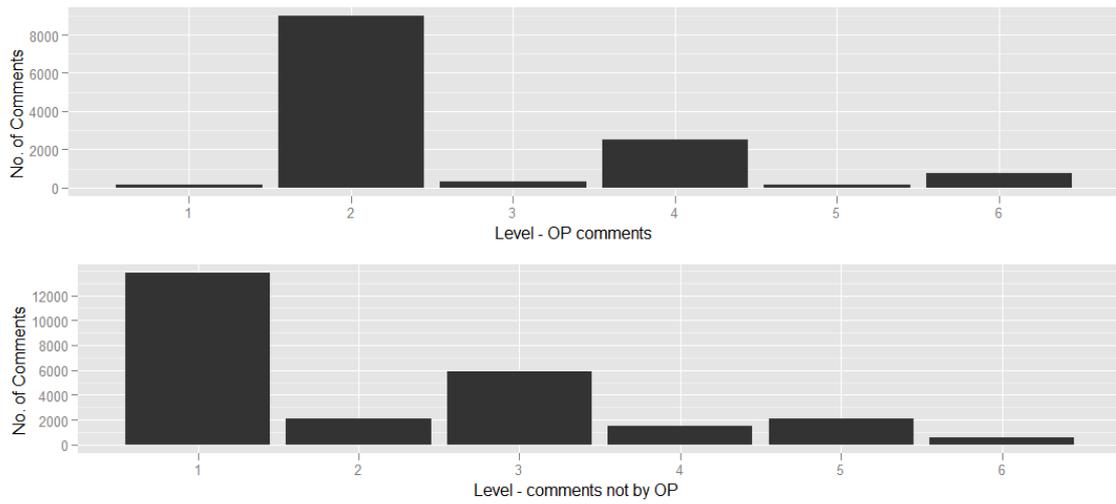
Figure 5.21: Showing the level of comments submitted to Main page IAmA posts - for comments by the OP and by other users.
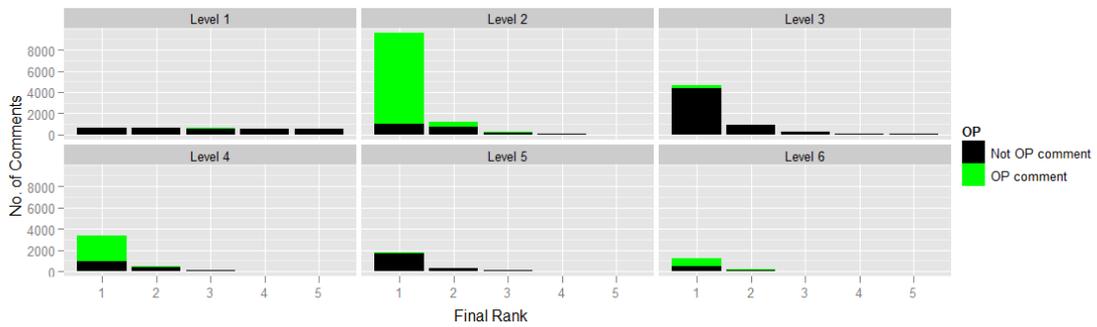
Figure 5.22: Showing the final ranks of comments submitted to Main page IAmA posts by level. Comments with final rank greater than 5 have been excluded, N = 12,875.
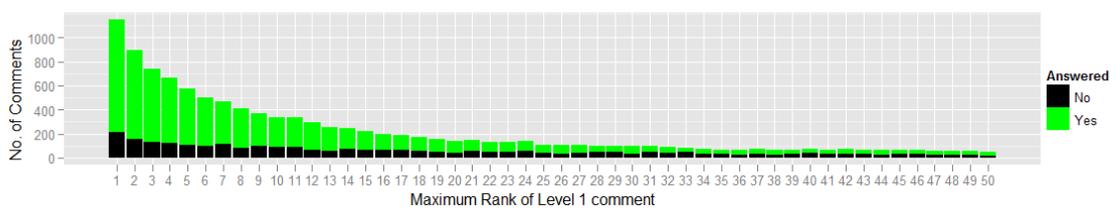

Figure 5.23: Showing the maximum ranks of level 1 comments which appeared at some stage in the top 50 comments - and whether they received a response from the OP

5.1.13 Comments - a short summary

The analysis of comments in this chapter has shed light on how the voting system works with respect to comments. We have seen that comments have considerable mobility, particularly those comments relating to a post which appears on the Front page. There is however also a strong effect of submission time, and on a moderately active comments page the comments will tend not to move very far from their initial rank. `Reddit`'s comments pages, based on this data, certainly seem capable of representing roughly 'what `reddit`'s users think' about the post's subject at a glance. There is evidence which hints that a 'mature' comments page will give a much more considered opinion than the score for a post (which on the whole looks like quite a shallow measure).

However, low levels of attention and voting for comments which are submitted later in a post's lifespan indicate that this system is far from 'perfect'. It is not clear what standard `reddit`'s comment voting system should be judged against, I do not know of any non-DM systems which even attempt to allow thousands of individuals to participate in a unified discussion of a single topic. The strength of `reddit`'s comment voting system is that the discussion can be re-shaped by a large number of voting users such that it more accurately reflects the feelings of the community. This ability to re-order the components of the discussion means that the capacity to have a completely unified discussion (where *all* comments are part of the same conversation) is lost. Instead, each thread (all of the lower-level comments which relate to a single top-level comment) stands as a seperate conversation, and at the highest level it is these threads which are being ranked and sorted. Conventional discussion fora would seem more suited to facilitating unified discussions on a topic - but when a thread on a conventional discussion forum reaches hundreds of contributions one suspects that this too will contain sub-sets of posts where groups of users have pursued their own tangent. Here the difference would be that on a conventional forum these 'sub-discussions' are scattered throughout the larger thread, and that there is no mechanism whereby the community can determine that a 'sub-discussion' is of poor quality and exclude it from being displayed prominently.

On `reddit` there is an advantage to entering the discussion early in that one's comment(s) have a better chance of being displayed prominently and 'setting the agenda' of the discussion. On conventional discussion fora without DM this advantage is absolute, early contributions *will* be displayed more prominently because contributions are simply displayed in the order they were submitted.

Understanding the role of comments pages on `reddit` will require much more consideration of the qualities of comments. We have seen that `reddit`'s users can collectively lift a comment from obscurity and raise it to prominence. If we are to relate this to the social implications of Distributed Moderation more generally, we must look at the *types* of comments which `reddit` users raise to prominence and what the outcomes of making these comments visible might be. This will be

considered further in the context of the 'SOPA' case study presented in Chapter 7.

The analysis of IAmA posts allowed us to explore the ineraction of `reddit`'s social and technical components. The activity on IAmA comment pages was found to meet all of the expectations which are specific to the subreddit. As the technical component is held constant across all subreddits, these distinct patterns of behaviour can only be the result of different social norms about how one votes on this subreddit. `Reddit`'s users can collectively use the website's software infrastructure for this purpose that it was not specifically designed for. There are a number of other subreddits which one suspects have their own sets of voting norms (e.g. /r/AskReddit), and `reddit`'s users regularly employ the voting system for other purposes (e.g. conducting polls) on specific posts, but these situations do not generate such easily testable hypotheses.

5.1.14 Re-visiting `reddit` Research Questions

We began by considering whether the distribution of voting activity between items of content was skewed and this was found to be the case. The distribution of voting activity between posts in March 2009 was found to be roughly approximated by the power law, and actually appears to be more skewed than a typical power law. An often referenced aspect of the power law is that 20% of cases account for 80% of the quantity being distributed, in `reddit`'s case 7.8% of posts accounted for 80% of votes.

This was seen as a central aspect of how `reddit` functions, by focusing the attention of its users on a small sub-set of submitted posts these users can have shared understanding of what's happening on `reddit`. If `reddit` was a newspaper the Front page would be equivalent to the articles which made it into the daily edition, item's which did not appear on the Front page did not make the cut. This widespread knowledge of what's on `reddit`'s Front page is likely implicated in the sense of community which (at the time when this data was collected) seemed stronger than one would expect of a website with such a large user-base. The focusing of user attention on the Front page is also heavily implicated in `reddit`'s capacity to serve as a vehicle for social endeavours (explored in Chapter 7).

Analysing the distribution of votes between posts also uncovered a potential weakness in `reddit`'s voting system - a large number of posts (particularly in more active subreddits) were not voted on by any users. If some proportion of submitted items are being ignored at random then this suggests a weakness in `reddit`'s ability to place the *best* items in the highest-visibility locations. However, in 'front-end' data collected more recently there are far fewer posts which received no votes. A number of factors may be in play here. Back-end data excluded votes which had been nullified by `reddit`'s anti-cheating code whereas the vote counts collected through the API included 'spam' votes which were not actually being used to calculate the item's score (in `reddit`'s terminology the

scores are 'fuzzed'). It is possible that some of the posts which appear (through front-end data) to have been voted on actually received no legitimate votes (and therefore if back-end data for this period was available they would appear to not have been voted on).

There is also the ever-present fact that `reddit` is and has been expanding and evolving rapidly. The back-end procedural data which the analyses in section 5.1.1 were based on is the oldest data considered by the present research, and was collected more than three years before some of the front-end data considered in this chapter. During this time `reddit` has seen a strong and sustained surge in its usage statistics, meaning that there are more users to cast votes (but also more posts to vote on). There have also been direct interventions from the moderators of some subreddits (/r/pics, /r/videos), placing messages at the top of their pages which asked users to 'help out' by voting on the 'New Queue' - these are considered in section 6.1.4.

There are further developments on `reddit` which one suspects are occurring but which cannot be verified with the available data. As the website's user-base has grown there seems to have been a growing tendancy for users to manage their subreddit subscriptions and to unsubscribe from many of the default subreddits. The evidence for this is chiefly observed in popular comments where users advise others to un-subscribe from default subreddits (particularly /r/politics and /r/atheism) and that `reddit` is best when one subscribes to smaller subreddits which match one's interests - the opinion that default subreddits have been declining in quality is also regularly voiced. Figure 5.24 was taken from one of the increasing number of websites which collect certain types of data through `reddit`'s API and display these, it shows the increase in subscribers for subreddits in descending order. As expected, the top 20 subreddits are the current default set (because when a new account is created it is automatically subscribed to these). One way to read this is that there have been around 150,000-160,000 new accounts created over the preceeding month - most users remain subscribed to /r/funny whereas /r/politics and /r/atheism see the greatest levels of un-subscription from new accounts (but still have much more 'growth' than the highest-placed non-default subreddit (/r/lifeprotips).

This suggests that as `reddit`'s user-base has grown it has also become more fragmented. The concept of posts being broadcast to all users through the default Front page is perhaps more applicable to the `reddit` of 2009 than the `reddit` of 2012 - and one senses that the idea of a single 'Reddit community' is not as prevalent on the website as it once was. Chapter 8 is devoted to the consideration of longitudinal trends on `reddit`.

In this chapter we have also considered voting levels on particular page types and the passage of a post to the Front page. The most active subreddits were found to have a greater 'depth' of voting activity (i.e. relatively high levels of voting on the New and Rising pages as compared to less active subreddits). This was found to produce a 'hurdling' effect on the highly active subreddits whereby posts had to progress through several pages (New -> Rising -> Main -> Front) and many were

30 days			
/r/funny	↑ +157,589		
/r/todayilearned	↑ +155,598	/r/videos	↑ +140,623
/r/bestof	↑ +154,552	/r/wtf	↑ +140,506
/r/pics	↑ +153,850	/r/worldnews	↑ +140,170
/r/announcements	↑ +152,352	/r/aww	↑ +134,054
/r/iama	↑ +150,348	/r/music	↑ +133,609
/r/science	↑ +149,706	/r/gaming	↑ +133,070
/r/technology	↑ +149,631	/r/adviceanimals	↑ +127,447
/r/askreddit	↑ +149,219	/r/politics	↑ +115,168
/r/blog	↑ +145,157	/r/atheism	↑ +97,746
/r/movies	↑ +144,725	/r/lifeprotips	↑ +33,905

Figure 5.24: A screen capture from redditmetrics.com taken on 14th December 2012 - showing the number of new subscribers for subreddits in descending order

filtered out at each stage. The hurdling effect was much weaker for less active subreddits and this resulted in a larger number of 'poor' (i.e. negatively scoring) posts appearing on the Main page for these subreddits.

One of the problems in assessing how well `reddit`'s post voting system works is that the only available measure of a post's 'quality' is its score - we are simply assuming that posts which receive a negative score are poor and those which receive the largest scores are good. To address the question of how *well* Reddit's post voting system works an external measure of 'post quality' is required. Section 7.3 details an experiment which was intended to produce such a measure.

Similarly, the analyses of comments presented in this chapter concern their mobility (i.e. whether they move up into highly visible locations or down into obscurity) but do not speak directly to the qualities of these comments. The role of comments and comment voting will be considered further in the context of specific case studies in chapter 7.

Finally, the analysis of /r/IAmA comments suggested that `reddit`'s users can collectively employ the comment voting system in specific ways to achieve specific ends. Chapter 7 will also re-visit this phenomenon whereby `reddit`'s users in some instances seem to operate as a unitary social entity.

5.2 Stack Exchange

This section explores the workings of the `Stack Exchange` family of websites introduced in section 3.6. The analyses concentrate on `Stackoverflow.com`, as this is the longest-running and most active of the `Stack Exchange` websites. The `Stack Exchange` websites periodically release publicly accessible data dumps which provide a good record of certain activites on the websites. *However*, when something has been deleted from the website (e.g. questions, answers, comments, user accounts) it will not appear in this data, and the data has no information on user voting behaviour as this is considered private. The data considered here was dumped in September 2011, and cover the period from the site's launch (31st July 2008) until 31st August 2011.

The intention here was to compare and contrast `Stackoverflow` with `reddit`, but in practice this is difficult because although both websites utilise voting systems these are used in very different ways and the data-sets which can be analysed concern different aspects of their use. Instead, we will consider `Stackoverflow` from the perspective of using DM to achieve a very different end. `Stackoverflow` is also interesting because it combines a variety of Social Media mechanisms with the voting systems that the present research is primarily concerned with.

What is `Stackoverflow`'s 'purpose'?

As we consider the analyses of `Stackoverflow` one of the dominant themes which emerges when one tries to compare it to `reddit` is that the two websites behave very differently despite sharing many of the same 'Web 2.0' tools. An important question to begin with therefore concerns the 'purpose' of `Stackoverflow`. Figure 5.25 shows an excerpt from the website's 'About' page. It is clear from the outset that this is a website which caters to programmers, the opening sentences contain sveral snippets of programming-related jargon and the reader is informed that `Stackoverflow` is collaboratively produced by 'your fellow programmers'. Taking the website's own description and structure into account, it seems reasonable to ascribe a primary and secondary purpose to `Stackoverflow` - the primary purpose being to provide answers to programming-related questions and the secondary purpose being to index these answers in such a way that a user can easily locate a previous answer to their question if one exists (thus avoiding a scenario where the same common questions are asked repeatedly).

The 'About' page acknowledges that `Stackoverflow` "synthesize[s] aspects of Wikis, Blogs, Forums and Digg/Reddit" - arguably Social Bookmarking should be added to this list because the tagging of user-submitted content was popularised on these websites and appears to be a central aspect of how `Stackoverflow` meets its secondary purpose of making answers navigable. `Stackoverflow` differs from `reddit` in that it provides comprehensive documentation on how the website operates and how a new user might go about contributing e.g. Atwood (2009).
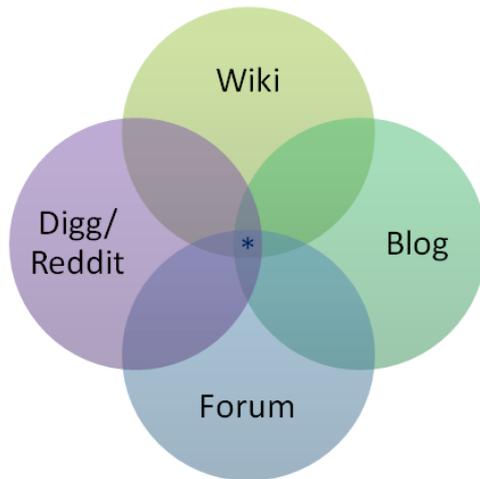
Figure 5.25: Excerpt from the `Stackoverflow` 'About' page

Stackoverflow's activity levels

All of the analyses of `Stackoverflow` are based on the data dump which was released in September 2011 (so covering the period from the website's launch (31st July 2008) until 31st August 2011. In this data-set there are records of 2 million questions, 4.4 million answers, 7.5 million comments, 16 million votes and 740,000 users. Of the 2 million questions 1,863,587 (92%) have at least one answer, and 1,268,745 (63%) have an answer which has been accepted by the user who asked the question. For questions with an accepted answer; the time between question submission and the submission of the answer which is ultimately accepted has a range of between 4 seconds and 2.9 years. This measure is highly skewed, as shown in figure 5.26. The median time until the accepted answer is submitted is 22.15 minutes. This is an impressive average response time, and suggests that `Stackoverflow`'s method of directing its users' attention is fulfilling the goal of having questions answered quickly. The presence of very long 'time until answer' intervals also suggests that questions do not suffer the same slide into obscurity that occurs for posts on `reddit`

- and that there are at least some users who browse older questions and submit answers where they can. There is in fact a 'badge' for answering a question which is more than 60 days old.

Time Until Questions are Answered
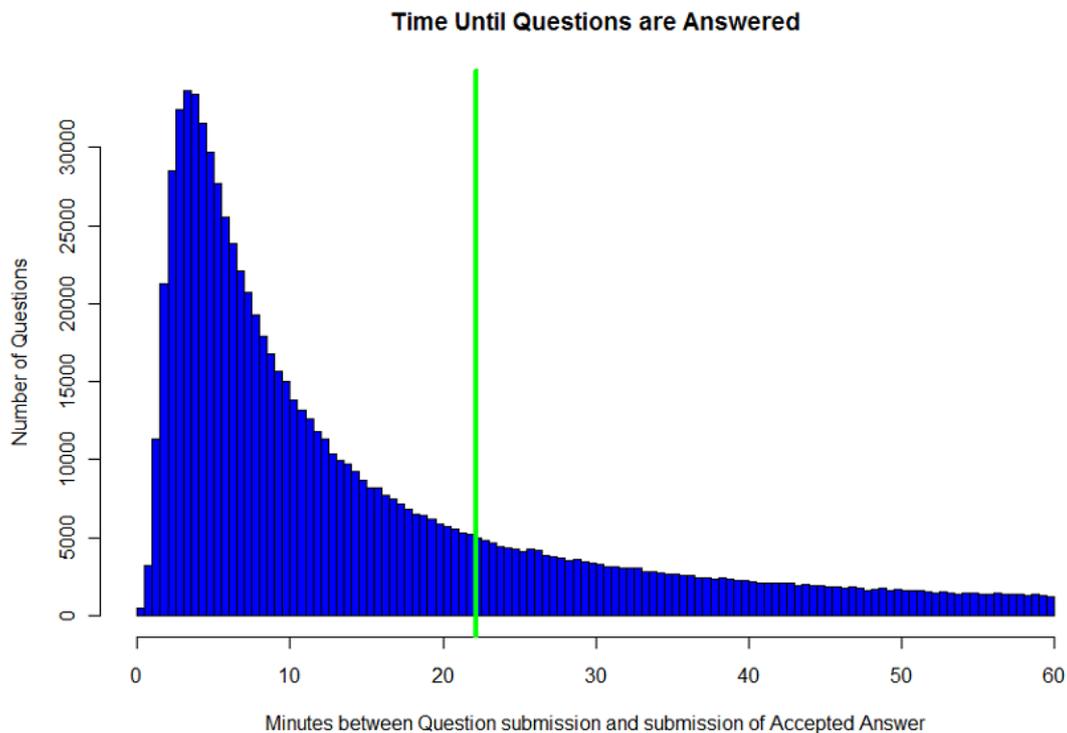
Figure 5.26: [Showing the time between the submission of questions and the submission of the answer which is ultimately accepted.]Showing the time between the submission of questions and the submission of the answer which is ultimately accepted. The green line shows the median time until accepted answer is submitted.

5.2.1 Answering Questions, Voting, and Accepting Answers

On the `Stack Exchange` websites the user who asked a question can select one answer as the answer they accept. In addition, `Stack Exchange` users can also vote to increase or decrease the score of answers; by default these scores are used to determine the order in which answers are displayed for a question (although an answer which has been accepted by the user who asked the question will appear at the top of the page regardless of its score). Of the 1,267,880 questions with an accepted answer; for 78.4% of these the accepted answer is also the answer with the highest score. If we exclude questions with just a single answer this leaves 817,419 questions that have an accepted answer, and for 63.7% of these the answer with the highest score is the answer which has been accepted.

Does the timing of answers influence their voting performance or chance of acceptance? This seems likely to be the case; only one answer can be accepted for each question and if an early answer is

accepted subsequent answers cannot also be accepted. The research of Oktay et al. (2010) indicates that the rate of answer submissions for a question on `Stackoverflow` is not affected by the presence of high-quality answers; users still submit answers to a question even once that question has been answered well. Figure 5.27 shows answer acceptance rate as a function of submission order for the first 10 answers submitted for questions which received at least 10 answers (N = 191,985). A logistic regression model was fitted to verify the robustness of this trend, and order of submission was found to be a highly significant predictor of whether an answer would be accepted ($p < 0.001$). The first answer submitted for the questions in this sample had a 13.8% probability of being accepted, as order increases the odds of each subsequent answer being accepted are 14% lower, so for instance the 10th answer to be submitted would have a 3.5% probability of being accepted.
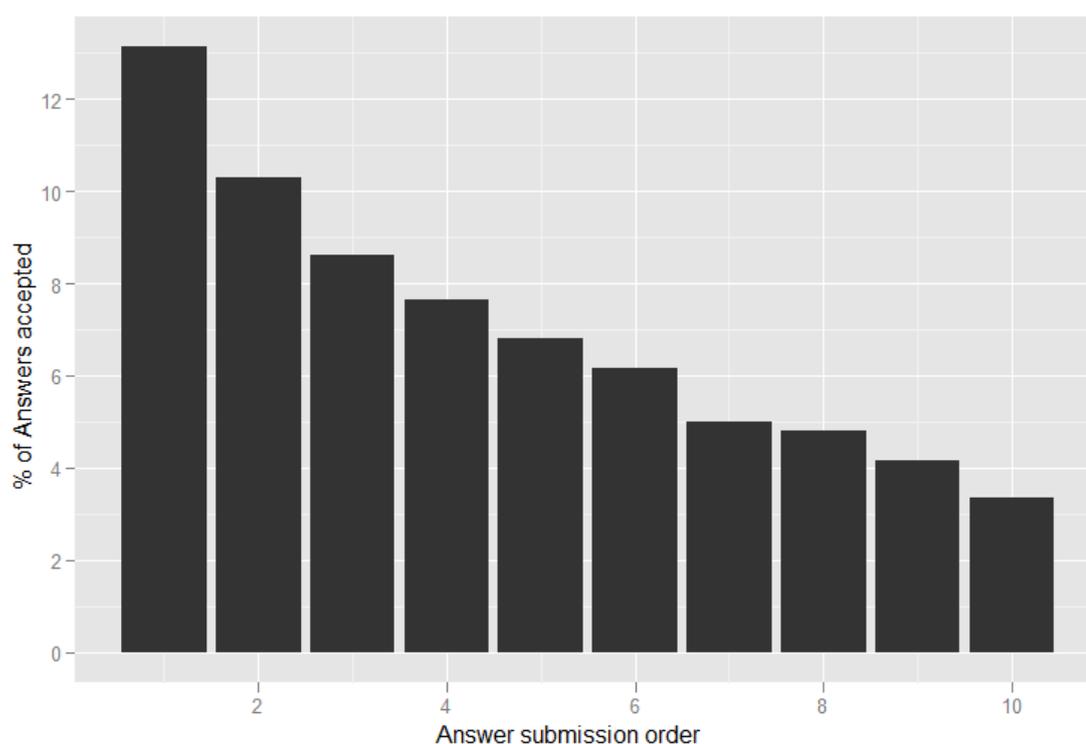
Figure 5.27: Showing Answer acceptance rate as a function of submission order for the first 10 answers submitted for questions which received at least 10 answers.

5.2.2 Which questions receive answers?

This section models the number of answers received by questions, searching for variables which can predict whether a question will receive many or few answers. As with many other measures taken from DM systems the distribution of answers between questions is highly skewed; many questions (36.5%) have received just one answer while a small number have received many answers (in this case the largest number of answers for a single question is 637). To deal with the skewness of

the data 'number of answers' has been converted to an ordinal variable with 6 levels (0 answers, 1 answer, 2 answers, 3-5 answers, 6-10 answers, more than 10 answers). Table 5.10 shows the number of questions from the random sample of 100,000 which fell into each category.

Answers received	Frequency
0 Answers	457
1 Answer	37,775
2 Answers	27,757
3-5 Answers	27,964
6-10 Answers	5,169
> 10 Answers	878

Table 5.10: Showing the number of Questions from the random sample of 100,000 which fell into each level of the ordinal 'number of answers received' variable.

As noted previously less than 1% of questions do not receive any answers, making this category much smaller than the others considered. These questions were included as a seperate category because these represent important cases (rare cases where `Stackoverflow`'s approach has failed to generate any answers).

Given that voting scores for questions influence their visibility it is important to first look at the relationship between voting score and number of answers. Voting scores are also highly skewed and therefore have been converted to an ordinal variable with six levels (negative score, 0, 1, 2-5, 6-25, 25+).

Ordinal logistic regression models work by splitting the ordinal dependent variable (number of answers) at a number of cut points. The coefficients for explanatory variables denote the log odds of posts with a given level of the explanatory variable (the reference category or intercept) falling on the larger side of these cut points. The ordinal logistic regression models fitted to `Stackoverflow` and described here were fitted using the `lrm` function in `R` to a random sample of 100,000 answers. The intercepts (or 'cut points') for these models relate to a series of dichotomies of the ordinal 'Number of Answers' variable. For each cut-point the coefficient describes the probability of a Question post receiving a number of answers which falls on the larger side of the cut-point. The vast majority of questions fall on the larger side of the first cut point - 226 questions received one answer or more for every question which received no answers. Conversely, only 0.005 % of questions received more than 10 answers and consequently fall on the larger side of the last cut point.

As expected there is a strong relationship between voting score and number of answers; the model containing just this explanatory variable accounts for 14.8% of the deviance in number of answers. Coefficients for each level of the Score variable can be used to generate predictions by summing

179

	Coefficient	Std. Error	p-value
Answers >= 1	5.4209	0.07207	< 0.001
Answers >= 2	0.4234	0.05505	< 0.001
Answers >= 3	-0.8257	0.05511	< 0.001
Answers >= 6	-3.1775	0.05673	< 0.001
Answers >= 11	-5.3932	0.06619	< 0.001
Score < 0	Reference Category		
Score = 0	-0.4356	0.05574	< 0.001
Score = 1	0.8628	0.05912	< 0.001
Scores 2-5	1.9929	0.11401	< 0.001
Scores 6-25	4.1112	0.17141	< 0.001
Scores 26+	7.0457	0.24544	< 0.001

Table 5.11: Showing model parameters for an ordinal logistic regression with number of answers as the response variable (as this is a 6-level variable it is parameterised as five binary variables or cut-points) and question score as the categorical indicator variable. There is a strong relationship between a question's voting score and number of answers.

them with the cut-point of interest. For example, to generate a prediction of whether a question with a score of 26 or greater would receive more than 10 answers we sum (7.0457-5.3932) and exponentiate the result - such a post is likely to receive more than 10 answers five times more often than not.

One slightly counter-intuitive finding derived from the model is that questions with a negative score (the reference category) tend to receive more answers than those with a score of zero (as indicated for the negative coefficient for the Score = 0 category). For any given cut point questions with a score of zero have 35% lower odds of receiving a larger number of answers than those with a negative score.

The relationship between question score and number of answers is a clear indication that up/down voting on questions affects their likelihood of receiving answers. The locus of this effect could simply be an increase in the question's visibility, or it may be the case that users are more likely to submit an answer to a higher-scoring question. For a website like `reddit` there is an obvious mechanism whereby voting response to an item determines its visibility (through the 'hot' ranking algorithm and Front page). For `Stackoverflow` this relationship cannot be assumed because although pages exist which display items by score these are not 'default' options as they are on `reddit` - if score influences visibility it suggests that these pages are in fact viewed by `Stackoverflow`'s users.

A further possibility concerning the relationship between question score and number of answers is that users are simply more likely to provide answers to good questions, and that the score a

question receives through the voting system is a good estimator of its 'quality'.

Another question one might ask concerning the answering of questions is whether there is an effect related to the status of the user who asks the question. `Stackoverflow` appears as a peer-produced resource which is accessible to all - any individual can make use of `Stackoverflow` by submitting their question and having it answered by knowledgable users. However, it is conceivable that established users may be more likely to receive answers to their questions than new users. If `Stackoverflow`'s community of users were concerned that their collective question-answering resources were being stretched or exhausted, they may choose to give priority to other users who themselves have a history of contributing on the website. In other words a new user might have to earn the 'right' to have their questions answered - in a scenario where `Stackoverflow` had a perceived problem with 'freeloaders' this may appear as a valid community-level strategy to deal with such a problem. Do `Stackoverflow`'s users give priority to questions asked by established contributors when they answer questions?

To answer this question a user's post order is considered. A user's first post (question or answer) has been assigned an order of 1, with the order variable increasing by one every time a user submits a question or answer. This order variable is again highly skewed and is again treated as ordinal to circumvent this problem. A model which included the order variable as a second explanatory variable only accounted for an additional 0.1% of deviance in number of answers, suggesting that the effect of this variable is quite weak.

In this model, users making their first post are the reference group, and users asking a question which is their first post have 8% higher odds of receiving more answers than users asking a question with an order of 2-8. Beyond this category of order 2-8 there is a trend whereby users who have made more previous posts receive more answers; but this trend is quite weak. At the high end of this scale users with more than 100 previous posts have 43% higher odds of receiving answers greater than a given cut point when they ask a question. There is a trend whereby established users receive more answers to their questions but this trend is quite weak and does not constitute evidence that users must be recognised as community members before they will receive answers to their questions.

Questions on `Stackoverflow` can be tagged to describe the topics or areas a question relates to and thus make it easier to find for users who are interested in those topics. One might expect a higher number of tags for a question to be associated with increased visibility and therefore a larger number of answers. However, when a variable representing the number of tags is added to the ordinal logistic regression model this measure is actually found to be associated with a lower number of answers. The effect is however quite weak; for every tag added for a question the odds of it receiving answers on the large side of a given cut point decrease by 5%. The direction of this effect could be due to a greater number of tags being related to more specialised questions - with

the lower number of answers being an indication that there are fewer users who feel that they are capable of answering more specialised questions.

If instead of considering the number of answers a question received as the response variable we consider the number of times the question was viewed - we find a different effect of the number of tags. The number of Views per question was split into a seven-level ordinal variable and a model with number of tags as the explanatory variable was fitted. This model produced a significant effect of number of tags; with each additional tag for a question being associated with a 20% increase in the chance that it would have a number of views falling on the larger side of a given cut-point. This supports the idea that tagging on `Stackoverflow` serves primarily to make questions easier to navigate for users who are searching through existing questions; as opposed to helping would-be answerers to find questions they can answer. Therefore we might say that tagging on `Stackoverflow` facilitates the meeting of the website's 'secondary' aim (producing a knowledge-based resouce which is easily navigable) ahead of its primary aim of producing answers to programming-related questions.

5.2.3   What makes a good (high-scoring) answer?

This section looks for predictors of whether answers will receive a high or low score. Answer scores are highly skewed and have been classified as falling into one of six levels of an ordinal variable. These are the same six levels used to categorise question scores (negative score, 0, 1, 2-5, 6-25, 25+). Table 5.12 shows the number of answers from the random sample of 100,000 which fell into each category.

Answer Score	Frequency
Score < 0	1,345
Score = 0	36,561
Score = 1	27,189
Scores 2-5	28,373
Scores 6-25	6,149
Score > 25	338

Table 5.12: Showing the number of Answers from the random sample of 100,000 which fell into each level of the ordinal score variable.

The scores of questions are a good indicator of how visible these questions have been made, and questions with higher scores are for this reason likely to have answers which generally have higher scores. If a question has a high score it seems sensible to assume that more voting will have taken place on its answers than for a question with a lower score - and therefore that the answers for such a question will tend to have higher scores.

	Coefficient	Std. Error	p-value
Score >= 0	5.4209	0.07207	< 0.001
Score >= 1	0.4234	0.05505	< 0.001
Score >= 2	0.4234	0.05505	< 0.001
Score >= 6	-0.8257	0.05511	< 0.001
Scores >= 26	-3.1775	0.05673	< 0.001
Question Score < 0	Reference Category		
Question Score = 0	-0.3952	0.04889	< 0.001
Question Score = 1	0.8863	0.05354	< 0.001
Question Scores 2-5	1.9090	0.10123	< 0.001
Question Scores 5-25	3.1499	0.15076	< 0.001
Question Score > 25	4.6074	0.20404	< 0.001
Accepted Answer	1.4105	0.01359	< 0.001
User Order = 1	Reference Category		
User Order 2-10	-0.1439	0.05487	< 0.01
User Order 11-50	0.4035	0.06312	< 0.001
User Order 51-150	0.7662	0.11652	< 0.001
User Order 151-500	1.1649	0.17084	< 0.001
User Order > 500	1.6875	0.22552	< 0.001

Table 5.13: Showing model parameters for an ordinal logistic regression with answer score as the response variable (as this is a 6-level variable it is paramaterised as five binary variables or cut-points) - explanatory variables are question score (ordinal), whether the answer was accepted (binary) and user's post order (ordinal). This model accounted for 22.5% of the deviance in answer score.

The score of the question each answer was submitted for is the first explanatory variable to be considered; and the ordinal version of this variable accounts for around 7.2% of the deviance in answer scores. The relationship between question score and answer score is similar to that between question score and number of answers. Answers submitted to questions with a negative score tend to receive a more positive score than those submitted to questions with a score of zero. When questions have a positive score, the larger this score is the more likely its answers are to also have a large score. For example, answers submitted to a question with a score of greater than 25 have 75 times the odds of having a score on the large side of a cut point than answers submitted to a question with a negative score.

The next variable to be added to the model was a binary variable indicating whether an answer had been accepted by the user who asked the question. This variable significantly improved the utility of the model, increasing the amount of deviance it accounted for to 19.5%. There is a strong trend

associated with accepted answers; these have odds 4.5 greater of having a score which falls on the high side of a cut point than equivalent answers which were not accepted. This supports the idea that there is a strong relationship between an answer's score and whether it will be accepted; and this in turn supports the notion that answer scores and acceptance are related to a more abstract concept of answer quality (which we do not attempt to measure here).

Next a variable representing the order of answer submission was added to the model. This variable produced a coefficient which suggests that answers submitted later in the lifespan of a question do not receive scores as high as those submitted earlier. However, in the presence of the question score variable and the answer acceptance variable this answer order variable did not improve the utility of the model significantly.

The next variable to be added to the model considers the `Stackoverflow` experience of the user who submitted the question; as measured by the 'user post order' variable. This variable was broken down into six levels (user's first post, order 2-10, 11-50, 51-150, 151-500 and 500+). Adding this variable to the model increased the amount of deviance it accounted for by 3% to 22.5%. There is a generally positive relationship between users' experience and the scores of their answers. The exception to this trend is a user's first post, which is around 20% more likely to receive a large score than posts with a user order of between 2 and 10. This may be a result of users who had previously been 'lurkers' creating a `Stackoverflow` account for the purpose of submitting an answer to a specific question which they felt well qualified to answer.

Beyond the answers submitted by users with 2-8 posts, answers from users who have made more previous posts tend to receive higher scores; with answers from users who have more than 500 previous posts having 7.6 times greater odds of receiving a larger score than answers submitted by a new user. This is a strong indication that users who contribute heavily to `Stackoverflow` are especially valuable to the website; not only do they submit more answers but their answers are generally of higher quality than those submitted by their less experienced counterparts.

Table 5.13 gives details of the final model of answer score.

5.2.4 Voting on `Stackoverflow`

Data on voting is anonymous with respect to user ID and this limits the analyses of voting which can be conducted. We can however present a general overview of voting. In total there are records of 12,943,596 votes and of these only 6.3% are down-votes - down-voting is much less prevalent on `Stackoverflow` than on `reddit`. We can also consider the ratio of votes to items of content (questions/answers) and on `Stackoverflow` there were 2 votes for each question/answer post - in March 2009 on `reddit` there were almost 10 times as many votes as there were posts. This is indicative of the much greater centrality of voting on `reddit`.

5.3 Your Freedom

The 'Your Freedom' website was introduced in section 3.2.1. It was launched by the current British coalition government in 2010 (and hosted at www.YourFreedom.hmg.gov.uk). Your Freedom solicited political policy ideas from the general public and allowed users to rate each other's ideas; in this case the intention was that users would use the system to collectively highlight pieces of legislation which should be repealed or submit ideas for the enhancement of civil liberty. Your Freedom opted for five-star ratings of ideas, as opposed to up/down voting, and this appears to have been a sub-optimal choice. Your Freedom also had limited facility for commenting and no method of effectively rating and ranking comments; this limited any capacity Your Freedom users may have had to discuss and develop their ideas as a community.

The Your Freedom website used a piece of software ('Dialogue App') produced by Delib Ltd. Although the Your Freedom website has long since been disbanded one can still browse an archived snapshot of the website through the national archives (YourFreedom, 2010).

While Your Freedom was active requests were made for access to some of its procedural data. Although no such data was provided, some time after Your Freedom was shut down a data-set containing details of the interaction which had taken place there was released. This data-set is comprised of 13,954 rows each representing an 'idea' (there will be referred to as posts) with some measures of its voting performance. No data was provided on the comments which accompanied these posts. Your Freedom was split into three subject areas, and table 5.14 shows the level of activity in each of these - the area for discussion of "Cutting Business and Third Sector Regulations" has seen much lower levels of activity than the other two areas.

Subject	No. of Posts	No. of Ratings	No. of Comments
Repealing Unnecessary Laws	5,308	94,527	36,942
Restoring Civil Liberties	6,166	122,157	48,073
Cutting Business and Third Sector Regulations	2,480	20,548	9,564

Table 5.14: Showing activity levels for each of Your Freedom's three subject areas.

Figure 5.28 shows the distribution of ratings between posts and this is highly skewed but does not appear to follow a power law. In total 237,232 ratings were submitted. 80% of these ratings were distributed between the top 2,943 (21% of) posts - close to the 80-20 rule associated with power law distributions. 25% of all ratings accrued to the top 202 (1.5% of) posts. The 'low' end of the distribution is where it deviates from what would be expected of a power law - this is an indication that the areas of the website which showed new content received more user attention (relative to the areas showing popular content) than their counterparts on reddit.

Figure 5.29 shows the distribution of comments between posts and again this is highly skewed,
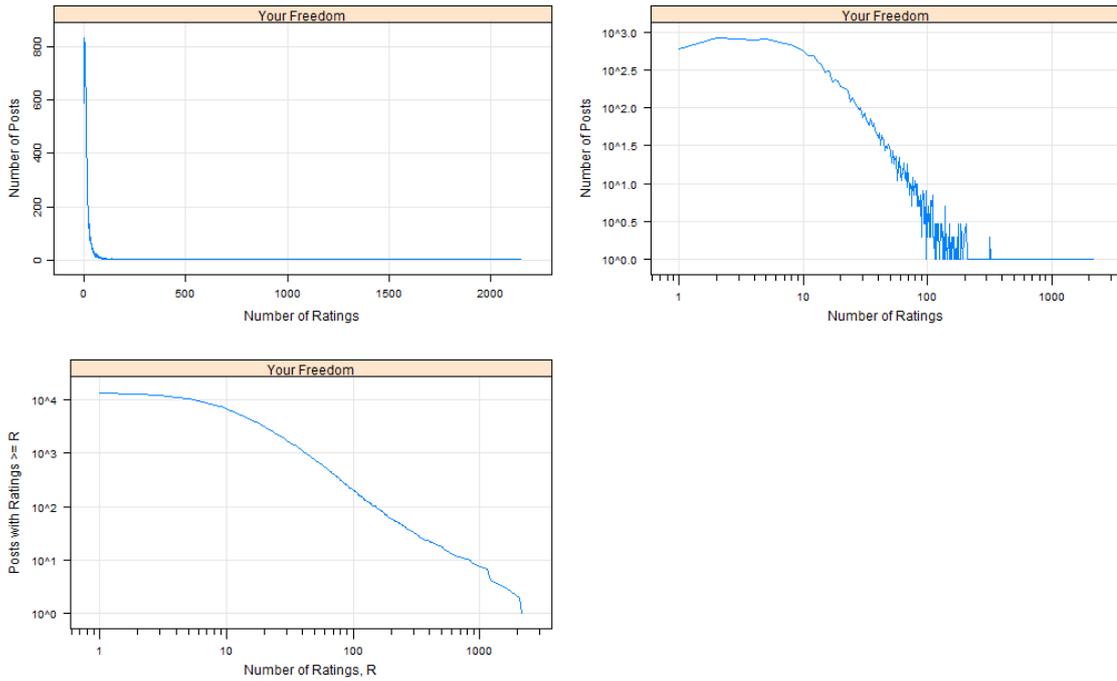
Figure 5.28: Showing the distribution of ratings between posts for `Your Freedom`

following a very similar distribution to that of ratings between posts. Furthermore, Pearson's correlation coefficient between posts' number of ratings and their number of comments was calculated (r = 0.82) and this suggests a strong relationship between these variables. The level of commenting activity on a post closely parallels its level of rating activity - supporting the idea that these measures reflect the visibility of posts on `Your Freedom`.

For each post there is a variable representing the average of all ratings made for that post on the 5-point scale. In the analysis of average ratings the 670 posts which received no ratings have been excluded. The mean of the average ratings for posts which received at least one rating is 3.68. Figure 5.30 shows the distribution of the average rating variable and it is immediately apparant that most of these average ratings are quite high. There are also 575 posts with an average rating of 1, the lowest possible rating.

Next we consider whether there is a relationship between average rating and the number of ratings a post received. The presence of such a relationship would indicate that posts with higher average ratings were being seen by more people. If one of `Your Freedom`'s goals was to focus attention on the best ideas then a relationship between idea quality (as measured by average rating) and attention (as measured by number of ratings) would be a desirable characteristic. In order to look for such a relationship the highly skewed number of ratings variable was converted to an ordinal variable - for graphing purposes this has six levels, for modelling a seventh level has been added to allow for the consideration of highly active posts (100 or more ratings) seperately.

186
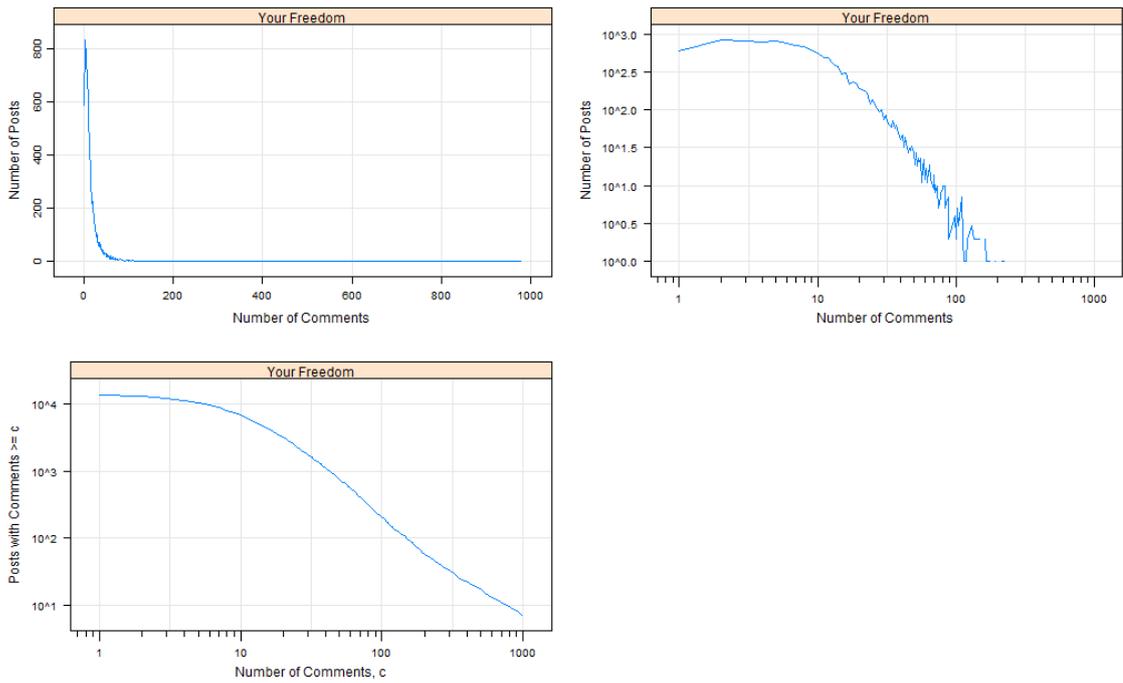
Figure 5.29: Showing the distribution of comments between posts for `Your Freedom`.
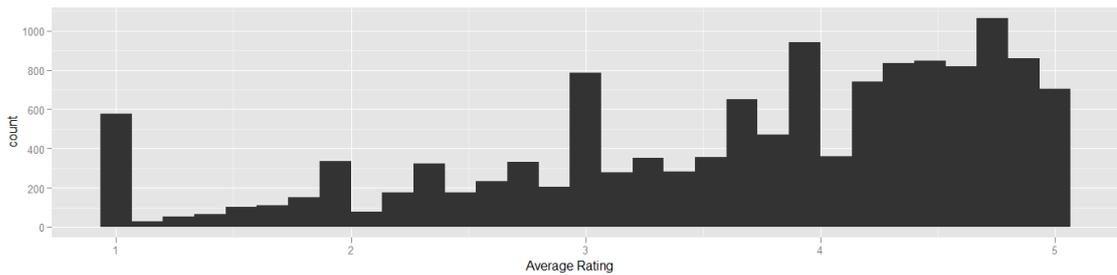
Figure 5.30: A histogram showing the average ratings achieved by posts to `Your Freedom`.

Figure 5.31 suggests that the relationship between average rating and total number of ratings is weak. Posts with an average rating of greater than four seem equally likely to have received a number of ratings falling in any of the largest three levels of total ratings. A linear model assuming a normal distribution was fitted to average rating with total number of ratings as the explanatory variable. This model shows a significant effect of total number of ratings in that all of the ordinal groupings with more than 2 total ratings are significantly different to the reference group (posts with 1-2 ratings). The magnitude of this effect is however weak, particularly at the higher end of the number of ratings distribution - where posts which received 100 or more ratings had on average only a slightly higher rating than those which received 33-99 ratings.

Figure 5.32 shows average ratings for only posts with 100 or more ratings (the top 1.5%) and even within this set there are a substantial number of posts with low average ratings. Furthermore, of the seven posts which received more than 1,000 ratings one of these had an average rating of just 1.8.
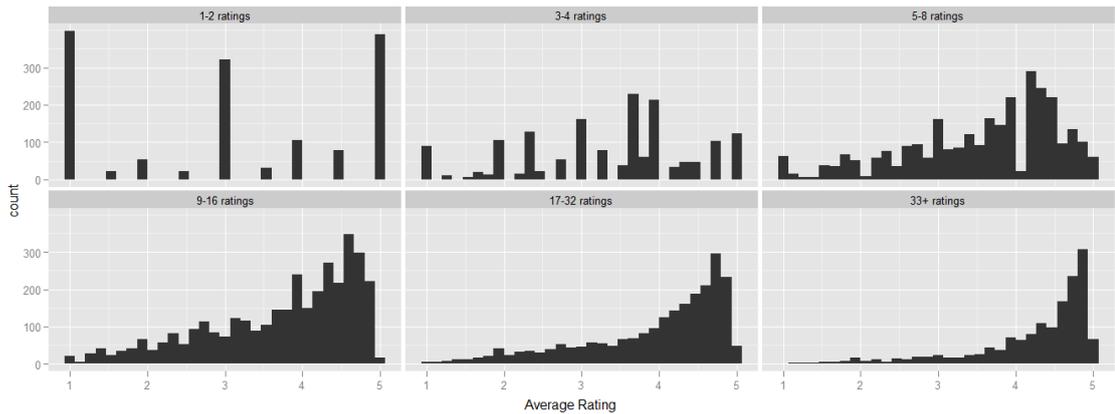
Figure 5.31: A histogram showing the average ratings achieved by posts to `Your Freedom` with panes representing posts which received a certain number of ratings in total.

	Coefficient	Std Error	p value	Est. Average Rating
Intercept - Posts with 1-2 ratings	3.01	0.03	< 0.001	3.01
Posts with 3-4 ratings	0.27	0.04	< 0.001	3.28
Posts with 5-8 ratings	0.48	0.03	< 0.001	3.49
Posts with 9-16 ratings	0.64	0.03	< 0.001	3.65
Posts with 17-32 ratings	0.86	0.03	< 0.001	3.87
Posts with 33-99 ratings	1.14	0.03	< 0.001	4.15
Posts with 100+ ratings	1.26	0.03	< 0.001	4.27

Table 5.15: Showing model parameters for a linear model of average rating with total number of ratings as the explanatory variable.

This shows that low-quality posts (those with low average ratings) were not being excluded from high-visibility areas of the website. This is indicative of a problem, because presumably when one *designs a system which focuses user attention on a small sub-set of posts*, it is the highest-quality posts which should be displayed in the high-visibility areas.

Returning to figure 5.31 one can see a clue to the nature of this problem. The number of posts with an average rating at or approaching 5 (the maximum) is higher within the categories of post which have low numbers of total ratings. `Your Freedom` allowed users to sort posts (optionally within certain sub-categories) by three criteria - new-ness, average rating and total number of ratings. There is an obvious problem in combining these display mechanisms with a 5-point rating scale, because the measure of a post's 'quality' does not seem to take into account the number of ratings which have been used to calculate the average score. Therefore one assumes that posts will only be able to attain the maximum measure of quality (an average rating of 5) for a limited amount of time until they receive their first rating of less than 5 (from which point onwards their average will be less than 5).
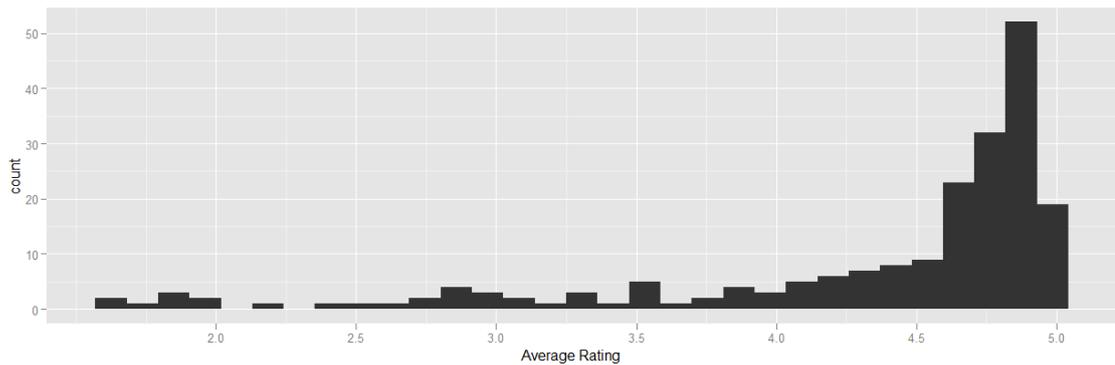
Figure 5.32: A histogram showing the average ratings achieved by posts to `Your Freedom` which received 100 or more ratings

The result is that pages showing where the community's attention is (posts with the greatest number of ratings) show a distinct set of items from pages showing the posts that have been most highly rated (posts with the greatest average score). By not combining these measures a large number of potentially high-quality posts have likely disappeared into the void between posts with high average ratings and posts with a high total number of ratings. Without a link to post quality, the pages showcasing posts with the highest number of ratings likely became static. Posts appearing here became 'locked in' to their positions in a self-perpetuating cycle of gaining more ratings because they appeared on a high-visibility page. There was no page showing highly rated posts *with a large number of total ratings*, and therefore no real pathway by which a high-quality idea could supplant one of lower-quality on the pages showing the most frequently rated posts. The evidence for this is largely in the fact that posts with low average ratings were among the posts with the greatest total number of ratings. Also, of the 19 posts which received more than 500 ratings each - 17 of these were submitted on July 1st 2010, the day that `Your Freedom` launched.

One of the two posts which defied this pattern may be indicative of an issue which sites like `Your Freedom` face - an organisation or interest group's members could co-ordinate their actions to use `Your Freedom` as a means of pushing their own agenda. This idea was titled "Abolish Control Orders", it was submitted on 9th July and somehow managed to achieve more ratings than any other idea, along with a very high average rating (4.97). It is 'suspicious' merely because `Your Freedom`'s design did not seem to allow for 'new' ideas to make their way unto the most frequently rated list of ideas once this had been established on the first day. This idea would not have been displayed with a great degree of visibility on `Your Freedom` initially upon its submission, and it is difficult to imagine how it could have overcome this barrier *within the confines of* `Your Freedom`. A more likely scenario is that a group which existed externally to `Your Freedom` co-ordinated their actions with the aim of positioning this idea in a highly visible location. Within the context of `Your Freedom` and all of the problems which plagued it, this issue barely warrants a footnote.

However, this will tend to become a more serious problem for any instance of DM once it is widely perceived that placement in a high-visibility location has value. Both `reddit` and `Digg` have had problems with 'voting cabals', and these problems would have the potential to cause more damage on a DM website which dealt with politics (*if* anyone was taking its outputs seriously in the first place).
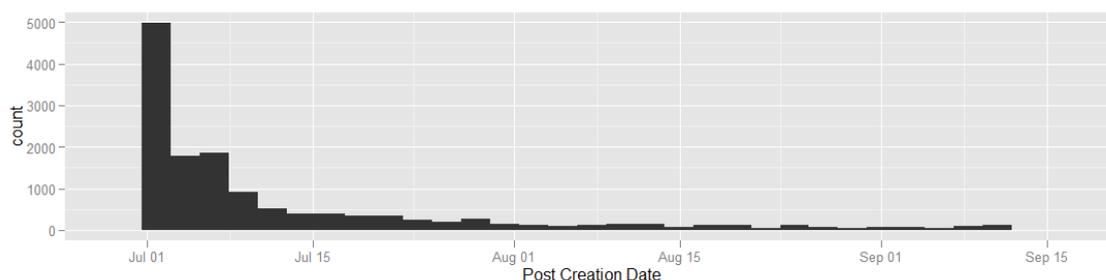
Figure 5.33: Showing the creation dates of `Your Freedom` posts

Figure 5.33 shows the creation dates of `Your Freedom` posts - after an initially high rate of new post submissions this rate has dropped off sharply. In total 1,973 posts were submitted on 1st July 2010, posts created on the first day account for 14% of all posts - meaning that posts from this day are still over-represented on the 'most frequently rated posts' page. `Your Freedom` was featured in several newspapers (Guardian, 2010b) on its launch day and this accounts for the initially high level of participation. Indeed, the level of participation was high enough to cause problems for `Your Freedom`'s servers on the first day and parts of the website were inaccessible at times. The subsequent drop-off in post submission may reflect dissatisfaction with `Your Freedom` or may indicate that many users came to `Your Freedom` just to submit one of their own ideas. Speaking as someone who was looking at `Your Freedom` in its first few days the website had obvious problems - it was not clear where one was supposed to be looking or where the community of users were active, and the website seemed to be plagued by duplicate ideas. Also, the five-star rating system for ideas was un-satisfying at best, in making one of these ratings one did not feel any sense of impact at all, this action did not serve the purpose of making the rated idea any more or less visible to other users and was also limited in being able to reflect users' opinions about the idea.

Which Ideas were popular on `Your Freedom`?

In the absence of a clearly defined and workable measure of quality let us consider the ideas which saw the most rating/commenting activity on `Your Freedom` as representative of the community's opinions. Table 5.16 shows the top 25 ideas by total number of ratings - posts also had sections in which they outlined their idea in detail and a further 'Why?' section in which they made a case for the idea. There are several ideas which appear on this list a number of times - chief among these are ideas about the 'Digital Economy' act - a piece of legislation which introduced harsh

anti-piracy measures and was widely perceived to have been pushed through parliament in the last stages of the Labour government. In total there are four ideas in the top 25 about this subject, and these have very high average scores (ranging from 4.95 to 4.97). If there is one idea which Your Freedom users are in agreement on it is that the Digital Economy act should be scrapped or heavily modified. There are also four ideas in the top 25 which concern drug prohibition (3 relate to cannabis specifically), these have average scores ranging from 4.64 to 4.75, suggesting strong support among Your Freedom users for a more liberal approach to drugs legislation.

There are two ideas about anti-terrorism laws - one concerning the abolition of control orders and the other expressing concern about expanded stop-and-search powers for the police - these have high average scores, 4.97 and 4.9 respectively. There are two ideas about compulsory worship in schools which also have high average ratings (4.66 and 4.61). There are two ideas about repealing the smoking ban which received a more mixed response (average scores of 3.05 and 3.35), and two ideas about repealing the 2004 hunting act which have average ratings of 2.89 and 2.92. Finally, the fourth most frequently rated idea concerned bringing back the death penalty but this had a low average score (1.81), suggesting that Your Freedom users did not approve of this idea at all.

One of the trends which emerges from this table is that similar ideas tend to have similar average scores - suggesting that this method was at least a *reliable* way of measuring users opinions on the ideas. However, for ideas with mid-range average scores it is difficult to interpret what this number means - it could have stemmed from polarised opinions (many 1 or 5-star ratings) or more neutral opinions (many 3-star ratings).

Figure 5.34 shows the average scores for posts with certain common key terms in their title. Ideas concerning certain topics show a remarkably consistent voting response - posts about the Digital Economy act, the right to take photographs (Photo), compulsory worship in schools (Worship) have received almost exclusively high average scores. Posts about Cannabis/Drug liberalisation and the Smoking ban appear to have a more mixed response - but when one inspects the posts those which had a low average score were calling for an opposite course of action to those with high average scores (the low-scoring ideas about cannabis have titles like "Nip cannabis smoking in the bud", low-scoring posts about smoking tend to call for an outright ban on smoking). Posts about repealing the hunting act or the human rights act tend to have mid-range average scores (with higher-scoring posts containing these key words tending to argue the opposite point).

Your Freedom also employed a tagging system and some analysis of tags has been conducted. However, the analysis of tags is limited for a number of reasons: 1) a moderator has been employed to re-tag some ideas, so some of the ideas with the most ratings have a single tag like 'duplicate', or 'not repeal' when a post did not relate to repealing legislation, and 2) generic tags like 'freedom' or 'justice' outnumber tags which more acurately reflect the substance of an idea.
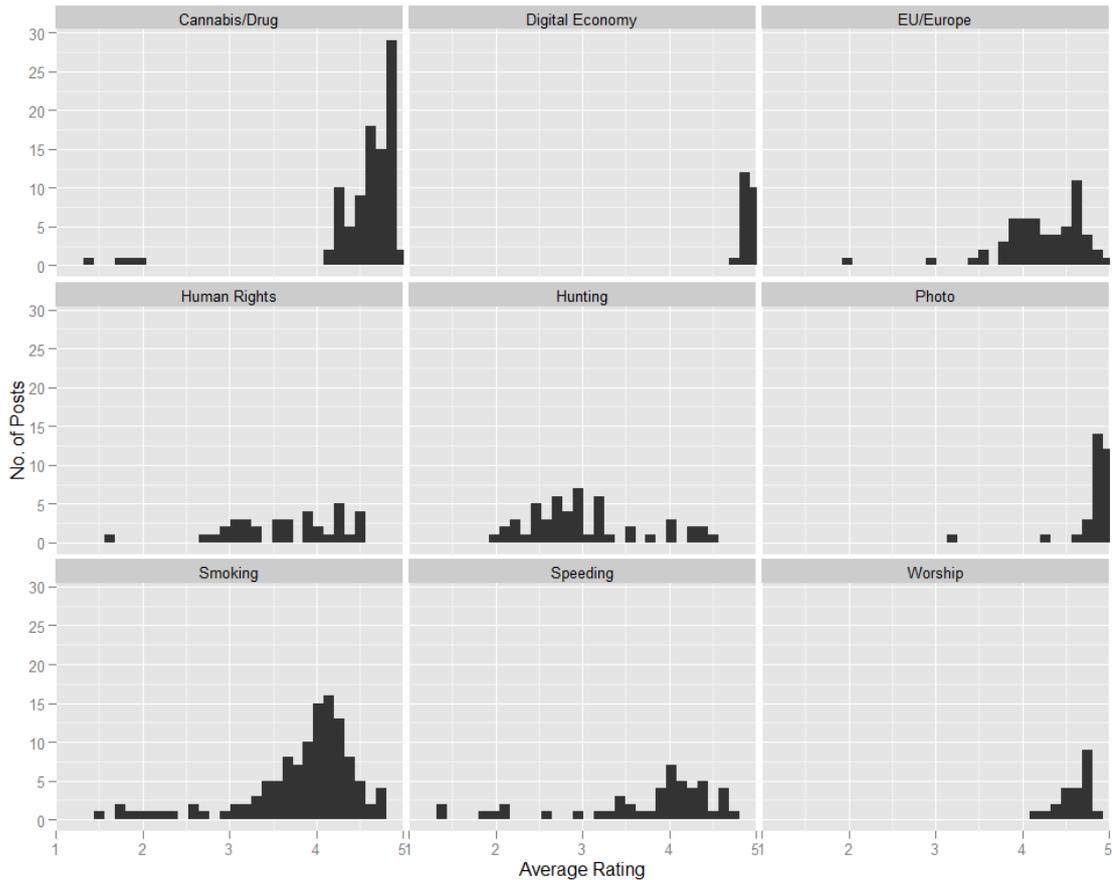
Figure 5.34: Showing the average scores for posts with certain key terms in their title. Posts which received less than 30 ratings in total have been excluded.

Your Freedom's purpose

The press release from Number 10 which accompanied the launch of `Your Freedom` concludes with the following statements: "Users of the site will be able to comment on and rate their favourite ideas and relevant departments will then respond to the most popular workable ideas. The views expressed on `Your Freedom` will be taken into account in the Freedom Bill later this year." One wonders which measure is being used to identify the 'most popular' ideas - are these the 581 ideas with a perfect average rating of 5 (based on a mean 2.35 ratings each)? Maybe they are the 88 ideas which achieved an average score of 4.9 or greater based on 50 or more ratings.

I do not personally recall any response whatsoever from the government to any of the ideas on `Your Freedom`, nor did I expect to see one after I had taken a look at `Your Freedom`. Furthermore, the lack of specificity as to what defined a 'popular idea', and the caveat that only those which were 'workable' would be responded to, left one feeling that should the government acknowledge the activity on `Your Freedom` at all they would merely cherry-pick a few ideas which matched their own goals. This lack of specificity in how the website's primary goal (bringing popular ideas

to the attention of government) would be fulfilled resulted in an air of futility hanging over `Your Freedom`.

I would argue that bringing individuals' ideas or opinions to the attention of the government was itself not the right goal for an endeavour like `Your Freedom`. There are already established mechanisms through which peoples' ideas and opinions can be brought to the government (letters, petitions, protests) - and if the government can ignore hundreds of thousands of people marching through London they will presumably have no problem ignoring the most popular ideas on a website like `Your Freedom` which is being used by a few thousand individuals.

The first problem with this kind of goal is that all of the activity on `Your Freedom` was, by implication, merely for the benefit of an external audience (the ministers who were going to be presented with the most popular ideas). In this sense `Your Freedom` was just another way of trying to have one's voice heard and in execution it was not dissimilar to the government's system for e-petitions (Miller, 2009) - except that it was possible on `Your Freedom` to register an opinion on ideas which was not one of support, and to comment on the ideas.

The lack of voting on comments is a missed opportunity to *begin* moving towards a more collaborative kind of space. Here `reddit`'s approach to comment voting and ranking is informative - if a similar system had been applied to the comment pages of `Your Freedom` this might have allowed users to better and more easily understand what the community as a whole thought of the ideas. Through democratically-mediated discussion threads and the capacity to edit an idea it may have been possible for users to not only express their idea and have it rated but to work on improving their idea - taking the opinions expressed by other users (and ratified through comment voting) on board and using these to refine the idea.

Indeed it is difficult to see what the advantages of the 'Dialogue App' software which powered `Your Freedom` were, as compared to a straight replication of `reddit`'s software. Reddit's software is open source and so could likely have been appropriated for this use at little cost - this would however have required greater planning and the hiring of employees to conduct the work, and in this sense would have been more difficult than paying Delib for their 'Dialogue App'.

Up/down voting systems assign each item a score, a single number which tells us something about the popularity of the item, large scores indicating that a lot of people endorse the item. If we wish to know what proportion of people endorse an item we can look at its ratio of upvotes to downvotes. The scores produced by up/down voting are also easily converted into ranks, and a time penalty can be applied if required. Up/down voting also offers a neater conceptual fit with political policy ideas; when politicians vote on legislation they can either vote for or against it, or abstain.

Five-star ratings do not offer the same conceptual fit with voting in the domain of politics. One

might imagine that a rating of 5 equates to full support and 1 equates to full opposition but what is the meaning of a 2, 3 or 4 star rating? In aggregating peoples' ratings the tendency is to calculate a mean rating for each item, but this is not as easily interpreted or used as an up/down voting score. For instance, an item with a full 5-star rating based on 5 votes probably doesn't warrant as much attention as an item with a rating of 4.9 that has received 500 votes. How would one interpret a mean rating of 3, which could be produced by many people with neutral feelings on the matter or two sets of individuals with strong feelings either way.

The statement which accompanied the launch of `Your Freedom` referenced 'popular ideas' and 'your views' but the software which was used did not offer any obvious manner in which the most popular ideas and views would be determined. Had `reddit`'s software been deployed for this purpose this would likely have produced clearer outcomes. The most popular ideas would have been the posts with the highest scores (ratio of up to down-votes could also have been considered), and the community's views on these ideas would have been represented by their highest-scoring comments. A 'Front page' using `reddit`'s 'Hot' algorithm would also have been very useful in showing users where the community's attention was currently focused. When `Your Freedom` launched I had already been observing `reddit` for some time and by comparison `Your Freedom` lacked any way to quickly get a feel for 1) who its users were and what they thought about these issues, and 2) where users' attention was or what they were discussing at that moment. There was, perhaps as a consequence, no sense that `Your Freedom`'s users represented a community - whereas on `reddit` at the time there was quite a strong sense of a 'Reddit community', and this on a website with more users than `Your Freedom` by orders of magnitude.

In applying `reddit`'s software (without modification) to political discourse there would likely be problems related to the suppression of minority opinions and the homogenisation of the user-base as a result of drop-out (see Chapter 8). `Your Freedom` did not encounter these problems because it failed even to convey what the opinions of its users were. Furthermore, given the stated aims of `Your Freedom` (determination of what the 'most popular' ideas were and what the 'views' of users were, so that these could be shown to ministers or government officials) - these would not necessarily have been problems at all.

Section 9.2 considers alternative ways in which Distributed Moderation might be applied and adapted to political discourse - and makes the case that 'presenting top-ranking ideas to an external audience' is not a particularly suitable goal for an application of DM to politics.

Idea	Avg. Rating	No. Ratings	No. Comments
Abolish Control Orders	4.97	2147	563
Repeal the Digital Economy Bill	4.96	2061	413
Digital Economy Act	4.95	1612	348
Bring back the Death Sentence	1.81	1254	460
scrap law that says school must hold collective worship	4.66	1206	411
Repeal and change the Smoking Ban	3.05	1180	978
Decriminalise Cannabis along the same lines as the Netherlands	4.63	1149	932
Repeal Drugs Prohibition	4.75	962	540
'Ban' smoking ban in pubs. Give the Landlord/Landlady the choice.	3.35	860	414
Stop compulsory worship in schools	4.61	847	259
Legalise and Tax Cannabis	4.74	737	501
Repeal the Digital Economy Act 2010	4.97	637	50
Repeal Section 44 (Terrorism Act 2000)	4.9	630	109
Save Britain's Digital Economy by Repealing the Digital Economy Act	4.95	580	71
Repeal the Ban on Tail Docking of Dogs	4.57	559	308
Leave the EU - that should stop most of the daft, expensive legislation	3.78	517	222
Scrap the Human Rights Act	2.95	512	191
Legalise cannabis for recreational and medical use	4.69	510	166
To gather and hold an Annual Free festival	4.81	483	316
End IR35 immediately	4.8	451	259
Repeal the Hunting Act	2.89	424	259
Repeal the 2004 Hunting Act	2.92	411	148
Abolish Car Tax and put on Fuel	4.64	410	175
Prostitution (legalize brothels)	4.75	382	84
Repeal retrospective elements of section 58, 2008 Finance Act	4.88	358	609

Table 5.16: Showing the 25 Idea posts with the highest number of ratings. Posts also had sections in which they outlined their idea in detail and a further 'Why?' section in which they made a case for the idea.

5.4   Chapter discussion

The analyses conducted on `reddit` and `Stackoverflow` have revealed very little similarity between the websites aside from the presence of voting systems. Differences in the types of data which are available from each website and the structure of the sites themselves make it very difficult to compare them directly. Instead the picture which emerges is one whereby these two websites have woven voting systems into their structure in ways which are tailored to the aims of the websites.

On `reddit` the voting system takes precedence over any other feature. The manner in which votes determine visbility and focus users' attention on a small set of posts is `reddit`'s defining characteristic. Converesely, on `Stackoverflow` the voting system appears to perform in a more peripheral role. `Stackoverflow` is structured in a way which is designed to meet what we have deemed the website's primary purpose - generating answers to questions. The `Stackoverflow` 'Front page' is a reflection of this - showing users new questions which have not yet been answered, as opposed to 'the best' (high-scoring) recent questions. There are other tools which `Stackoverflow` employs to meet this purpose (e.g. wiki features - the ability of established users to edit or delete other users' contributions) and these are arguably more important to the functioning of the website than the voting system.

Stackoverflow's voting system serves the website's secondary purpose - organising the information so that it is more easily navigable by users. The tagging system also serves this purpose, and where the tagging system acts as a tool for categorising content the voting system produces a measure of the content's quality. There is evidence that scores produced by the voting system influence a question's visibility and the number of answers it will receive, but one gets the impression that its role in this regard is not absolutely vital to the website's success. In this sense `Stackoverflow`'s use of up/down voting is an interesting example of how this approach can be utilised in a tertiary role to augment a form of communication which is not expressly dedicated to assessing the quality of items.

This is part of what could be described as `Stackoverflow`'s strategy for protecting its Common Pool Resource (Ostrom, 1990) (the capacity to provide good answers to programming-related questions) from over-consumption and the tragedy of the commons. By building a navigable archive of previously answered questions users may find the answer to their own question without having to ask at all, and when a user asks a question which has already been answered they are often directed to the relevant post and their own new question is subsequently deleted. This ensures that areas of the site where the task of answering questions is performed are not cluttered with old questions, and enables the providers of answers to focus their energy on questions which are genuinely new.

It is however possible that `Stackoverflow`'s voting system serves a further purpose external to

determining an item's visibility but related to judging its quality. Firstly, the presence of positive scores beside answers may re-assure readers that the answers are of legitimate quality (with negative scores suggesting that answers should be ignored). Secondly, `Stackoverflow` uses a reputation system whereby the scores of a user's contributions are used to calculate the user's reputation score (and also whether they will receive certain 'badges'). The voting system may therefore serve an important role as a motivator on `Stackoverflow` - facilitating the production of quantitative feedback on a user's contributions. This type of feedback may in itself encourage users to keep submitting content, and its use in calculating reputation scores is only likely to enhance this effect. The motivations of `reddit` and `Stackoverflow`'s users to contribute are considered in the following Chapter.

Finally, the `Your Freedom` website serves as an example of a Distributed Moderation system which was not suited to its intended purpose - and which in many ways was *not fit for purpose.* In this case there was no clear relationship between the design of the system (method of voting, pages used to display content) and its expressed goals (identifying the most popular ideas).

Chapter 6

Users of Distributed Moderation systems

This chapter will consider `reddit.com` and the `Stack Exchange` network from the perspective of their users. According to Benkler (2006), motivation and organisation are the main barriers to the success of a Peer Production endeavour. The broad questions under consideration in this chapter relate to why and how the users of DM systems participate.

The Chapter begins with a profile of `reddit`'s users based on a number of surveys of `reddit`'s users which have been conducted - on `Stackoverflow` there is some demographic information in users' profiles but of this only age is easily analysed. One common assertion about Social News websites specifically (among their users) is that they suffer from groupthink - in `reddit`'s case this is often labelled the 'Reddit hive-mind' defined by a set of dominant attitudes and opinions among users which are expressed through the voting system such that contrary opinions have little chance of appearing in prominent locations. A precise definition of 'hive mind' as it relates to `reddit` is not available, but this is a term which one often sees being used on the website - see section 7.4 for a more detailed discussion of this concept.

If this 'hive mind' scenario is the case it would likely result in homogenisation of the user-base with respect to these attitudes and opinions - it is often assumed that users who hold minority opinions/attitudes will seek out an alternative website rather than persist in a space where their contributions are routinely down-voted. While no access to useful data on user opinions is available, it is possible to look both at user homogeneity in terms of demographic variables, and at whether the level of homogeneity is increasing.

The bulk of this chapter deals not with who the users of these websites are but *how* they use the websites. To address this question I consult procedural data from the websites themselves, looking

directly at records of user behaviour to build profiles of participation. The picture which emerges is one of specialisation, with users tending to focus on one form of activity. Particular attention will also be paid to new users, considering how these users learn to participate effectively and looking at the effect of feedback received through the voting system on a new user's behaviour. Empirical findings on user behaviour will then be related back to the question of *why* people participate on these websites.

6.1 Reddit

6.1.1 Who are `reddit`'s users?

When a user creates a `reddit` account they are not asked for any personal details, only a username and password (and optionally an e-mail address). Reliable information on the demographics of `reddit`'s users is therefore difficult to find. In the time that `reddit` has been under observation there have been several attempts to survey users with varying degrees of success. Most of these were surveys created by an ordinary `reddit` user on an external website (e.g. `SurveyMonkey`, Google docs), with the user then submitting a `reddit` post which linked to their survey.

There are various issues with the design of these surveys and some survey-specific problems are noted below in the relevant sections. Many items on the original surveys have been omitted because they are either of little interest (e.g. `reddit` users' favourite cheese, whether they prefer cats or dogs) or provided poor response options. The data generated by these surveys is nonetheless valuable in building a profile of who `reddit`'s users are.

`Reddit`'s users surveying the community

The first instance where this has been recorded occurred in February 2010 and was a three-stage failure. On Monday 8th February at 00:09 UTC a user submitted a post (redd.it/az92g) to the (now defunct) /r/reddit.com subreddit which linked to a survey they had created on '`QuestionPro.com`' - the post was heavily up-voted and appeared on the Front page with a score of around 1,400. The post's title suggested that users should check back in 2 weeks for results. The comments on this post were quite critical of the way the survey had been constructed (e.g. location options which included specific US states, Antartica and 'Europe'). Fourteen hours later the same user submitted a second post (redd.it/azh5d) titled "ATTENTION: Many people expressed feelings of misrepresentation on the survey. Here is survey 2.0. Hopefully it is better than the last one. Take it and check back on Feb 21 for results!". This post was also up-voted to the Front page with a final score of around 1,500.

Another five hours later the same user submitted a third post (redd.it/azl7h) titled "OKAY `reddit`. My sincere apologies. The survey site I used stopped collecting data 20,000 people. Here is the absolute final survey with REAL TIME DATA viewing! So sorry for yet another survey." This suggests that the previous survey had been completed by more than 20,000 people but no results were reported. At this stage `reddit`'s patience for repeated survey attempts seems to have worn thin, the third post only received a score of 35 points and it appears that those users who completed this iteration of the survey often submitted nonsense responses. The user who created these surveys and posts then appears to have deleted their `reddit` account and no results were ever reported for the first two iterations of the survey.

The next attempt to survey `reddit`'s users which has been recorded did actually produce results. This survey was created by the `reddit` user JavaLSU (written by himself in `PHP/MySQL` and hosted on his own server) and a post linking to it (redd.it/dmkjj) was submitted to /r/AskReddit on 4th October 2010 at 11:54 UTC. It was up-voted to the Front page with a final score of around 2,500. The questions on this survey largely concern demographic information and it was completed (at least partially) by 25,849 respondents. JavaLSU has kindly provided full access to the data generated by this survey.

The next user-produced survey to appear on the Front page and receive many responses was created by the user burgess_meredith_jr using `SurveyMonkey` and submitted to the /r/reddit.com subreddit on 28th April 2011 (redd.it/gzb2w). The execution of this survey could have been better, a number of problems are outlined by burgess_meredith_jr in a comment on the post. Firstly, this user did not realise they would have to pay to access responses beyond the first 100 so for a period of time it seemed that the data would be lost. However, Reddit's administrators stepped in and paid the $330 required to gain access to the data. burgess_meredith_jr later provided a summary of responses in a comment and this is the only aspect of the data which we have access to.

On July 13th 2011 `reddit`'s administrators launched their own survey attempt (jenakalif, 2011), giving credit to burgess_meredith_jr for inspiring this. This survey was completed by 32,756 users and some months later the individual responses were made publicly available.

We will now consider what these surveys can tell us about `reddit`'s users, concentrating on JavaLSU's survey in October 2010 and the `reddit`-run survey in July 2011 as full access to data is available - with a brief summary of burgess_meredith_jr's survey of April 2011. The sampling design of these surveys is atypical and would best be described as an unusual form of convenience sample, any individual could complete the survey (multiple times if they wished) and the set of potential respondents will have been defined by where the survey appeared. Here we consider surveys which appeared on the default Front page so one can assume that a considerable percentage of individuals who were browsing `reddit` at the time were aware that the survey existed and chose whether or not to participate. However, it is important to note the subreddit which each survey

was submitted to, because users who had unsubscribed from that subreddit would not have seen the survey unless someone else showed it to them. It is hoped that considering data from several surveys conducted at different times and through different subreddits will circumvent any problems caused by the unusual sampling method.

JavaLSU's survey - October 2010 - submitted to /r/AskReddit

JavaLSU's survey was submitted to /r/AskReddit on 4th October 2010 at 11:54 UTC - it was completed (at least partially) by 25,849 respondents - with the great majority of responses coming within the 24-hour period after this post was created. There follow bar-plots of some of the variables on which data were collected. Some editing of responses has taken place (some categories have been collapsed, and for questions with many possible responses those responses with lower than a given frequency have been labeled 'Other'). There follow bar-plots for a selection of variables from the survey.
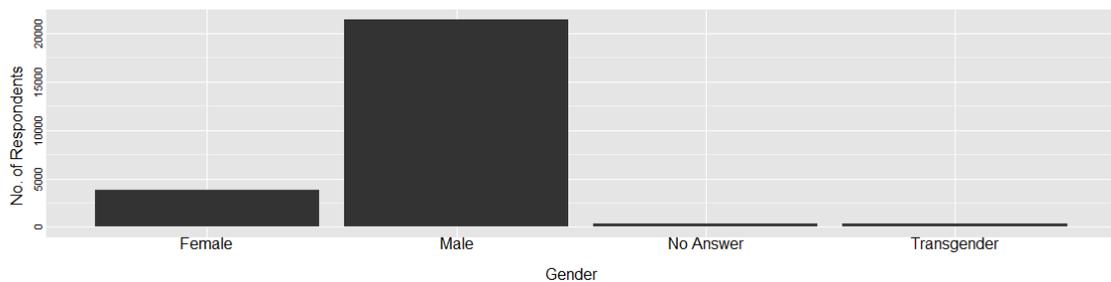
Figure 6.1: Gender of respondents to the JavaLSU survey - N = 25,849
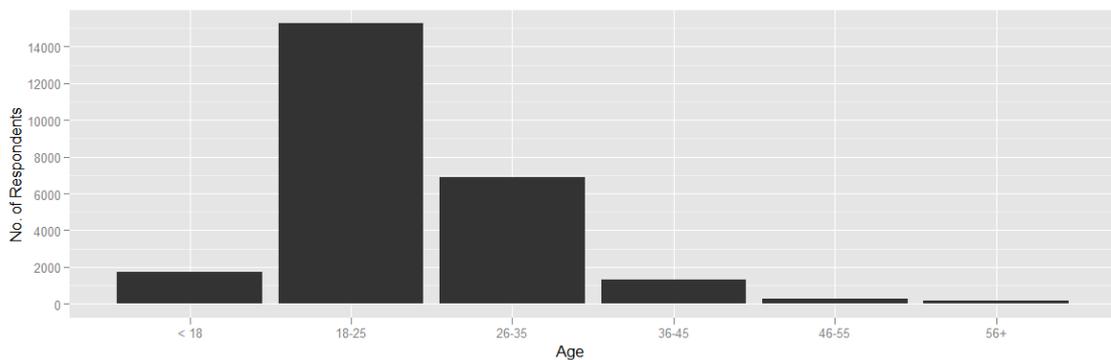
Figure 6.2: Age of respondents to the JavaLSU survey - N = 25,554

It is immediately clear that many of the items on this survey had one dominant response. 85% of respondents are male, 60% fall into the 18-25 age category, 68% live in the United States and 83% are white. 7,140 respondents in total (28% of the sample) are white males aged 18-25 living in the United States - if we expand the age range to 18-35 this demographic accounts for 40% of the sample. In terms of these demographic variables `reddit`'s users are a surprisingly homogenous
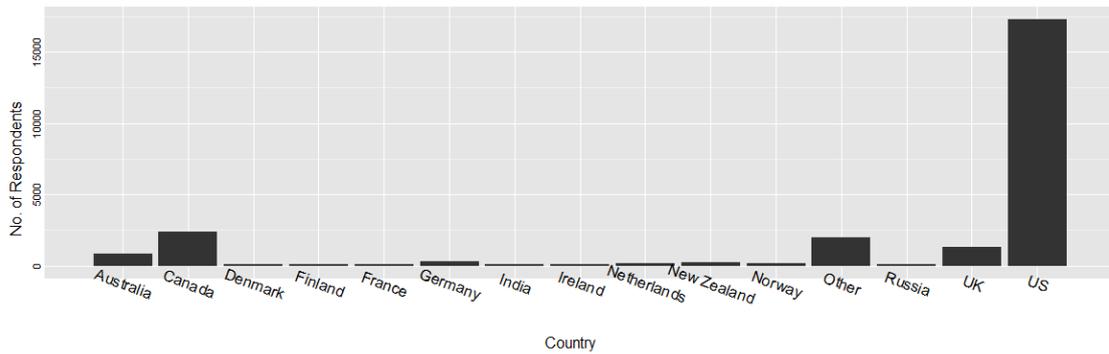
Figure 6.3: Country selected by respondents to the JavaLSU survey - N = 25,598. Countries which were selected by fewer than 100 respondents have been grouped as 'Other'.
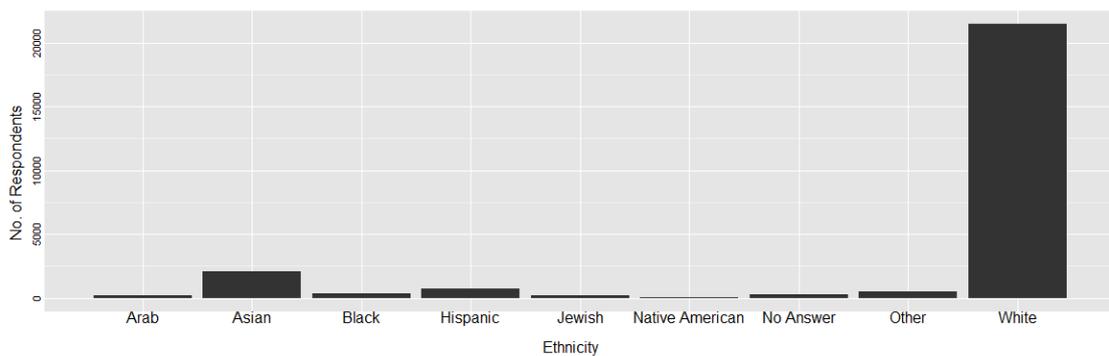
Figure 6.4: Ethnicity of respondents to the JavaLSU survey - N = 25,849

cohort.

If we consider the countries where respondents live, language seems to be the dominant factor, the top four countries (US, Canada, UK, Australia) all share English as their dominant language (figure 6.3) - and `reddit` is a website where the English language dominates. Taking the countries `reddit` users hail from into account one might have expected more respondents to select Christianity as their religion (figure 6.5) - although the prevalance of Atheism and Agnosticism (in combination accounting for 67% of the sample) will likely not come as a surprise to anyone who has spent time browsing `reddit`. The same could be said for the political leanings of respondents (figure 6.6) - with very few respondents choosing to identify as Republican. American respondents are also more likely to live in 'Blue states', although a fair proportion live in states which Romney won in the 2012 presedential election (figure 6.7).

A clue as to why so many respondents fall into these categories may be found in an unusual place, the item concerning preferred web browser. `StatOwl.com` gives the following figures for web browser usage in the USA in October 2010 - 63% for Internet Explorer, 20% for Mozilla Firefox, 7% for Google Chrome and 9% for Safari. For respondents to the JavaLSU survey conducted in the same month these figures are - 1% for Internet Explorer, 46% for Firefox, 44% for Chrome
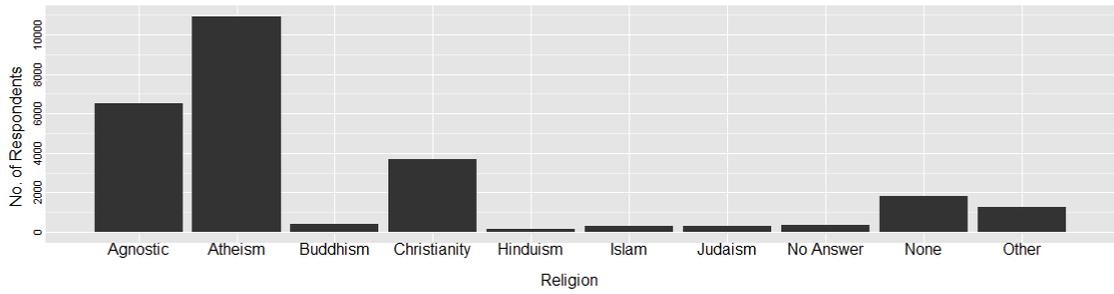
Figure 6.5: Religion of respondents to the JavaLSU survey - N = 25,849. Religions selected by fewer than 150 respondents have been grouped under 'Other'.
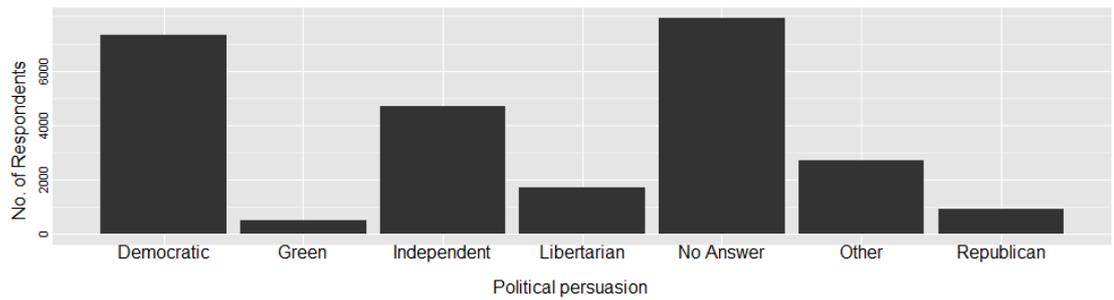
Figure 6.6: Political persuasion of respondents to the JavaLSU survey - N = 25,849.

and 5% for Safari. Reddit's users differ from the general population in that they do not use Internet Explorer. Casual observation of `reddit` does suggest a technologically savvy user-base, and perhaps the choice of a browser other than Internet Explorer lends some support to the idea that `reddit`'s users are more technologically savvy than the general population.
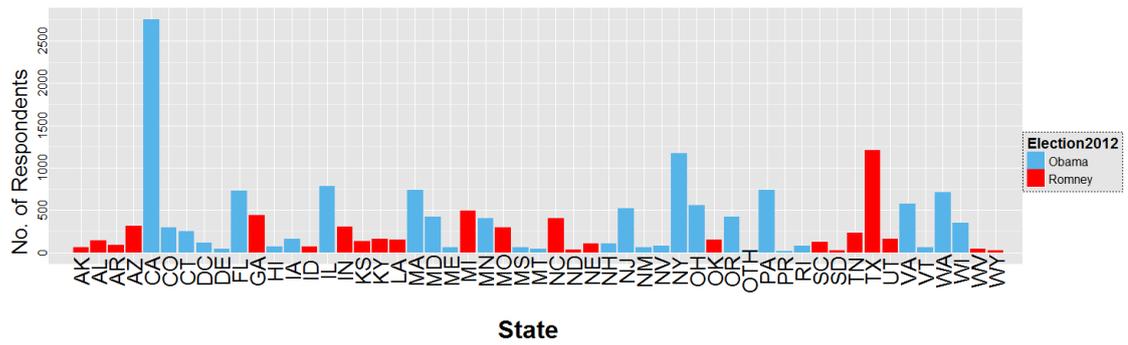
Figure 6.7: State selected by respondents who chose United States as their country, N = 17,511. States selected by fewer than 10 respondents have been grouped under 'OTH'. Bars have been coloured to reflect whether Obama or Romney won the state in the Presedential election of 2012.
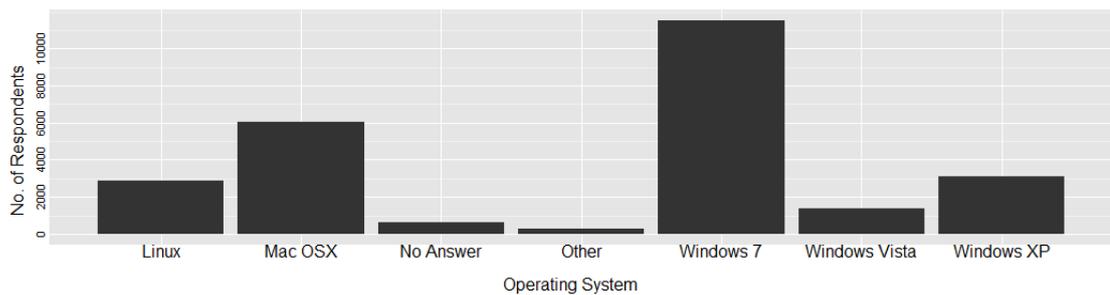
Figure 6.8: Operating System of respondents to the JavaLSU survey - N = 25,849.
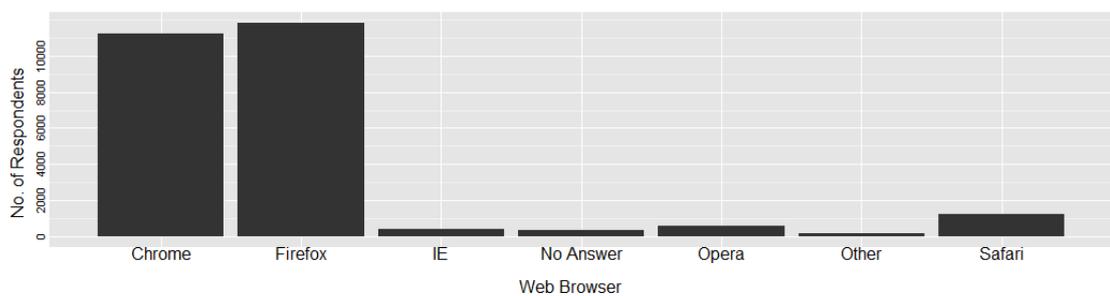
Figure 6.9: Web Browser of respondents to the JavaLSU survey - N = 25,849.

burgess_meredith_jr's survey - April 2011 - submitted to /r/reddit.com

The figures reported for this summary are taken from burgess_meredith_jr's comment summarising the results after 26,887 responses had been collected. Table 6.1 reproduces selected descriptives. There are some obvious issues with this data - for instance the overlap of Age categories. Respondent location was also collected but it is problematic because it was collected using free text entry and burgess_meredith_jr has manually calculated percentages based on the first 500 responses - reporting 80% for the US, 10% for the UK and 4% for Canada and Australia.

It was decided that this survey was worth consideration primarily because it collected information on how long respondents had been using `reddit`, a variable not measured in other surveys. Responses to this question revealed that many respondents had been using `reddit` for quite a short period of time. A question on household income also revealed that 44% of respondents presently classified themselves as a student. Responses to questions of Gender and Age showed congruency with those for the JavaLSU survey.

Question	Response	%
Gender	Male	84
	Female	16
Age	Under 18	6
	18-25	56
	25-35	29
	35-45	5
	45-65	2
	65+	0.5
Politics	Centrist Left	54
	Socialist	22
	Centrist Right	13
	Anarchist	5
	Conservative Right	2
	Ultra Conservative	1
	Communist	1
	Fascist	1
Time as Redditor	'Since the beginning of time'	1
	5 years	2
	4 years	4
	3 years	9
	2 years	18
	1 year	30
	6 months	17
	Less than 6 months	18

Table 6.1: Percentage of the 26,887 respondents to burgess_meredith_jr's survey who selected each response option.

Official `reddit` Survey - July 2011 - submitted to /r/blog

On July 13th 2011 the `reddit` employee jenakalif published a post to the `reddit` blog which asked users to complete a survey (largely concerning basic demographic information). A link to this post was submitted as a post to the /r/blog subreddit (where only Admins can submit posts) and from here appeared on the Front page with an eventual score of around 1,700. This survey was completed by 32,754 respondents. There follow bar-plots of some of the variables on which data were collected.

For items which appeared on both surveys (e.g. Gender, Age, Country) there is a strong similarity in the profile of responses - 81% of respondents are male and 70% live in the United States. In terms of age there is a slightly different profile for the `reddit` survey responses, with 48% of respondents in the 18-24 category and 35% in the 25-34 category. This could be accounted for by the fact that the categories are themselves slightly different (with an Age of 25 falling into different bands) - or that the survey was circulated through a different part of the website (/r/blog rather than /r/askreddit). For the `reddit` survey 41% of respondents are Males aged 18-34 living in the United States - a similar percentage to the JavaLSU survey where White ethnicity was also specified.

The relatively high levels of similarity for items which appeared on both surveys, despite these being seperated in time by 9 months and appearing in different parts of the website, suggests that this method of recruiting respondents is reliable. This consistency also speaks to an important question about `reddit` and Social News websites more generally. The mechanisms which these websites are built on (i.e. a feedback loop where popular items are 'broadcast' and in a sense define the website's identity) would seem to tend towards a user-base which becomes more homogenous over time - if a new user or potential user disagrees with popular opinion on `reddit` they are unlikely to find much value in the website. Comparing these surveys suggests that `reddit`'s users are not becoming more homogenous in terms of the measured demographic variables (e.g. the percentage of female respondents has increased as compared to the JavaLSU survey). It is however still possible that `reddit`'s user-base is becoming more homogenous with respect to their opinions, none of the items included on both surveys can speak to this.
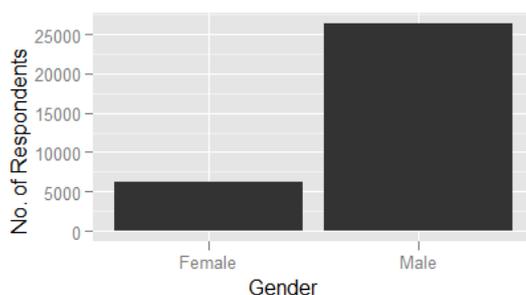
Figure 6.10: Gender of respondents to the `reddit` survey - N = 32,553
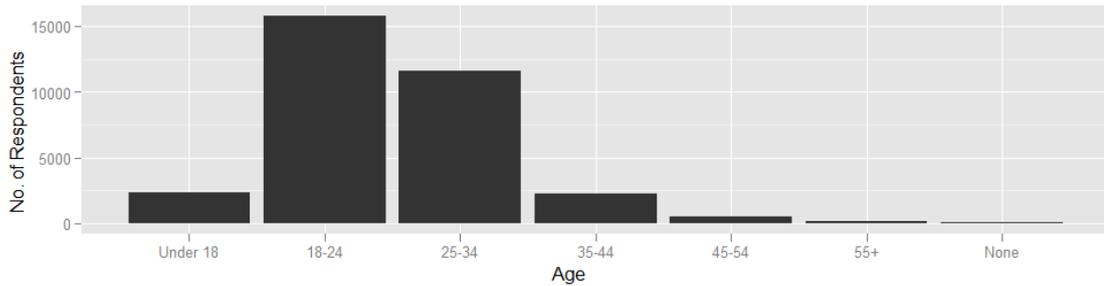
Figure 6.11: Age of respondents to the `reddit` survey - N = 32,754
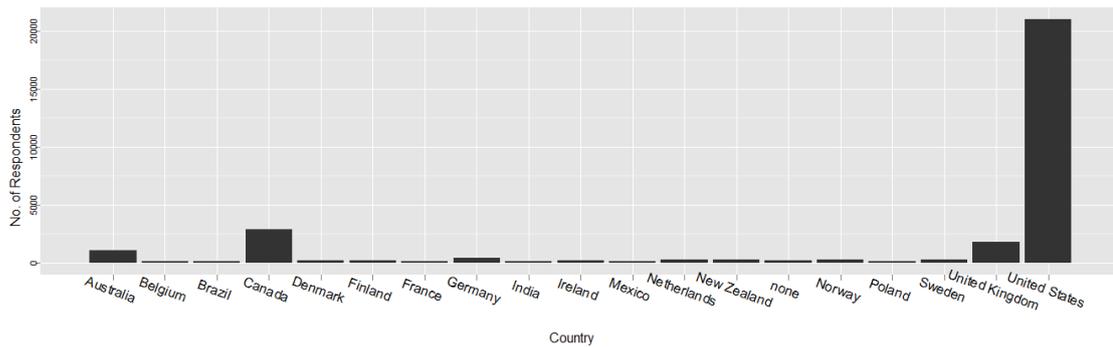
Figure 6.12: Country selected by respondents to the `reddit` survey - N = 29,817. Countries which were selected by fewer than 100 respondents have been grouped as 'Other'.

The `reddit` survey included a number of items which the JavaLSU survey did not (with the converse also being true). Of these, Household income (figure 6.14) and Education level (figure 6.15) are of particular interest. Figure 6.15 suggests that 48% of respondents have a Bachelor's or Graduate degree - the US Census bureau reported a national figure of 30% having a bachelors degree or higher in 2011. There are more respondents claiming a Graduate degree than there are respondents whose eduction finished after High School. For another item concerning occupational status 40% of respondents indicated that they were currently a student. Reddit's users appear to be relatively well-educated.

Finally, perhaps one of the most interesting question asked by the `reddit` survey concerned a user's favourite subreddit. This item consisted of the question "What is your favorite `reddit`? (Please no r/ prefixes.)" and a text entry box - so users were free to type any subreddit. Free text entry resulted in some messiness in the data and although a level of cleaning has been performed there are still some cases where subreddits were mis-spelled. There is also a great variety of subreddits which have been given as a user's favourite, here we will only consider subreddits chosen as favourite by at least 100 respondents. Subreddits which were chosen by less than 100 respondents or mis-spelled account for 8,570 responses.

Figure 6.16 shows the number of respondents who selected as favourite each of the remaining
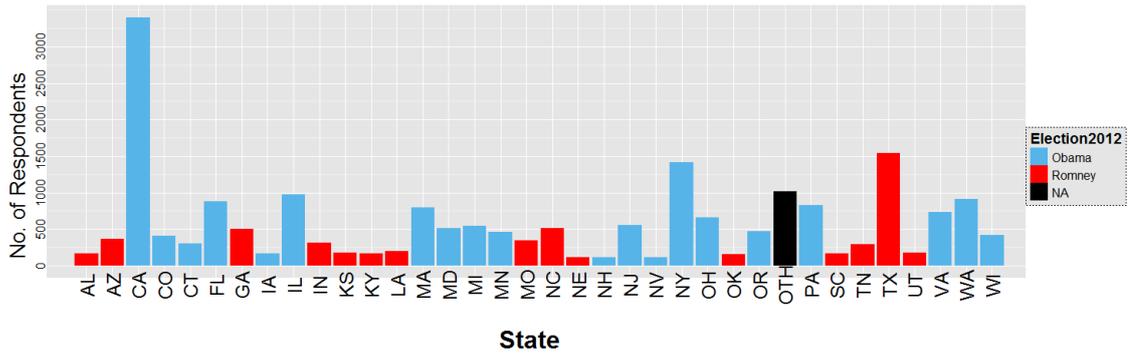
Figure 6.13: State selected by respondents who chose United States as their country, N = 20,846. States selected by fewer than 10 respondents have been grouped under 'Other'. Bars have been coloured to reflect whether Obama or Romney won the state in the Presedential election of 2012.
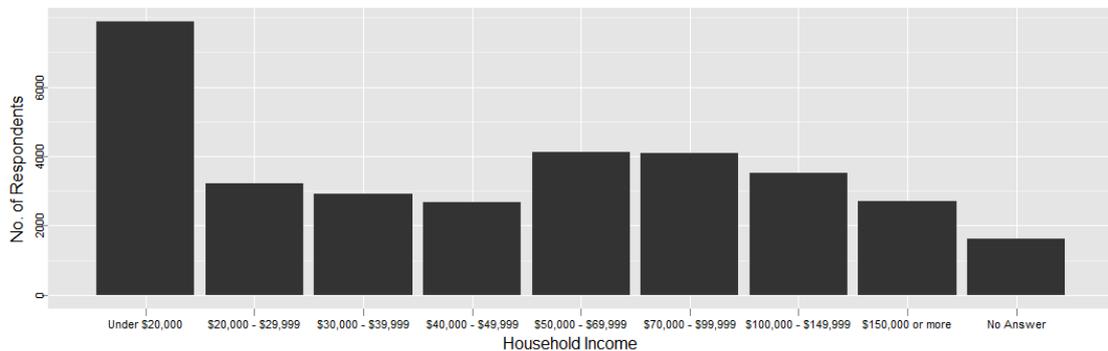
Figure 6.14: Household Income of respondents to the `reddit` survey - N = 32,754.

subreddits. These subreddits have been manually classified as falling into one of six categories. 'Default' subreddits are the subreddits which were defaults at the time when the survey was conducted, this category takes precedence over all others. 'Game' subreddits relate to specific computer games. 'Over 18' subreddits cater to adult content and a user must click through an Over-18 disclaimer to access content from these subreddits. The 'Page' category exists because some users entered page types instead of subreddits, perhaps mis-understanding what a 'reddit' was in this context. 'UGC', or User Generated Content, is a category of subreddits which cater to posts 'produced' predominantly or exclusively by `reddit` users themselves. Some of these subreddits cater to text posts written by users (e.g. on /r/nosleep users submit ghost stories they have written, on /r/askscience users submit science-related questions and answers), with some variety in the nature of conent on other subreddits (e.g. 'rage comics' on /r/ffffffuuuuuuuuuuuuu). The final category, 'Non-default', includes all of the subreddits which were not defaults at the time and which do not fit neatly into any other category.

It is no surprise that the Default subreddits were entered as favourite most often with 41% of respondents choosing a default subreddit as their favourite, one can assume that exposure to these

Figure 6.15: Education level selected by respondents to the `reddit` survey - N = 32,610.

Figure 6.16: Favourite subreddit chosen by respondents to the `reddit` survey (free text entry) - N = 19,849. Subreddits entered by fewer than 100 respondents are not included - 8,570 responses have been excluded in this manner.

subreddits is highest among `reddit` users. It is perhaps a little surprising that default subreddits did not account for a greater percentage of 'favourite subreddit' considering the level of activity on the default subreddits. Subreddits which cater to User-Generated Content account for 26% of users' favourite subreddits - and some popular default subreddits would also fall into this category (e.g. /r/Askreddit, /r/Iama). Figure 6.16 suggests that many respondents do not see `reddit` as being primarily a source of 'News' from external sources which has been aggregated - but are more interested in content which has been produced *and* aggregated by other users.

It is unfortunate that these surveys which were completed by very large numbers of respondents did not ask more questions about how and why the individuals used `reddit`. A number of surveys have asked such questions but they have tended not to appear on the Front page and consequently have not received anything approaching the same level of response. These surveys will be considered as part of section 6.1.7.

6.1.2 User activity levels

It is important to note at the outset of this section that there is a type of user who will not appear in any of the analyses of user activity. 'Lurkers' are people who visit or read the website without necessarily having a user account and, by definition, without participating by submitting or voting on content. It is difficult to estimate how many of `reddit`'s 'users' fall into this category, but one suspects that it is a considerable proportion. For instance, in section 5.1.4 we saw that an /r/pics post which appeared on the Front page was receiving hits at a rate which was around 25 times higher than the rate at which it was receiving votes - to put this another way only around 4% of the people who saw the picture voted on its `reddit` post, the other 96% must incorporate both lurkers and those users who saw the picture and decided not to vote on it. Although lurkers do not interact with the voting system directly it must be assumed that, collectively, they have an affect on `reddit` by serving as an audience for the activity which takes place there. If one conceives of `reddit` as a mechanism for management of its Common Pool Resource (Ostrom, 1990) (human attention) - these 'lurkers' are actually valued contributors, they provide their own attention and do not attempt to gain attention for their own submissions. In this work lurkers are sometimes referred to as readers or visitors, and the label of 'user' is often reserved for those who have and use a `reddit` account.

This section relies exclusively on the back-end procedural data for March 2009 provided by a `reddit` adminsitrator for research purposes. Voting is an anonymous activity on `reddit` and therefore there is little information of relevance here which can be extracted through `reddit`'s API. In March 2009 on the `reddit.com` website 102,232 users cast 3,346,062 votes on 352,902 items of content (or 'posts'). Of these users 33,589 (33%) only acted once (one vote or submission), whereas the most active user for the month registered 23,776 actions. Figures 6.17, 6.18 and 6.19 illustrate the highly skewed distribution of activity between `reddit` users.

When the distributions of votes and submissions between users are presented on logarithmic axes it is clear that they do not follow the straight line which would be expected from a typical power law. Figure 6.18 shows quite a pronounced 'hump' in the distribution of votes between users. This suggests that users with more than around 250 votes have made more votes than would be expected under a power law, and that these very active users may warrant particular attention. There were 2,690 users who made 250 or more votes, and between them this group (2.6% of active users) accounts for 44% of all the votes cast in the month.
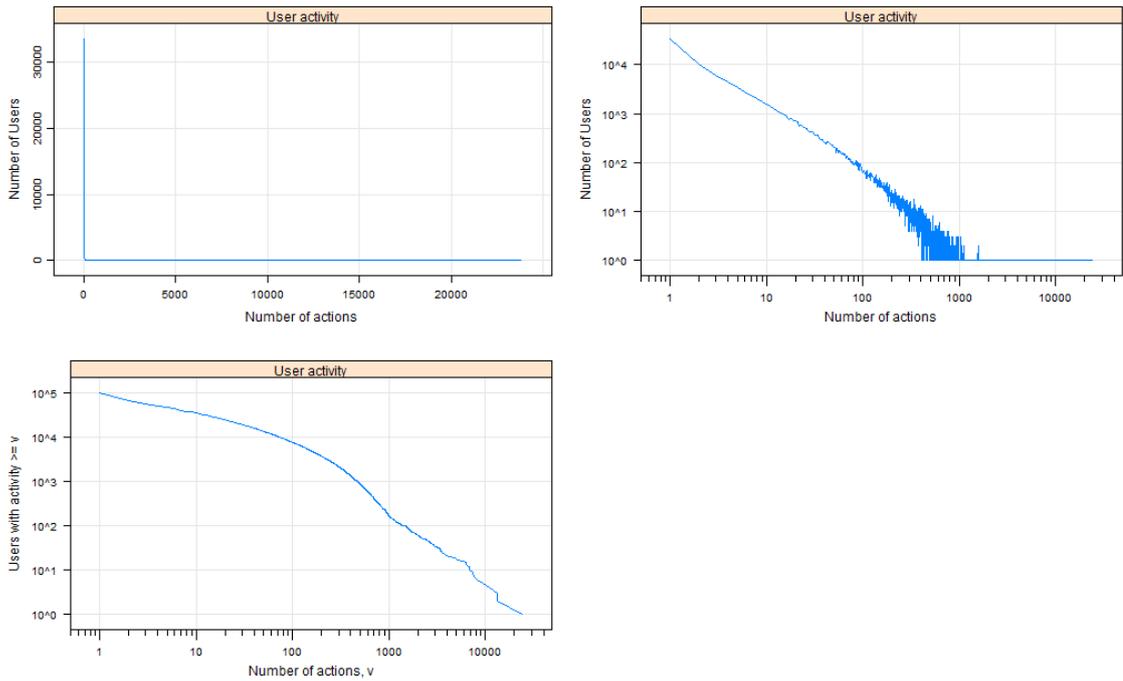
Figure 6.17: The distribution of activity between `reddit` users. Top-left: Raw data; Top Right: Data presented on logarithmic axes; Bottom Left: Inverse cumulative distribution on logarithmic axes
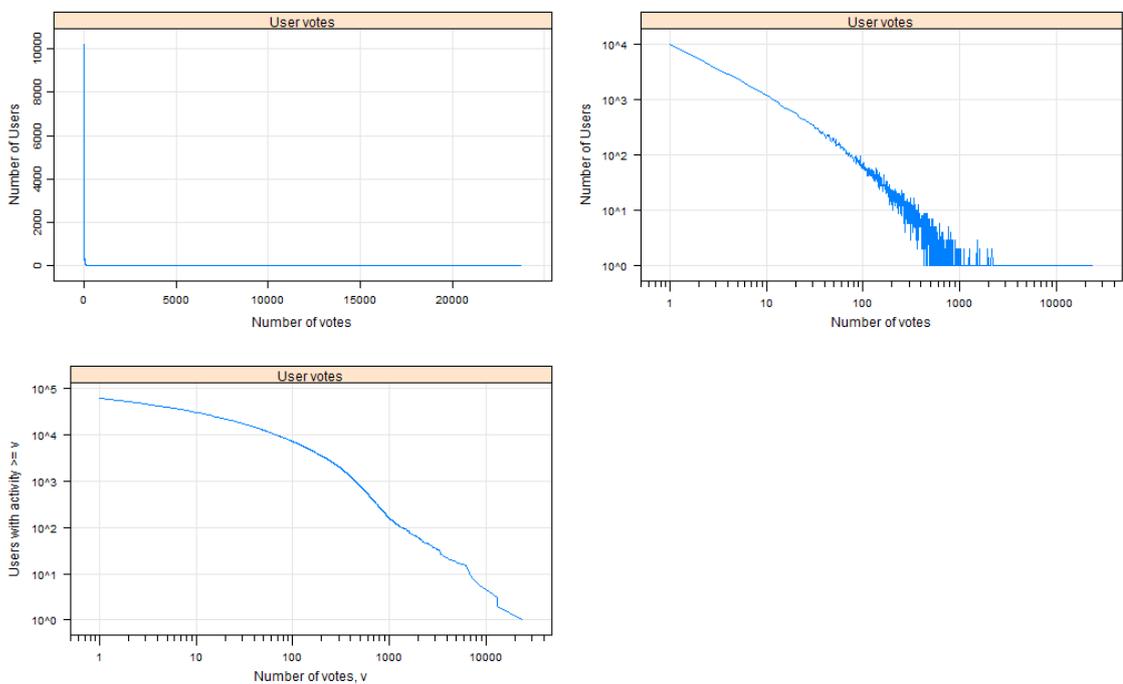
Figure 6.18: The distribution of votes between `reddit` users. Top-left: Raw data; Top Right: Data presented on logarithmic axes; Bottom Left: Inverse cumulative distribution on logarithmic axes
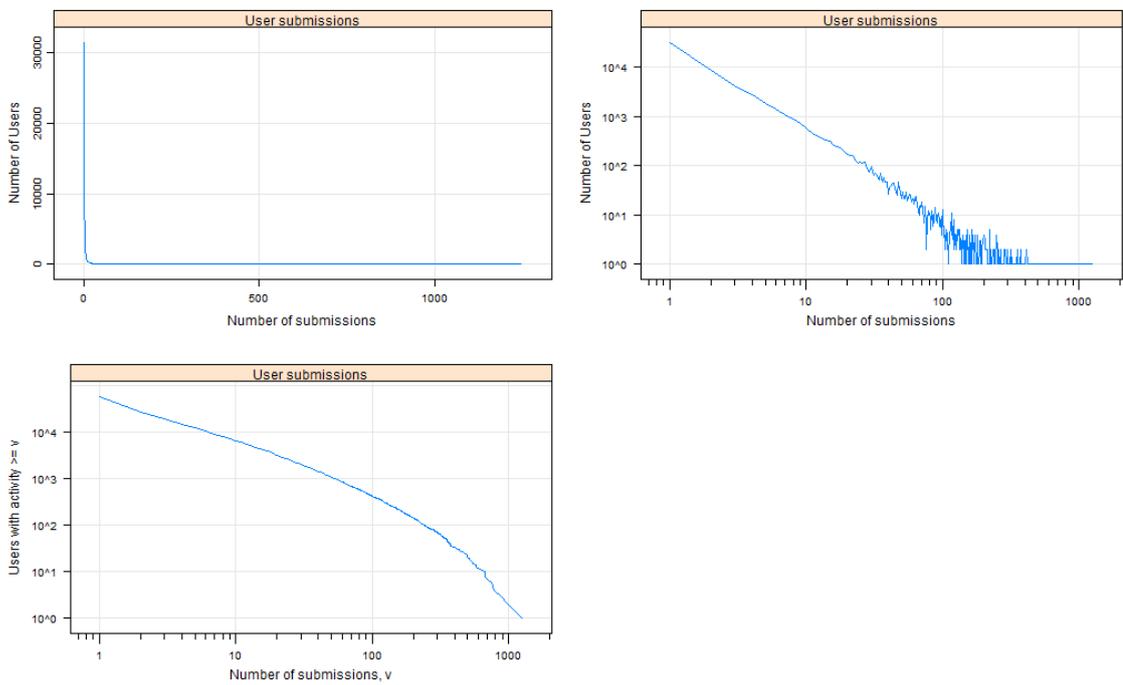
Figure 6.19: The distribution of post submissions between `reddit` users. Top-left: Raw data; Top Right: Data presented on logarithmic axes; Bottom Left: Inverse cumulative distribution on logarithmic axes

6.1.3 User activity types

`Reddit`' users show considerable variability in their level of participation on the website - do they also exhibit variability in the types of activity they engage in? Cursory examination of the 102,232 users who were active in March 2009 suggests that they do. 80,211 of these users (78.5%) have engaged in just one of the two forms of activity recorded (posting and voting). The 33,589 users who only acted once were bound to fall into this category, if these users are excluded this leaves 68,643 users who acted more than once, and 68% of these users have only engaged in one form of activity. Of these users who exclusively vote or submit 53% are voters and 47% are submitters.

Does a user's tendency to engage in one form of activity to the exclusion of the other vary with their level of activity? Are more active users more likely to be voters or submitters? If we compare very active to moderately active users we might expect those who are very active to be more rounded participants, submitting posts and also voting on the posts submitted by other users. We might also expect more active users to engage more in voting than submitting, as voting is less effortful and it is easier to register a large quantity of votes than it is to make a large quantity of submissions. To address these questions users were divided into 14 'bins' based on their activity level for the month.
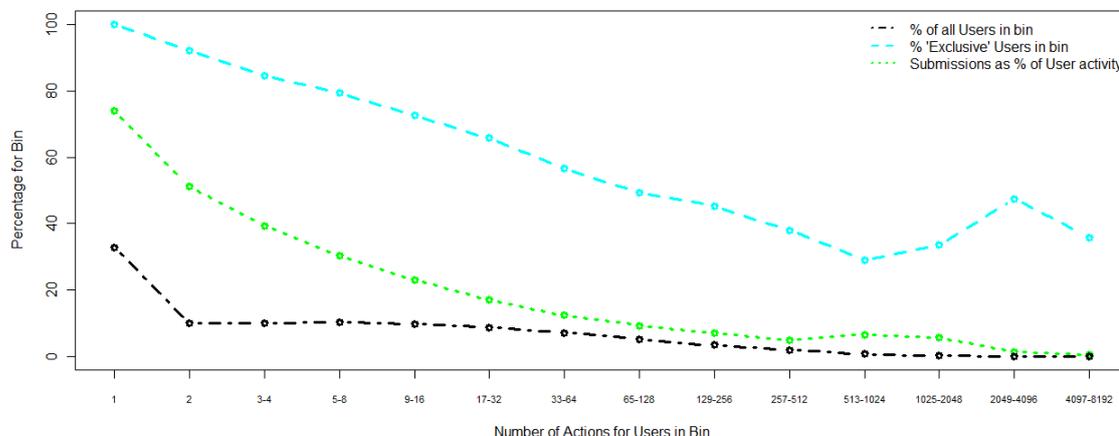
Figure 6.20: Showing the percentage of all Users who fall into each activity bin, the percentage of exclusive users within a bin who engaged in either voting or submitting but not both, and the average proportion of actions which are submissions for users within a bin.

Figure 6.20 indicates that the above-mentioned trends are present in the data. Users who are less active are more likely to engage in one form of activity exclusively, and more likely to participate by submitting than voting. However, the proportion of highly active users who only engage in one form of activity is surprisingly high. The percentage of 'exclusive' users within activity bins only drops below 35% for the bins which represent users with between 512 and 2,048 actions, and actually increases again for the bins representing users with more than 2,048 actions. Many of

214

the most highly active `reddit` users only participated by voting, they did not make a single post submission within the month considered here.
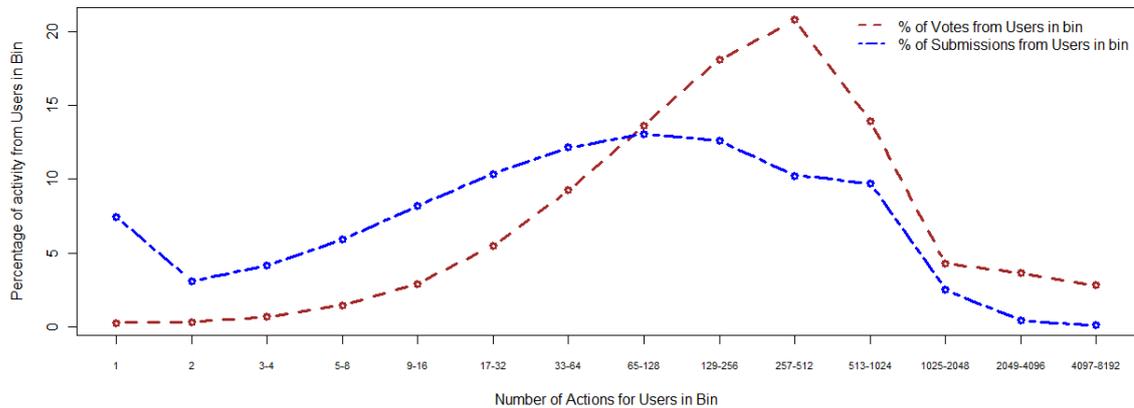
Figure 6.21: Showing the percentage of global votes and submissions which originate with users in each activity bin.

Figure 6.21 shows where `reddit`'s submissions and votes come from, in terms of the activity levels of the users who contribute them. Most of `reddit`'s votes come from moderately to highly active users, whereas submissions are more evenly distributed between low and high-activity users. To illustrate how the power law manifests on `reddit`; the 14 users who fall into the 4097-8192 actions bin together made more votes than the 62,890 users with between 1 and 8 actions. On the other hand, those users with between 1 and 8 actions made 183 times as many post submissions as the users with 4097-8192 actions. If we only consider the volume of submissions and votes, it would appear that many of these originate with two distinct groups of users.

6.1.4 Users who vote

`Reddit` allows users to up-vote or down-vote items of content; do users who vote make both up and down votes, or prioritise one form of voting over the other? In the month for which voting data are available, 77 % of all votes cast were positive, with 23 % being negative. Up-voting occurs more frequently on `reddit` than down-voting.

Figure 6.22 shows that users who are more active are more likely to use a mixture of up and down votes. For most of the activity bins considered, users make an average of 20% negative votes. However, highly active users (those with more than 512 actions in the month) show a stronger tendency to vote negatively. This provides an initial indication that highly active voting users may be acting in a manner which is distinct from their less active peers. To pursue this further the timing of users votes is considered.
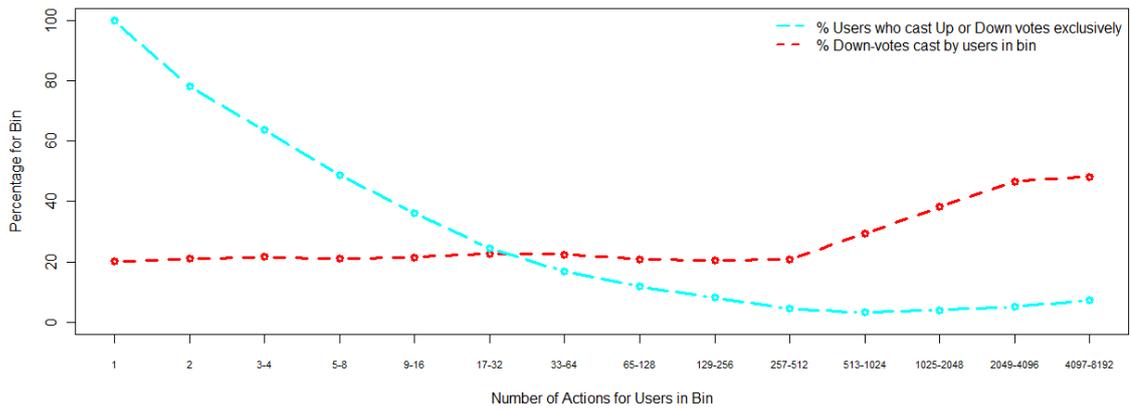
Figure 6.22: Showing the percentage of voting Users in each bin who cast Up or Down votes exclusively, and the percentage of votes cast by voting users in each bin which are Down-votes.

Early Votes

Every new post submitted to `reddit` appears initially on the New page, where recent submissions are displayed in reverse chronological order. Whether a post is displayed on any other page is determined by the voting response it receives from users. To stand a chance of being successful and reaching `reddit`'s Front page, a post must appear on another page (most likely the 'Rising' page) before it is pushed off the New page by subsequent submissions. It was shown in Chapter 5 that a post's early voting reception (within the first 30 minutes) has a relationship with its chance of reaching the Front page; and that the level of voting activity on the New page is very low as compared to other pages. Other research on `Digg` (Lerman and Hogg, 2010) has likewise suggested that early voting for a post is a strong predictor of ultimate success.

To a user who understands this aspect of Social News it suggests a possible strategy to maximise the effectiveness of ones' votes. Suppose a user sees a post which they do not like on the website's Front page. The user can down-vote this post, but the power of their vote will be diluted by the hundreds of other votes being cast every hour on Front page posts - their down-vote is unlikely to lead to the post's removal from the Front page. Suppose the same user sees this post when it is on the New page, where posts tend to receive votes at a much slower rate; here a down-vote is more likely to make the difference between that post slipping from the New page into obscurity, or going on to appear on the Rising page (and from there having a chance to reach the Front page).

This can be thought of as the first obstacle between a new post and the Front page. Every post is given the same chance on the New page, but placement on subsequent pages must be earned through the accumulation of a positive score. Are there users who operate specifically in this capacity - filtering out 'poor' submissions before they have a chance to reach higher visibility pages of the website? Figure 6.23 shows that posts are much more likely to receive down-votes early in their lifespan, and that many of these early down-votes come from a small number of users.
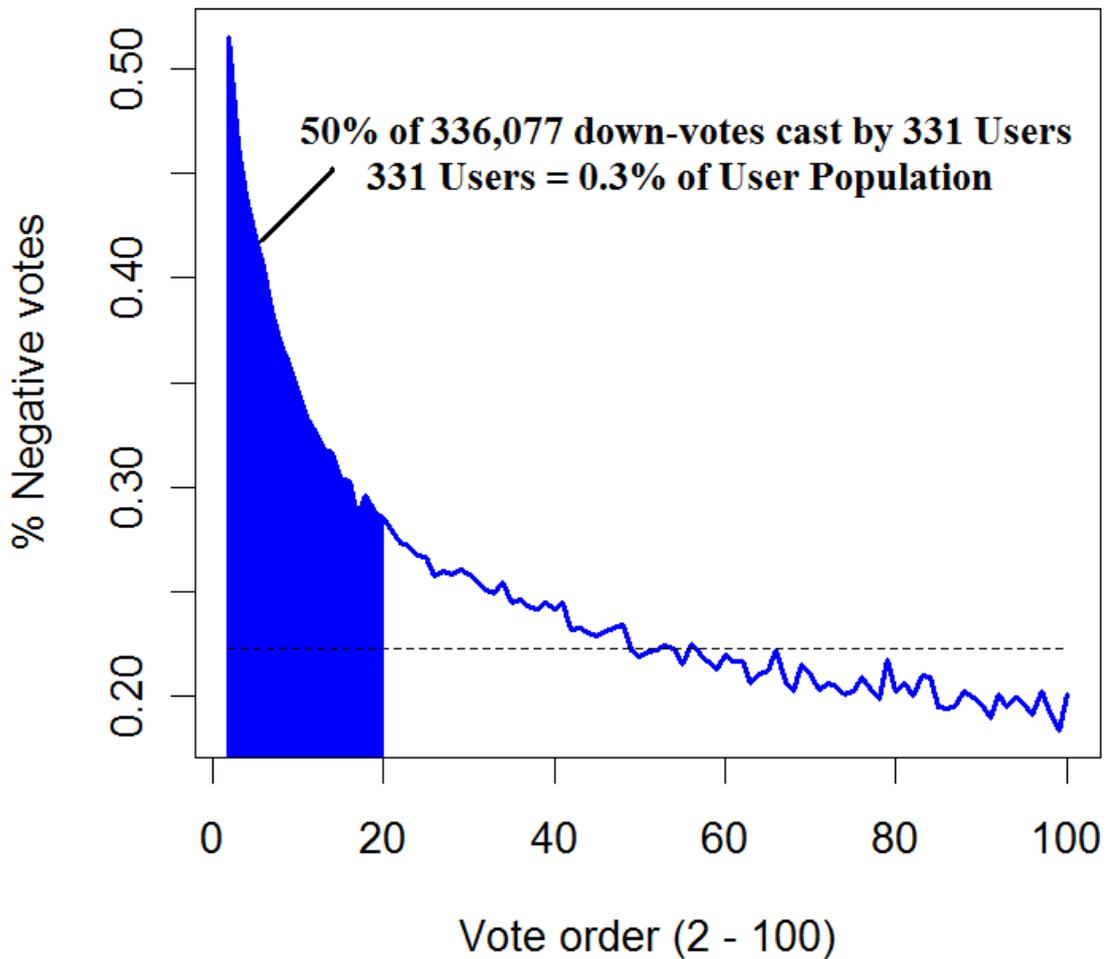
Figure 6.23: Showing the proportion of votes which are negative (down-votes) as a function of post voting order. Votes with order 2 are the first votes to be cast on a post by a user other than the submitter. The dashed line represents the global site average of 23 % negative votes. The coloured area represents the 336,077 negative votes which were among the first twenty to be cast on a post, 50 % of these early negative votes originate with just 331 users.

Collectively, these users appear to be imposing an early negative voting filter on new submissions. Are these users acting altruistically or in a self-interested fashion? `Reddit` receives thousands of submissions every day, users who vote early could do so altruistically because they wish to exclude low-quality submissions quickly. In fact it would be expected that early votes are more likely to be down-votes, because low-quality posts which receive mostly down-votes at first will not receive any votes with a higher vote order. To put this another way, a post which appears on the Rising page has already passed through a filter of sorts - more people up-voted than down-voted it while it was on the New page.

Users might also vote early because they wish to maximise the impact of their voting activity on the collective decision about which posts reach the Front page. A user might wish to filter out posts which they find personally objectionable (e.g. on political grounds), in addition to posts

which are of generally low quality. Irrespective of the motivation of these users, they appear to collectively wield the power to 'veto' a post - they can ensure that it does not progress further but cannot dictate whether it appears on the Front page.

There is some anecdotal evidence which suggests that `reddit` users are willing to vote on new posts to improve the functioning of the website. In May 2011 the pics subreddit began printing a message at the top of its pages which asked users to vote more on the New page. This page was observed and records created in February 2011 and June 2011; in February posts received a mean of 5.3 votes the last time they were observed on the New page, by June this figure had increased to 7.2. Between February and June the rate of new submissions to this subreddit also increased - the posts submitted in June received more votes on the New page *despite appearing there for less time on average*. This suggests that at least some `reddit` users are willing to vote on new submissions when requested, however it does not speak to the motivations of users who were engaging in this behaviour before it was called for by moderators.

Quick Votes

`Reddit` allows users to vote as often as they please (although only once per post) - a user who wishes to exert maximal influence on the decision-making process could do so by voting more frequently than other users. This could be achieved by spending more time on the website perusing content and voting, and/or by voting at a faster rate while browsing content on the website. The votes of a user who scans the titles of posts and votes quickly based on these titles alone will bear the same weight as the votes of a user who devotes several minutes to assessing the quality of a post before voting.

Inspection of voting data revealed that 652,873 votes (19.5 % of all votes which were not a user's first vote in the month) were cast within 10 seconds of the user's previous vote, and 283,907 votes (7.6 % of all votes which were not a user's first vote in the month) were cast within 2 seconds of the user's previous vote. These quick votes have likely been cast on the basis of a post's title alone, ten seconds seems scarcely enough time to open a post, consider its content, and decide which way to vote. On certain subreddits (e.g. those dealing with pictures) it may be possible to consider a post's quality within ten seconds. Inspection of the subreddits where these quick votes (within 10 seconds) occur revealed that the percentage of votes on a subreddit that are 'quick' ranges between 14 and 28 %. However, the 'pics' subreddit (dealing exclusively in pictures) actually has a lower percentage of quick votes (19 %) than subreddits which deal with written material (e.g. 'politics' at 23 %, 'science' at 25 %).

It is also possible that `reddit`'s user interface plays a part here - if a user follows a post's hyperlink and spends several minutes assessing the quality of the linked-to resource they may well find

that when they return to the `reddit` page to vote on this post it has moved and is not easy to find so that they can cast their vote. It is likely that users with more experience have already found solutions to this problem - either by installing an add-on like the `reddit` Enhancement Suite (which superimposes a bar at the top of a page which was linked to by `reddit` - showing the post's title along with buttons whereby a user can up or down vote from that page) or by opening the linked-to resource and the comments page for the post in seperate browser tabs.
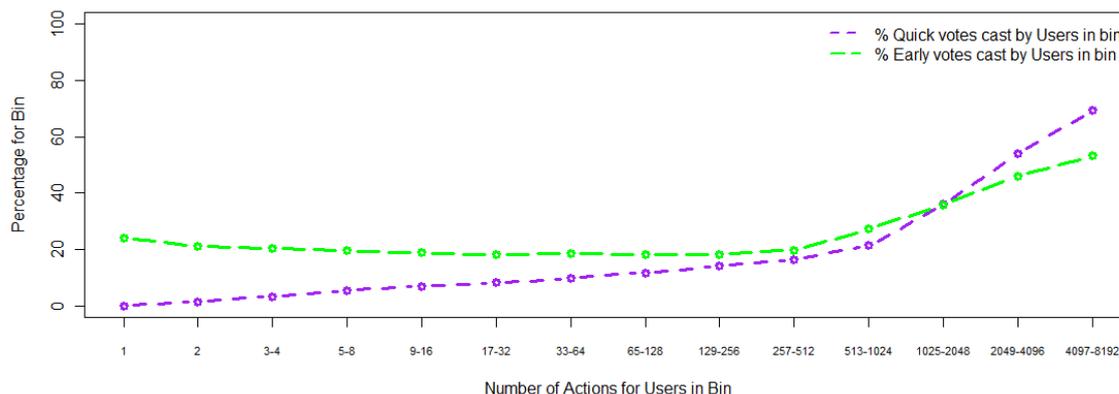
Figure 6.24: Showing the proportion of 'quick' (within 10 seconds of user's previous vote) votes and 'early' (one of the first 20 votes to be cast on a post) votes cast by users in each activity bin.

Figure 6.24 considers the origins of early and quick votes in terms of the activity levels of the users who cast them. Both of these types of vote appear to be more common among highly active users (specifically users with more than 512 actions in the month). The figures considered thus far suggest an association between early votes, quick votes and down votes. Table 6.2 shows a cross-tabulation of binary variables representing whether a vote was a down-vote and whether it was classified as an early or quick vote.

	Up-Votes		Down-Votes	
	Not Quick Vote	Quick Vote	Not Quick Vote	Quick Vote
Not Early Vote	1106907	170260	230548	75055
Early Vote	331888	87840	188808	112779

Table 6.2: A cross-tabulation of binary variables representing whether a vote was a Down-vote, Quick vote, Early vote - votes which accompanied a post submission, votes on 'obscure' subreddits, and votes which were a user's first in the observation period have been excluded. N = 2,304,085

Log-linear Poisson regression models have been fitted to test the association between these three types of vote. A model was fitted with binary main effects for whether votes were up-votes, quick votes and early votes, along with 2-way interactions between each of these variables - all of the effects were highly significant and this model had a residual deviance of 78. Next, each of the interactions were removed from the model one at a time - where the removal of an interaction

term results in a significant increase in deviance for the model this indicates that an association between the two variables is present. Table 6.3 shows the resultant increase in deviance when each interaction term is removed from the model. The removal of any interaction term results in a highly significant increase in deviance for the one additional degree of freedom gained. Therefore all interaction terms are required in the best model and this indicates an association between the three types of vote. The strongest association is between whether votes are Early and whether they are Up-votes - when a vote is Early it is much more likely to be a Down-vote than when a vote is not Early. There is a weaker relationship between Early votes and Quick votes but these do tend to co-occur more often than would be expected by chance.

Interaction Removed	Degrees of Freedom	Deviance
Early votes X Up-votes	+1	+102889
Early votes X Quick votes	+1	+25029
Up-votes X Quick-votes	+1	+ 44993

Table 6.3: Increase in deviance observed when interaction terms were removed from the log-linear model of vote type frequency. Conclusion is that all interaction terms are required in the best model and therefore there is an association between these types of vote.

A small minority of highly active users account for a significant proportion of user voting activity on `reddit`. These users are also more likely to engage in forms of voting which are consistent with attempting to maximise ones' input into `reddit`'s decision-making process.

Are these users behaving in this way in an attempt to 'game' `reddit`'s voting system such that they exert more influence than other users? Or are they simply finding the most productive way for them to contribute to `reddit`'s utility? It is also important to ask whether these voting behaviours actually do give these users a greater say in `reddit`'s decision-making process.

In the case of early votes it is conceivable that these are pro-social - users may see this as pitching in to ensure that 'poor' submissions do not reach high-visibility locations. In the case of quick votes an altruistic motivation is harder to envisage - voting in quick succession based on the titles of posts seems more in line with a user who wishes to maximise their influence, as opposed to helping ensure that the website's collective decision about a post is correct.

In section 3.3.3 the prospect of 'superparticipants' (Graham, 2013) who are defined by their voting behaviour was raised. The sub-group of users on `reddit` who cast very large numbers of down-votes which tend to be 'early' and 'quick' represent a exactly such a class of superparticipant. Furthermore, the defining characteristics of their voting behaviour are consistent with an intent of 'curbing' the capacity for others to be involved in the discourse and have their contributions seen. Their votes are largely down-votes (a curbing mechanism), they tend to be cast in locations which matter, and at high speed.

6.1.5 Users who submit Posts

This section considers users who submit posts on `reddit`; are some of these users more successful than others? Are there attributes of *users* which are associated with more successful submissions? Is there any evidence that new users learn to submit 'better' content as a result of receiving feedback through the voting system?

Are some users more successful than others?

A highly skewed distribution of activity again forms the background to these analyses. Of all the users who made at least one recorded action in the data, 23,776 (24 %) made just one post submission. 17,997 (74 %) of these user accounts were new accounts registered within the month of March 2009. This is likely a reflection of the ease with which an individual can create a user account on `reddit`. All that is required for the creation of a new account is a user name and password, `Reddit` does not employ e-mail address verification. `Reddit` users often create 'throwaway' accounts when they wish to make a post about a sensitive issue - because they do not wish this post to be tied to their usual `reddit` identity. It is also conceivable that many of the user accounts which made a single post submission have been created by individuals who are not regular `reddit` users but who have some web resource that they wish to garner attention for (these posts often being referred to by `reddit` users as 'blogspam').

In considering the 'success' of users' submissions here the final observed score for their posts is always taken as the measure of success (often referred to as the final score or score). Voting on `reddit` posts does not close after a certain period of time and so the expression of a post's score being 'final' is not meant to be taken literally but rather shorthand for 'final observed'.

The mean score achieved by new users who made a single submission was 0.04 - compared to 'returning' users whose single submissions achieved a mean score of 2.6. This suggests that users with a history of participation on `reddit` have an advantage when they submit content. This could be because `reddit`'s voters are wary of newcomers, or because more experienced users have a greater understanding of which posts are likely to fare well on `reddit`.

Multi-level modelling has been employed to determine whether there is a relationship between the success of a user's submissions and a selection of explanatory variables. It is expected that individual users will have some level of inherent ability to submit posts which score highly. A mixed-effects model with a random effect for Users (could also be described as a multi-level model with users on the second level) should account for these individual differences at the level of the user, allowing better estimates for the explanatory variables' coefficients to be produced. A random effect for users is sufficient here because we are not interested in predicting an effect for each individual user, only in controlling for these effects.

A further problem with modelling the scores of posts relates to their power law distribution. The distribution of scores between posts is highly skewed such that it cannot be treated as normal or transformed to be normal - the small number of posts with a very high score would skew results derived from models which treat these as normally distributed. Here we are not interested in which posts reached the Front page as this is a very small proportion of total posts, we are interested instead in whether users with certain characteristics submit posts which score 'well'. Post scores have therefore been converted to a four-level ordinal variable - with levels for posts which receive more negative than positive votes, posts which receive the same number of positive and negative votes (or no votes), posts which achieved a score of between two and twenty (moderately positive voting response), and posts which achieved a score of greater than twenty (very positive voting response). STATA's GLLAMM package was used to fit a mixed-effects model with a random effect for users and an ordinal logit link.

Formally, the model can be written as:

$$\log_e y_{\{ij\}} = \gamma_{00} + \gamma_{q0}(X)_{i(j)} + u_{1j} + \epsilon_{ij}$$

where i indexes the post, j indexes the user, $(X)_{i(j)}$ is a vector of explanatory variables of length q at the user level, γ_{q0} is a vector of unknown fixed effects paramaters associated with X, u_{1j} is a random effects term associated with the user (assumed to have normal distribution) and ϵ_{ij} is residual variability (also assumed to have normal distribution).

In fitting these models, only submissions to the 20 most active subreddits of the time are considered as more obscure subreddits tend to behave differently (see Chapter 5). As a result, these models are fitted to a dataset containing 268,187 posts submitted by 51,646 users. First a model with two levels (posts on level 1, users on level 2) but with no explanatory variables was fitted to determine whether a level for users was required - the users level accounted for a significant amount of the variance in post scores ($\chi^2 = 47.75$, df $= 1$, p < 0.001). This is a strong indication of a relationship between the scores of posts which were submitted by the same user. It would seem that users vary in their overall success profile when they submit posts. This model had a deviance of 554994.

This model then served as a base model - variables were added one at a time and only retained when they reduced the model's deviance by a significant amount for the additional degrees of freedom which they required (tested against the χ^2 distribution). Table 6.6 contains details of the model building process, models which did not produce a significant reduction in deviance are shown in *italics* and the subsequent model is therefore tested against the last 'good' model.

Table 6.4 shows the breakdown of the ordinal Post score response variable. Table 6.5 provides descriptives for the count variables which were utilised in modelling. The maximum values for variables relating to votes all come from the same user - this user voted much more often than any

other user in the data. As a sensitivity analysis the final model was fitted again without this one user (who could be considered an outlier) but the model paramaters remained stable when this user was excluded - so the user has been included in the final model fit reported below.

Category	Number of Posts
Score < 1	86,526
Score = 1	181,157
Scores 2-20	71,753
Score > 20	12,184

Table 6.4: Number of posts falling into each ordinal grouping of final observed post score

Variable	Mean (Std. Dev)	Range
Total Votes	90 (450)	0 - 23,704
Total Post Submissions	76 (133)	1 - 1,246
Total Quick Votes	30 (325)	0 - 20,230
Total Early Votes	39 (299)	0 - 15,714
Total Down-votes	32 (331)	0 - 22,479

Table 6.5: User counts used as explanatory variables

In addtion to the stepwise addition of variables a full model (with all available explanatory variables) was tested by removing one variable at a time and observing the resultant increase in deviance. This is a 'fairer' method of deciding which variables should be retained in the final model as it gives a less biased account of how useful these variables are. For example, the effect of Early Votes seems quite weak when the model is 'stepping forward' because it has been added after Quick Votes and these variables are highly correlated.

Details of the final model are included in table 6.8. In order to interpret this model the most important consideration is the signs of coefficients and the signs of cut points. Where a coefficient is negative it is associated more strongly with those cut points whose coefficient is also negative. In this case negative coefficients for explanatory variables are associated with negatively-scoring posts.

The first variable considered was the proportion of a user's actions which were post submissions, this was initially entered as a binary variable denoting whether a user had voted at all but the proportion of their activity which was post submissions has more explanatory power. The coefficient for this variable is negative, indicating that users who had a higher proportion of post submissions tended to submit posts which scored poorly - whereas users who voted more than they submitted would be predicted to submit higher-scoring posts. The variable representing a user's total number of post submissions also has a negative coefficient, suggesting that users who submitted more posts

Model	Change in Deviance	Change in df	p-value
1. Binary variable representing whether a user voted at all	-1123	1	< 0.001
1b. Replace binary 'post submission' variable with proportion submissions (0-1)	- 705	0	< 0.001
2. + Account Age (4-level ordinal variable treated as continuous)	-400	1	< 0.001
2b. Treat 4-level ordinal Account Age variable as categorical	*-4*	*+2*	*0.135*
3. + User's total number of votes	-431	1	< 0.001
4. + User's total number of 'quick' votes	-332.5	1	< 0.001
5. + User's total number of down-votes	*-3.6*	*1*	*0.0578*
6. + User's total number of 'early' votes	-29	1	< 0.001
7. + Binary variable - User was active on 'Obscure' subreddit	-275	1	< 0.001
8. + Total number of Post submissions	-29	1	< 0.001

Table 6.6: Steps in the construction of an ordinal logistic regression model of Post's score with a random effect for the submitting users. Rows in *italics* did not improve the model and were not retained.

tended to submit poorer posts - this may be an indication that users who submit many posts are prioritising quantity over quality, or that some of them are spammers.

The next variable to be considered was the age of users' accounts. In the raw data supplied by `reddit` user identity numbers are ordered by account creation date. This fact was used to identify those users who created their account during the observed month (March 2009), and to divide pre-existing accounts into three age bands. Thus a four-level ordinal variable representing the age of user accounts was created. This age variable was added to the model as both a continuous variable and a factor. In principle, it would be better to compare each level of account age to a reference group (new accounts). However, this paramaterisation of the model yielded a deviance just 4 smaller than the model with a continuous 'account age' coefficient, not a significant improvement given the additional model complexity (two additional degrees of freedom required).

We consider here only posts submitted to the 20 most active subreddits of the time - a binary 'Active on obscure subreddit' variable was created which denotes a user who had participated by voting or submitting on one of the more 'obscure' subreddits. At the time when this data was collected finding out about non-default subreddits was difficult and observations suggested that people often learned about non-default subreddits from the comments of others or through an

Variable Removed	Change in Deviance	Change in df	p-value
Proportion Post Submissions	33	1	< 0.001
Account Age	192.4	1	< 0.001
Total Post Submissions	28.8	1	< 0.001
Total Votes	213	1	< 0.001
Total Quick Votes	155.8	1	< 0.001
Total Early Votes	14.4	1	< 0.001
Total Down-votes	*0.78*	*1*	*0.377*
Activity on 'Obscure' subreddit	288.5	1	< 0.001

Table 6.7: The change in deviance when explanatory variables are deleted from the full model

external website like `MetaReddit.com`. Where Account Age can be thought of as respresenting the duration of a user's experience of `reddit`, activity on an 'obscure' subreddit can be thought of as representing the depth of a user's experience. Both of these variables are associated with higher-scoring post submissions - users with older accounts and users who had some activity on an 'obscure' subreddit both tended to submit higher-scoring posts.

A user's total number of votes has a lot of explanatory power and is associated with higher-scoring submissions - users who vote more tend to submit higher-scoring posts. However, this effect is offset by negative coefficients for the 'expert' voting behaviours (Quick Voting and Early Voting both have negative coefficients, as did Down-voting but this variable did not have a significant effect). The coefficient for quick votes is slightly stronger than that for all votes, so each quick vote is associated with a net decrease in the probability of the user submitting high-scoring posts. As we saw earlier, quick votes and early votes tend to co-occur, and where they do this is associated with an even larger net decrease in the probability of the user submitting high-scoring posts.

As an aid to interpreting this multi-level model the probability distributions for the post submissions of four hypothetical users (calculated with the final model) are considered in figure 6.25. These predicted probabilities reveal the strength of the variable representing a user's number of votes. Users who vote frequently have a much better chance of submitting successful posts themselves. However, this effect is counteracted by the effect of quick and early votes - where a user votes frequently but quickly and early this is associated with a net reduction in their probability of submitting successful posts.

The combination of these variables can be thought of as representing a user's level of involvement with or commitment to the website. Users who have older accounts, who are active on non-default subreddits and who vote more (but give reasonable consideration to their votes) fare better when they submit content. It is interesting to note that the variable representing users number of submissions was associated more with low-scoring submissions. These results suggest that the users

	Coefficient	Std. Error	95% Confidence Interval	p-value
Post Submission Proportion	-0.0962	0.0167	-0.13, -0.06	< 0.001
User's Total Post Submissions	-0.0007	0.0001	-0.001, -0.0005	< 0.001
User's Account Age	0.0956	0.0069	0.08, 0.11	< 0.001
user Active on 'Obscure' SR	0.3193	0.0188	0.283, 0.356	< 0.001
User's Total Votes	0.00116	< 0.0001	0.001, 0.0013	< 0.001
User's Quick Votes	-0.0012	0.0001	-0.0014, -0.001	< 0.001
User's Early Votes	-0.0004	0.0001	-0.0006, -0.0002	< 0.001
Cut-point 1 (Score < 0)	-1.008	0.0171	-1.18, -1.13	< 0.001
Cut-point 2 (Score <= 0)	1.8	0.0175	1.76, 1.83	< 0.001
Cut-point 3 (Score <= 20)	4.061	0.021	4.02, 4.1	< 0.001

Table 6.8: Showing final model parameters for a multi-level model with a random effect for Users and Post Score (ordinal) as the response variable. There is a small but significant ($p < 0.001$)) negative relationship between total submissions and submission score, and a strong negative relationship between submission proportion and post score. Voting is associated with higher-scoring post submissions, but this is counter-acted by negative effects for 'early' and 'quick' votes. This model had a deviance of 552780.7

who define `reddit` are the users who vote. Users who only submit posts, or who vote rarely, seem to operate on the periphery - and this effect occurs without `reddit` users being able to see how many votes other users have cast. It seems likely that through voting regularly and thoughtfully `reddit` users gain experience which leads to a better understanding of what the website is about and how it works; which in turn gives them an advantage when they submit posts.
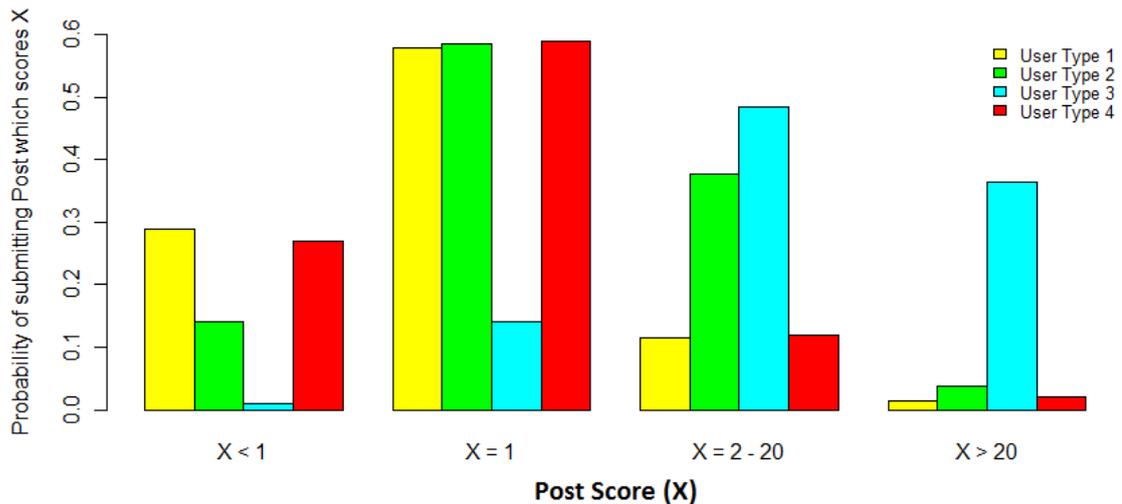
Figure 6.25: Showing probabilities predicted by the multi-level model for four hypothetical users. User type 1 was a new user who made only one post submission and no votes. User type 2 had an account in the oldest 25 %, made 200 votes and 10 submissions and had some activity on an 'obscure' subreddit. User type 3 had an account in the oldest 25 % and made 2,500 votes and 10 submissions and had some activity on an 'obscure' subreddit. User type 4 had an account in the oldest 25 % and made 2,500 votes (2,000 of these were quick votes and early votes) and 10 submissions.

New Users and feedback received through the voting system

`Reddit`'s voting system allows users to express their collective will, but the judgments registered through this voting system may also act as feedback for submitting users. An up-vote increases the chances of a post being seen more widely, but it also lets the user who submitted the post know that it has been seen and appreciated by another user. This section considers new users' experience of receiving this kind of quantitative feedback. Does such feedback encourage new users to continue submitting content? Do users learn by receiving this feedback to submit posts which will be appreciated more by other users? Does `reddit`'s voting system facilitate the induction of new members into the community?

During the month for which back-end data are available (March 2009) 28,908 new user accounts were created on `reddit.com` and used to submit or vote on a post at least once. A further 30,000 user accounts were created but were not used for post submission/voting - these accounts were either not used to participate or were used for comment submission/voting (no back-end data on comments is available for analysis). 19,200 (66.4%) of the new user accounts which have recorded actions were used just once. To compare these figures to those observed on `Slashdot` by Lampe and Johnston (2005) - Reddit had a much higher rate of new account creation during the observed month, and also a slightly higher drop-out rate (around 55% of new `Slashdot` users made just

a single recorded action). However, it should be noted that these comparisons are not entirely equivalent between websites, Lampe and Johnston (2005) looked at comment submissions whereas this research has considered post submissions.

Does feedback received through the voting system encourage new users to continue submitting posts? A logistic regression model was fitted to address this question. In order to do so accurately it was necessary to exclude user accounts created near the end of the study period. In this data the median delay between a new user's first and second submission is 99 minutes, and the mean is 41.7 hours with a standard deviation of 87.6 hours. User accounts created during the last week of the study period have been excluded from the following analyses - it is assumed that where a user has not submitted a second post within a week of their first post they have ceased submitting posts.

	Coefficient	Std. Error	p-value
Intercept (No Votes)	-0.91	0.02	< 0.001
Positive score	0.51	0.04	< 0.001
Negative score	0.26	0.06	< 0.001

Table 6.9: Showing a logistic regression on the probability that a new user would make a second submission, based on the voting performance of their first submission.

A logistic regression model was fitted on whether a new user would make a second submission, based on the voting performance of their first submission. The model, presented in table 6.9, reveals an association between voting feedback on a new user's first post and their propensity to submit a second post. New users whose first post received no votes are the reference category and the positive coefficients for both of the other variables suggest that these users were least likely to continue submitting. New users whose first post achieved a positive score were 67% more likely to submit again than users whose first post was not voted on. New users whose first post received a negative score were 29.5% more likely to submit again than new users whose first post was not voted on. These results suggest that receiving feedback through the voting system encourages new users to continue submitting content. Lampe and Johnston (2005) looked for but did not find this effect with new `Slashdot` users' comments. However, both Lampe and Johnston (2005) and Halavais (2009) reported effects, on `Slashdot` and `Digg` respectively, whereby voting feedback on a user's first post was associated with a decrease in the time taken to make a second post.

Feedback received through the voting system encourages users to continue submitting posts. Do new users also learn from this feedback which types of content are appreciated by other users? If this were the case one would expect to see an upward trend in the scores received by new users' posts as they gain experience. Halavais (2009) reported such a trend on `Digg`, with mean scores of comments increasing with their order of submission.

Multi-level models similar to those utilised in the previous section were deployed to look for such a trend in the `reddit` data. New users likely have some level of intrinsic ability to make successful posts - by accounting for such variance in innate ability at the level of the user it should be easier to detect an effect whereby a user's posts get better with practice. These models were fitted to a sub-set of the data concerning new users. Only new users who made at least five submissions were considered, and only the first five submissions of each user were considered. This decision was taken to marginalise any effect of self-selected dropout. If one assumes that new users who make post submissions which receive consistently low scores are less likely to continue submitting posts than more 'gifted' users - then if all post submissions by new users were considered one would expect posts with a higher post order value to have higher scores, by virtue of the fact that only users who are relatively good at submitting posts made a high number of posts. As a result, these models are fitted to data concerning the first five posts of the 2,608 new users who made at least five post submissions. However only 12,835 posts are considered in the models because post submissions to 'obscure' subreddits are not included.

First a model with two levels (posts on level 1, users on level 2) was fitted to determine whether a level for users was required. The users level accounted for a significant amount of the variance in post scores ($\chi^2 = 15.4$, df $= 1$, p < 0.001). This is a strong indication of a relationship between the scores of posts which were submitted by the same user. It would seem that, as with the analyses of all submitting users, some new users are more likely to submit successful posts than others. This model had a deviance of 20455.

The first explanatory variable to be added represented the order of post submission ranging from one to five. The coefficient for this model was significantly different to zero, and reduced the model's deviance by 25, a significant change on the one additional degree of freedom required. The coefficient for post order is positive, indicating a trend whereby users' post submissions tend to become more successful with practice. However, the strength of this effect is not particularly strong. Without considering any other variables, an 'average' new user's first post has an 18.7 % chance of receiving a negative score whereas their fifth post submission has a 14.7 % chance of scoring negatively (with the chances of submitting a positively-scoring post being 5.5 % and 7 % respectively).

Next explanatory variables relating to the new users were added to the model. These were only retained when they reduced the model's deviance significantly (p < 0.05) taking into account the number of additional degrees of freedom required. The first variable to be added concerned the total number of submissions the new user made in the month. This variable did not produce a coefficient which was significantly different from zero. This is not surprising, as the model only includes posts with order 1-5; whether a new user made many more than five submissions is perhaps unlikely to be related to the performance of their first five submissions.

In contrast, a user's total number of votes in the month did produce a coefficient which was significantly different from zero (p < 0.001), and reduced the model's deviance by 36 on one additional degree of freedom. This is a somewhat surprising effect, the positive coefficient indicates that a new user's total number of votes in the month is associated with better performance for their first five submissions. This likely reflects a difference in the attitude of new users, those who become active voters are likely more committed to becoming rounded `reddit` users and genuine members of the user community - and this is somehow reflected in the scores of their first five submissions. Perhaps these users have developed a more informed view about which posts are likely to score well while they have been voting on posts and presumably witnessing the scores attributed to these posts by other users. To equate this effect with the effect of post order, every 20 votes a user made in the month is associated with a similar level of improvement in their submissions' probability of succcess as an increase of one on the post order variable. To put this another way, users 'learn' as much for every 20 votes they cast in the month as for making one submission.

In response to this finding a new variable was created representing a new user's number of votes cast *at the time they submitted each post*. A model was fitted with this new variable in place of the variable representing each user's total number of votes in the month. This model produced a deviance of 20420, 9.5 lower than the previous model. The measure of a new user's *current* number of votes at the time of each post submission is therefore preferred to the measure of their *total* number of votes in the month. The coefficient for this variable is larger, in this model a new user who has made 12-13 votes has learned as much as a new user who makes one additional post submission.

A variable representing whether a user submitted posts exclusively or also voted was included in the model at this stage, but this variable did not produce a coefficient which was significantly different to zero.

The next variable to be added was a binary variable representing users who made at least one vote or submission to an 'obscure' subreddit (not one of the 20 most active subreddits of the time). This variable produced a sizeable and significant (p < 0.001) coefficient, and reduced the model's deviance by 73.25 on one additional degree of freedom. As with the model of all users in the previous section, users who have some level of activity on 'obscure' subreddits tend to submit more successful posts. This is again taken as a sign that the user has a higher level of familiarity with `reddit`. In the context of analyses concerning new users, this effect might be interpreted as evidence that the new user account actually belongs to an existing user, perhaps serving as their alternate or a 'throwaway' account.

Details of the final model for new users are displayed in table 6.10. The main finding from this model is the weak but significant effect of post order - with each post a new user makes the chance

	Coefficient	Std. Error	95% conf. interval	p-value
Post Order (trend)	.0683	0.0142	0.04, 0.1	< 0.001
Votes at post submission	.004	0.0017	0.0006, 0.007	< 0.05
Active on 'obscure' subreddit	0.542	.063	0.42, 0.67	< 0.001
Cut-point 1 (Score < 0)	-1.253	0.054	-1.36, -1.14	< 0.001
Cut-point 2 (Score <= 0)	3.076	0.065	2.95, 3.2	< 0.001
Cut-point 3 (Score <= 20)	5.456	0.108	5.25, 5.67	< 0.001

Table 6.10: Showing a multi-level model fitted on a post's voting performance (ordinal) with User as the level 2 variable. New users only, and only the first five posts for users who have made at least five posts. Deviance for the model is 20346.75.

of that post scoring positively increases. Also of interest is the explanatory variable representing the user's voting behaviour - further evidence that `reddit` users who vote on others' posts are more likely to themselves submit positively-scoring posts. Finally the variable representing activity on an 'obscure' subreddit in this context likely identifies new user accounts which do not belong to genuinely new users.

6.1.6 Posting superparticipants

Section 6.1.4 found utility in the concept of 'superparticipants' with regard to user voting behaviour - and identified one type of superparticipant who may be influencing `reddit`. In the same dataset there is evidence of users who submit a lot of posts, but the negative relationships between submission volume/proportion and scores (table 6.8) argues against considering these users to be 'superparticipants' - they certainly cannot be considered to be 'dominating' `reddit` in terms of the posts which users are seeing. Section 3.3.3 considered Graham (2013)'s 'agenda-setting' superparticipant and how this concept might translate to `reddit` - with posts or top-level comments being the obvious places for would-be agenda-setters to operate.

One limitation of front-end data collected through `reddit`'s API for this research is that this has not been collected on the basis of users. Information on users' voting behaviour is unavailable through the API. Data on users' submitted posts and comments can be collected from a user's public 'profile' pages, but this approach has not been adopted because it would require a significant amount of time (in writing the necessary scripts) and computational load (to run these scripts periodically retrieving data from user pages) with no hope of building a 'complete' picture of users' behaviour. There would also be serious ethical considerations to doing so. Although profiles are pseudonymous and `reddit` strongly discourages the posting of personal information, users may over time drop pieces of information about their lives into discussions in comments - and in some cases piecing these together might enable one to make a good guess at their identity. If one was to

make records of user posts and comments in an indexable database this resource could potentially be exploited to identify individual users.

It is however possible to present some relevant data from the longitudinal data-set (on which much of Chapter 8 is based). In Chapter 8 data-sets are compiled from observations of reddit's Front page recorded at 30-minute intervals between 10th June 2010 and 20th August 2012 (with a watershed moment at 18th October 2011 when the makeup of the Front page was changed dividing this into two distinct data-sets). See Chapter 8 for details of the method of data collection/preparation - and notes on some gaps in data collection.

These longitudinal data-sets allow for the consideration of users who submitted posts which were ultimately displayed on the Front page (although there is no complete record of these users' submissions which did not appear on the Front page, or their comments). In the first period (10th June 2010 until 18th October 2011) 38,973 different posts from 26,717 different users were observed on the Front page - between them receiving roughly 15.5 million comments (mean 400 comments each). There were 4,672 users who submitted more than one post which appeared on the Front page. In the second period (18th October 2011 until 20th August 2012) 43,489 different posts from 34,202 different users were observed on the Front page - between them receiving roughly 23.8 million comments (mean 550 comments each). There were 4,289 users who submitted more then one Front page post in this period.

Considering individual users, there is one obvious candidate for the title of 'superparticipant' - maxwellhill had the greatest number of posts observed on the Front page in both time periods. In the first period this user had 302 front page posts through 7 different subreddits, receiving in combination 85,355 comments and a cumulative score of 204k. In the second period maxwellhill had 195 Front page posts through 4 different subreddits, receiving 112,994 comments and a cumulative score of 295k.

Front Page Posts	Number of Users	% of FP posts
1	45,598	55%
2	7,020	17%
3-5	3,017	13%
6-10	610	5%
11-20	192	3%
21-50	60	2%
51+	27	4%

Table 6.11: Showing the number of Front page posts submitted by users per 'bin' between 10th June 2010 and 20th August 2012 - and the % of all Front page posts submitted by users in the 'bin'.

The distribution of front page posts between users is, unsurprisingly, highly skewed. Putting the two data-sets together, table 6.11 shows the number of users who submitted 'Front page posts' within given ranges - and the percentage of Front page posts which were submitted by users in each bin. More than half of all the posts observed on the Front page during this two year period were submitted by users who only had one of their posts appear on the Front page - suggesting that 'super posters' or 'power users' do not dominate `reddit`'s Front page. There are however a small number of users who seem to have a knack of submitting posts which reach the Front page - there are 279 users who submitted more than 10 Front page posts, between them accounting for around 9% of all Front page posts. These users could be described as 'agenda setters'. Although the number of their posts which appeared on the Front page is not particularly impressive, each occasion where this occurred is in many ways a significant event. If we consider the 27 users who submitted more than 50 Front page posts each - these users' (3,142) posts have received a combined total of 1.25 million comments, a tally which should easily qualify these users as 'agenda setters' in most contexts.

One further question which can be asked (with the available data) about these users who have submitted more than one Front page post is whether their seeming capacity to reach the Front page is confined to a single (or small group of) subreddits. Figure 6.27 (based on the data-set spanning 18th October 2011 until 20th August 2012) shows little evidence that this is the case. For example, for users who submitted two Front page posts these were more likely to be submitted to two different subreddits. Users who reached the Front page on more than one occasion do not appear to 'specialise' in submitting to a single subreddit.

These findings are also relevant to questions about `reddit`'s 'democratising' potential, going some way to refute the concern that Social News is a means of production easily co-opted by an elite of power users (see sections 2.11 and 3.4.1). The data also faciliates a direct comparison with a figure produced by the *Wall Street Journal* and quoted by Keen (2007) - namely that one third of Front page posts on Digg (over a certain interval) were submitted by just 30 users. Table 6.11 shows that `reddit`'s top 27 users over a two year period submitted just 4% of the posts which appeared on the Front page. This suggests that `reddit`'s approach to DM has not resulted in power users to the same extent that these dominated Digg - and is an indication that `reddit`'s formulation of DM was in this respect superior to Digg's, even during the peak of the latter's popularity.

Reading this another way - we can extrapolate that at any given moment over this two year period more than half of the posts appearing on `reddit`'s Front page were submitted by a user who had not previously submitted a Front page post. Submitting a post which hits the 'attentional jackpot' and appears on `reddit`'s Front page is not the privilige of a small set of power users, this is a resource which is available to all users in practice as well as in principle.
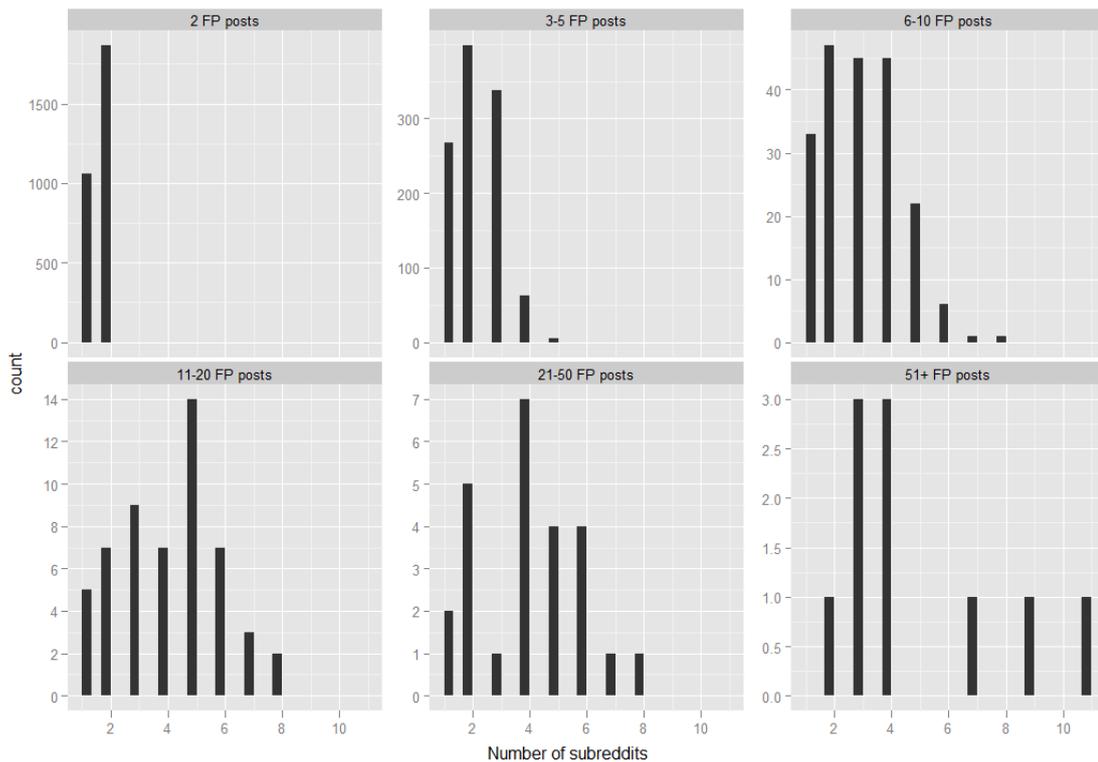
Figure 6.26: Showing the number of different subreddits which users reached the Front page through - panels show users with a specified number of Front Page posts

6.1.7 Questionnaires regarding `reddit` use

There are certain aspects of `reddit` use which it is not possible to address through the analysis of user behaviour - chiefly questions of motivation. This section discusses the results of two questionnaires, one created by myself, which included items which address aspects of `reddit` use that are not captured by the available procedural data. These questionnaires were submitted as posts to `reddit` in the same manner as the surveys discussed in section 6.1.1 - but unfortunately they did not appear in a high profile location and therefore were completed by a much smaller numbers of respondents. The fact that these questionnaires did not appear on the Front page also causes a secondary problem - the questionnaires were only seen (in one instance) by people who browse the new/rising sections of a large subreddit or (in the other instance) by people who subscribe to niche subreddits.

Questionnaire 1 - produced and administered as part of the present research

The first of these questionnaires was created as part of the present research and submitted to `reddit` (specifically the /r/reddit.com subreddit) in February 2010. The post received a moderately positive voting response and appeared briefly on the subreddit's Rising page in addition to it's New page - it was completed by only 139 respondents. This poses a problem because as we have seen

234

there is likely a distinction between users who are active in these locations and those who stick to higher profile locations like the Front page. The plan for analysing responses to this questionnaire involved labelling users according to where the post appeared *at the time when they completed it*, and this would have allowed us to contrast the 'expert' users who are active on the New/Rising pages with those users who read the Front page - indeed several items were included specifically with this in mind. As the post did not appear on the Front page it is not possible to make these comparisons, and instead we will only consider a selection of variables which are interesting in their own right and which have not been addressed by one of the surveys with a higher number of respondents.

In addition to demographic items similar to those in section 6.1.1 (which will not be reported on here other than to say that they exhibited similar trends to the large-scale demographic surveys), this questionnaire asked respondents about how and why they used `reddit`. For some of the items relating to how one uses `reddit`, analysis of the procedural data from March 2009 will provide a much better picture (as it does not rely on self-reporting of activity and all users are represented) - but for other aspects (e.g. commenting, multiple accounts) the procedural data is of no utility.
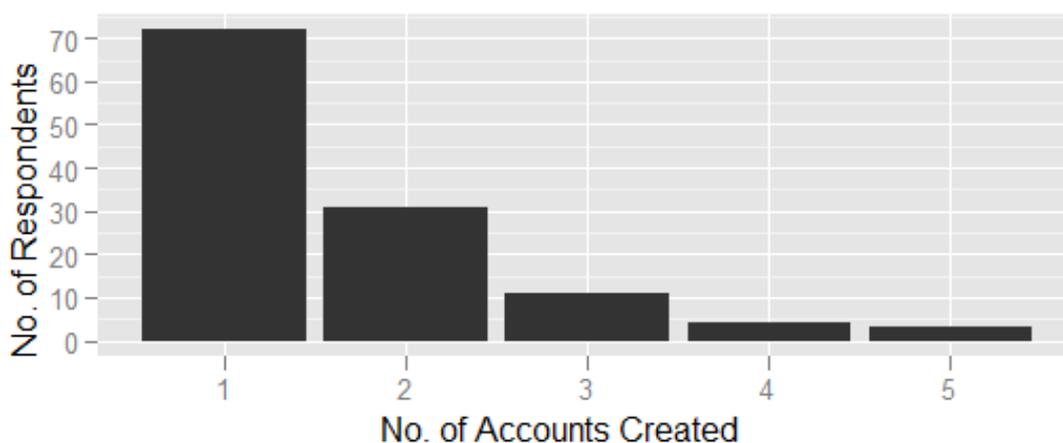
Figure 6.27: Showing the number of `reddit` accounts respondents had created

Of the 139 respondents 92% indicated that they had a `reddit` account. The questionnaire also asked respondents how many `reddit` accounts they had created and provided a box for them to enter their repsonse. I have previously suggested that users may have more than one account (based largely on the ease with which an account can be created) - figure 6.27 provides the only available evidence that this is the case and suggests that the practice is quite widespread (although bear in mind that respondents to this questionnaire are likely 'heavy' `reddit` users).

Figure 6.28 suggests that the majority of respondents spend between 1 and 4 hours on `reddit` every day. Figure 6.29 addresses the question of whether respondents use `reddit` for 'News' or some other purpose. Here there seems to be a trend whereby respondents do see `reddit` as a News type
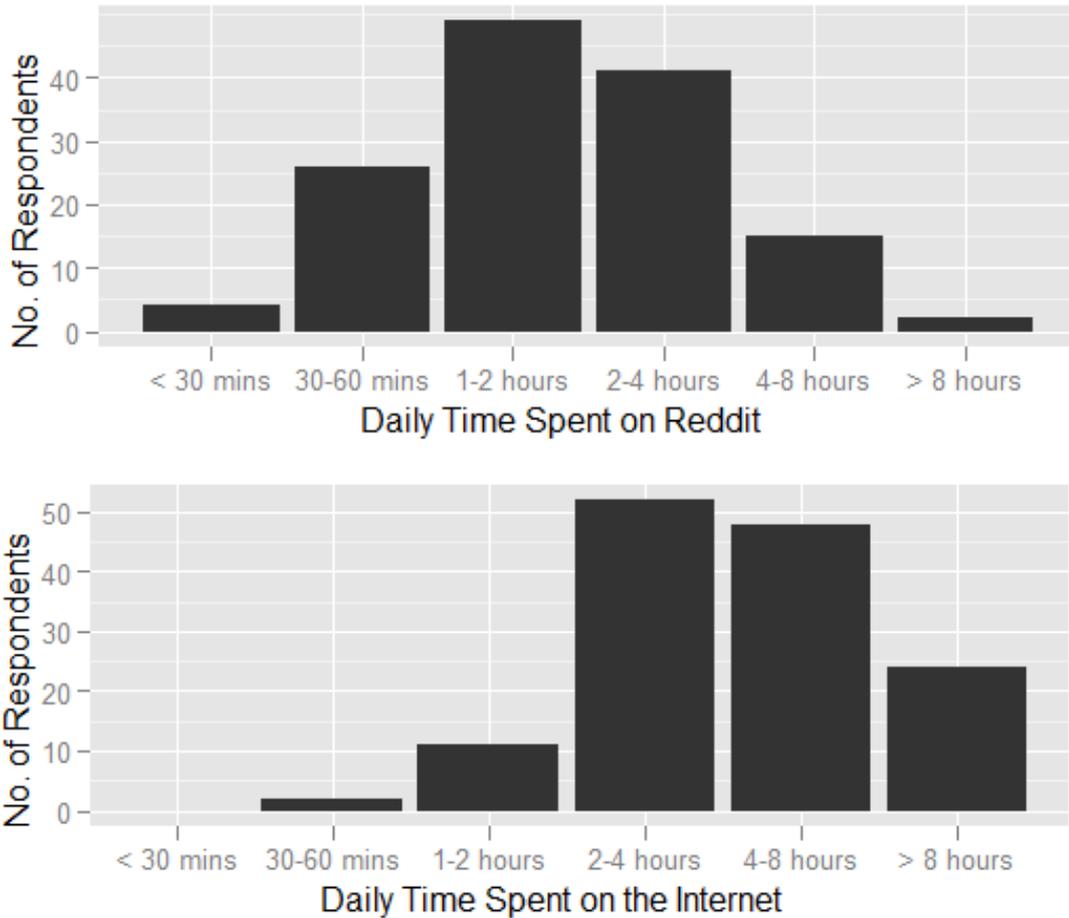
Figure 6.28: Showing the length of time respondents spent on `reddit` and the Internet daily

resource, with very few respondents choosing 'None'. One wonders whether the same proportion of respondents would say that they get 'News' from `reddit` were the survey to be performed again in 2012, given the apparant shift in the nature of Front page content.

Figure 6.30 shows responses to the items concerning the nature of one's activity on `reddit`. These responses fit with what would be expected of 'heavy' users based on the behavioural analyses, with many respondents claiming that they often vote but rarely submit posts. These items were intended to address one of the blind spots of the behavioural analyses - commenting. Here we are interested in whether comments tend to come from users who vote or users who submit posts. Spearman's rank correlations were calculated between responses to these items on a 5-point scale and there is roughly the same level of correlation between comment regularity and both voting (0.55) and post submission (0.54) regularity, respectively. There is a slightly lower level of correlation between submitting and voting (0.4). However, the small sample size limits the reliability of these correlations.
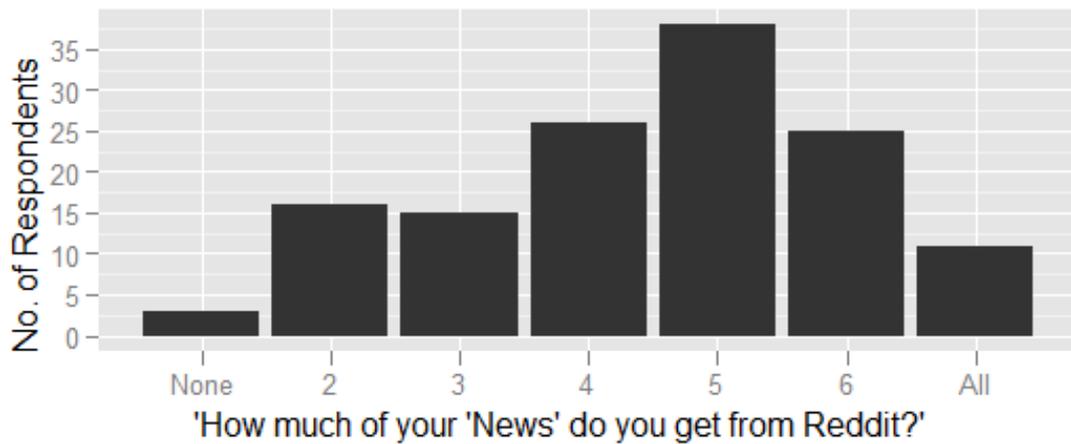
Figure 6.29: Showing responses to the item asking 'How much of your 'News' do you get from `reddit`?' Responses recorded on a 7-point scale with the scale's extremities labelled 'None' and 'All'
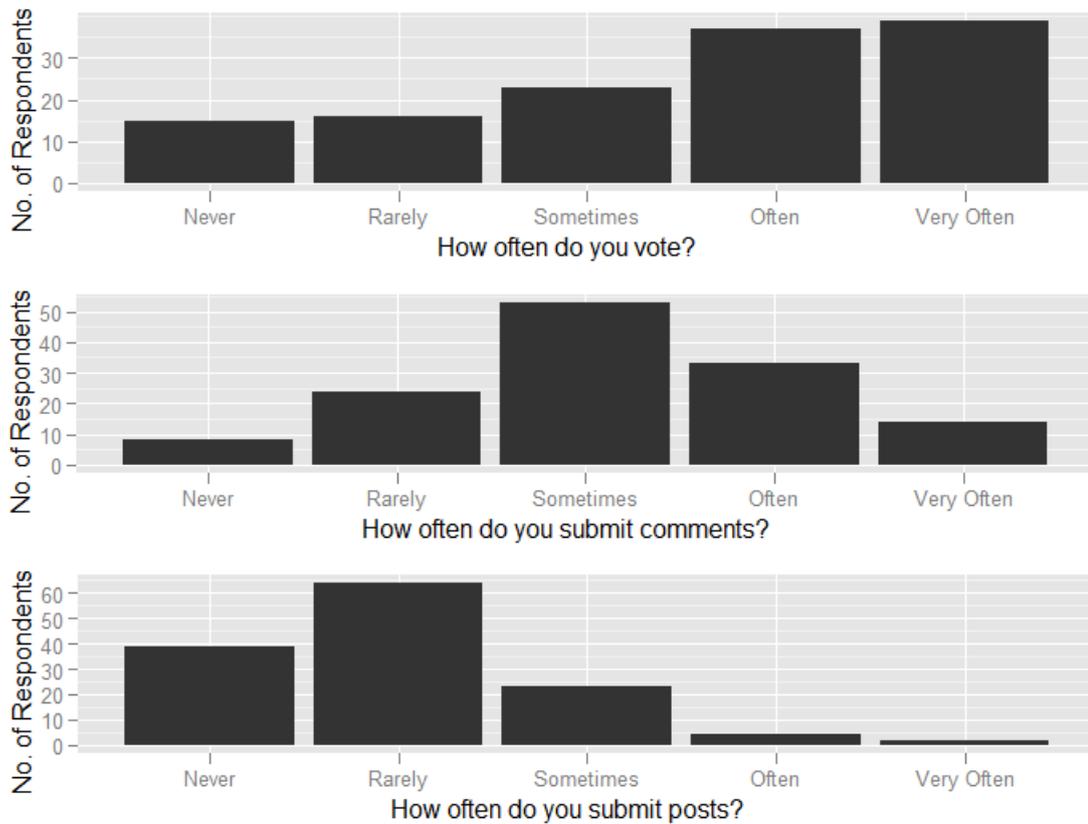
Figure 6.30: Showing the frequency with which respondents participated by voting, commenting and posting on `reddit`

Questionnaire 2

The second questionnaire which will be considered here was submitted to /r/TheoryofReddit, /r/Favors, /r/ineedafavor and /r/misc (all relatively small subreddits at the time) by students at the Danish Royal School of Library and Information Science in December 2011. The creators of the questionnaire returned several days later to post a summary of the around 440 responses they received (redd.it/n4vl2). One suspects that the respondents to this questionnaire were again quite 'heavy' users of `reddit`. Of the around 440 respondents 99% had an account, 99% had added subreddit subscriptions and 96% had un-subscribed from default subreddits (with 30% indicating that they were subscribed to more than 40 subreddits). Two sets of items are of particular interest here and the responses to these have been graphed and displayed in Figures 6.31 and 6.32.
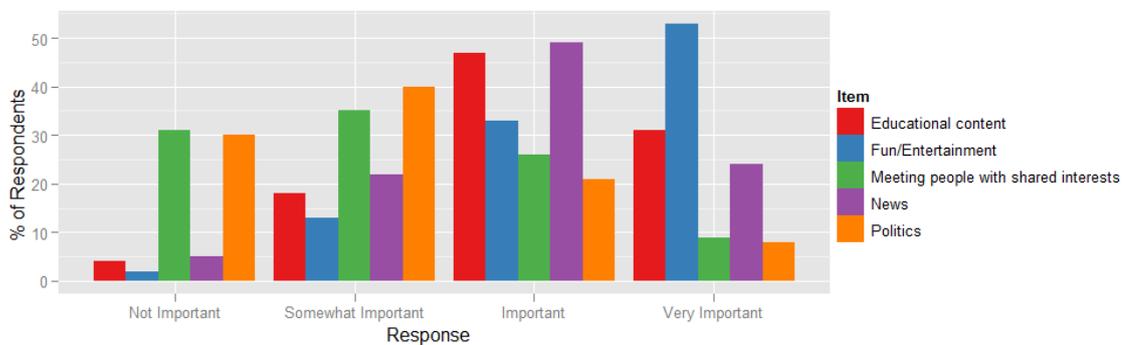
Figure 6.31: Showing responses to question: 'How important are the following when you browse `reddit.com`?' Bars represent the percentage of respondents who selected each response option - coloured to indicate the item which was being responded to.

Figure 6.31 shows that the breakdown of responses to the item about News is similar to that observed in responses to the earlier survey which was part of the present research. In addition, fun or entertainment appears to be a very important aspect of `reddit` for these respondents, with educational content having a similar profile of responses to that for News. Politics and 'Meeting other people who share my interests' are rated as less important by these respondents.

Figure 6.32 shows a similar pattern of responses regarding users' commenting, voting and post submission activity to that observed for the other questionnaire. One surprising aspect of these results is the fact that only 65% of respondents 'Always' view the Front page. Viewing the Front page and visiting subreddits have similar patterns of response - but one must bear in mind that this questionnaire was circulated through non-default subreddits so the sample is bound to exclude users who do not venture beyond the default subreddits and Front page. This is illustrated by the fact that no respondent indicated that they never or rarely visited subreddits. There may also be an element of confusion here as to what the 'Front page' is - some respondents may have interpreted this as what I refer to as the 'default Front page', and if they had significantly altered
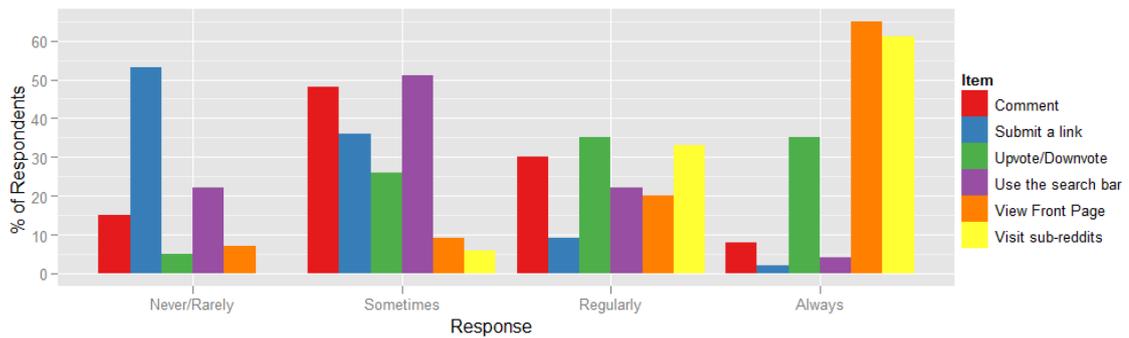
Figure 6.32: Showing responses to question: 'How often do you do the following when browsing `reddit.com`?' Bars represent the percentage of respondents who selected each response option - coloured to indicate the item which was being responded to.

their subreddit subscriptions (as most of these respondents had) they may believe that they do not 'view the Front page' any more. The alternative (and my own personal) interpretation of the 'Front page' is the aggregated 'Hot' page showing posts from subreddits the user is subscribed to. If respondents did interpret 'Front page' in this way then their responses suggest they are as likely to browse `reddit` one subreddit at a time rather than through the aggregated view - and this would be an unexpected trend.

6.1.8 `Reddit` Users Discussion

The portion of this chapter which deals with `reddit` has considered who `reddit`'s users are, how they use the website, and what their motivations are (so far as is possible with the available data). The results of demographic surveys disseminated through `reddit`'s Front page paint a picture of a user-base which is quite homogenous with respect to gender, age, ethnicity and location. The measured demographic variables are also consistent between the two surveys (circulated in different places and seperated by 9 months), suggesting that this is a reliable way to find out about `reddit`'s users and that the website's user-base is not becoming *more* homogenous demographically over time.

However, with respect to their level and type of activity, `reddit`'s population of users are a heterogeneous group. There is a power law type distribution of activity between `reddit`'s users, with many users who participate infrequently and a small number of participants who register a large number of actions. There is evidence that users at the extremes of this activity distribution are different in terms of how they participate, in addition to how much they participate. The majority of `reddit` users prioritise either voting or submitting to the exclusion of the other activity.

Among users who submit content, those who have been using the website for longer appear to have an advantage - the posts they submit tend to fare better than those submitted by newer users. This

effect has two possible origins which cannot be distinguished using the present data. It may be that individuals who have been reddit users for longer are simply more in tune with the website's community and are better able to decide which items will be appreciated by other users and how these should be presented to maximise their voting performance. Each reddit user account has an automatically generated 'profile' page which displays the user's recent posts and comments, and also the age of their account and their post and comment 'karma' scores. It is also possible that voters on reddit make use of this information, and give preference to more established reddit users when deciding which posts to up-vote and which to down-vote.

Users who participate in a more rounded fashion (i.e. voting and submitting) also seem to be at an advantage when compared to users who only submit content. As voting behaviour is anonymous this advantage most likely results from a greater affinity with the website's community or increased knowledge of how the website operates through experience gained while voting or browsing.

If we conceive of the attention available on reddit as a Common Pool Resource (Ostrom, 1990; Rheingold, 2003) - it appears that 'free riders' who would consume this resource (by submitting posts) but do not contribute their own attention to the common pool or participate in the administration of the resource (by voting *fairly*) have little success when they submit. This could be seen as a resistance to 'free riders' emerging from the collective voting behaviour of reddit's users - but the mechanisms through which this operates are unclear as users do not have access to complete information on the participation of others.

In addition to its primary purpose of ranking and sorting content, the voting system employed by reddit provides feedback to users about the content they have submitted. This feedback seems to encourage new users to continue submitting content;,with new users whose first post is voted on being much more likely to submit a second post. There is also evidence of a trend whereby users get better at submitting content which is well received with practice - similar to trends reported by Halavais (2009) and (Lampe and Johnston, 2005) on Digg and Slashdot respectively.

Among users who vote a set of behaviours were identified which are associated with more active users. Very active users (defined here as users who voted more than 500 times in the month) are more likely to use down-votes, more likely to vote on new posts, and more likely to vote several times in quick succession than their less active peers. These behaviours may be associated with a desire to increase the impact of one's actions on the collective decision being made through reddit's voting system. The association between votes which are down-votes, 'early' and 'quick' is indicative of a class of 'superparticipants' Graham (2013) who register a very large number of votes of these types. The nature of these votes is also such that it is consistent with an intent of curbing other users' capacity to contribute to the discourse.

Section 6.1.6 sought evidence of superparticipants characterised by their post submission behaviour

in longitudinal data on reddit's Front page. A relatively small set of users were identified who had submitted multiple posts which appeared on the default Front page. These users can be thought of as 'agenda-setters' (Graham, 2013) because when a post appears on the Front page it will inevitably receive a large number of comments. Graham (2013)'s concept of 'super-posters' seems more relevant to comments on reddit - as they were reported to predominantly post responses to other users and messages which were conversational in tone. The lack of suitable data on comments means that the presence of an equivalent class to 'super-posters' on reddit has not been investigated.

The structure of Social News websites is such that they give the impression that content has been submitted by community members and ranked and sorted by the same group of users. The emphasis these websites place on their Front page paints a picture of a relatively homogeneous community, and `reddit` users often make reference to the `reddit` 'hive mind'. Through this emphasis on its Front page `reddit` appears to have its own sense of identity (e.g. political content appearing on the Front page has a strong liberal bias, posts which concern the mass media are also common, and almost all of these are critical). Through the website's voting system users shape this identity, ostensibly on an equal footing. It is tempting to assume that `reddit`'s Front page reflects the opinions and wishes of its community of users.

Results presented here suggest that the reality of how `reddit` operates may be much more nuanced and fragmented. On `reddit` there are large sub-groups of users who only submit posts or vote, and also a small group of very active users who engage in a set of behaviours which appear to increase their influence relative to other users. The effectiveness of such behaviours or strategies remains to be seen, and this represents an important question going forward with research on these websites.

With regard to motivation, it is likely that the motivations of submitters and voters are different. Attention can be thought of as `reddit`'s 'currency' - the voting system serves to determine how much attention each item will receive. Attention is likely to be of central appeal to users who submit content, and `reddit` is attractive in this context because one has the capacity to reach a very large audience very quickly. The voting system which determines who or what will receive the huge volume of attention on offer is at the very least logical and understandable, it may also appear to users as 'fair' - with any attention being gained through `reddit` having been gained legitimately. The focus which `reddit` places on Front page content adds another dimension to this - if ones' post or comment appears in a prominent location and is sufficiently memorable this may well become part of `reddit`'s 'culture', at least for a while. It is common to see references to items which appeared on the Front page weeks or months later, in unrelated contexts.

There is one example in particular which springs to mind that illustrates the value of the attention garnered by `reddit`'s Front page. In this case a user (Prufrock451) submitted a series of comments to an /r/AskReddit post asking "Could I destroy the entire Roman Empire during the reign of

Augustus if I traveled back in time with a modern U.S. Marine infantry battalion or MEU?" (redd.it/k067x). Initially this user submitted a comment which took the stlye of a first diary entry from the hypothetical situation, positive feedback from other users (both comments and up-votes) encouraged them to continue writing these comments and also offered advice on factual inaccuracies. The endeavour soon spawned its own subreddit where it could be continued (now has 10,000 subscribers) and some time later the series of comments were optioned for a movie by Warner Bros (Solon, 2011). Reddit user Prufrock451 conducted an IAmA interview on the subject more recently (redd.it/11kt8o), and is known among at least some sub-set of `reddit` users as the individual who had their `reddit` comments turned into a movie. If and when this movie is released one would imagine that this incident might be recounted on the Front page again - and future users of `reddit` will come to know about Prufrock451 in this manner.

One aspect of submitting which has not been addressed here is `reddit`'s reputation system ('karma' scores). When one submits a positively scoring post or comment the score achieved by this item is added to the user's karma score, over time users who submit 'good' posts will steadily accumulate a larger karma score and this may be viewed by themselves or others as evidence that they are a valuable member of the community. Karma is, in one sense, a 'crystalised' form of the attention one has received, in a week's time your post may not appear on the Front page any more but you still have the 2,500 karma points to show for it. None of the analyses presented here address karma directly, but the frequency with which discussions of karma (and in particular 'karma whoring' - submitting content purely to accumulate karma points) have been observed on `reddit` suggest that it is on the minds of `reddit`'s users. Figure 6.33 shows an image post about one strategy for accumulating karma which appeared prominently on `reddit`, presented as evidence that `reddit`'s users think and talk about karma.

There is also anecdotal evidence that users may seek to punish another user by down-voting many of their submissions and thus reducing their karma score. This can happen when a user is seen to have wronged the community in some way (e.g. by soliciting donations for some cause fraudulently) - the offending user may find that not only is the post or comment where they committed the offense down-voted heavily - but that many of their previous contributions also suffer heavy down-voting. Where this behaviour is discussed it is described in terms of taking away the offending user's karma points as punishment.

While one can see how an 'attention-based economy' might serve to motivate users who submit content to `reddit`, it does not speak directly to the motivations of voting users. Voting is conducted anonymously, and this would seem to preclude all motivations relating to receiving attention or recognition for one's contribution. I would propose that there are broadly two sets of motivations (outlined originally in section 3.3.1) which are relevant to `reddit`'s voting users. The first set relates to being a member of the '`reddit` community' - `reddit` is a website which requires its

Figure 6.33: An image created by a `reddit` user about 'karma bombs'

users to vote in order for it to function, users may vote because they want to contribute to this endeavour, to play their part in ensuring that `reddit` continues to function, and maybe with the hope that they can help to improve it.

The second set of motivations I propose are related and concern being involved in a decision-making process which is seen as important or worthwhile. Users may vote because they want to be involved in deciding which posts appear in prominent locations. Some users may be happy to casually cast their votes and know that they have contributed, while others may strive to maximise their impact on the decision-making process, to imprint their own opinions or personality on `reddit`'s collectively determined 'identity'. The prevalence of 'expert' voting behaviours, in particular 'quick voting', suggests that there are at least some (heavy) users who aim to maximise their say in the decision-making process. Consider the user who cast 23,376 votes (20,230 of these being quick

votes, 17,596 being 'early' down-votes) in the month of March 2009. This feat must have required an effort of some magnitude, what did this user hope to achieve? Other `reddit` users do not know who this individual is, and the user's efforts have not been publicly displayed or acknowledged. Such behaviour can only be accounted for as at atempt to influence `reddit`'s voting system, and by extension leave an imprint on the website's identity, perhaps with the goal of pushing users' perception of `reddit` in a certain direction by trying to exclude certain types of post from prominent locations.

It is my assertion that voting on a Social News website like `reddit` is largely motivated by this concept of 'being involved in making a decision which is perceived as important'. Here the decision about which posts are displayed on the Front page is important in two ways - these posts will receive a lot of attention (sometimes enough to 'break' a story so that it filters into the wider public consciousness - see section 7.2), and they will also go on to shape peoples' perception of what `reddit`'s purpose is, indirectly defining the website's identity.

'Casual' voters are satisfied by knowing they have contributed to the endeavour, that they are 'part of `reddit`'. There also seem to be users who wish to dominate this decision-making process, to bend it to their will - these users understand how `reddit` works and are prepared to put considerable time and effort into leaving their mark.

The major questions in this area going forward are: how effective are heavy users who exhibit 'expert' voting behaviours? Are these strategies effective, and do these users exert disproportionate influence in the decisions about which posts appear most prominently? These questions take on a much greater urgency when we consider once again the application of a DM system like `reddit`'s in the domain of politics, if a small group of dedicated participants were able to effectively 'veto' items this would be highly problematic and would effectively nullify the system's potential.

6.2   Users of `Stackoverflow.com`

This section considers the users of `Stackoverflow` in terms of their level and type of activity. The analyses presented here diverge from those presented for `reddit` because the available data is different. Voting data for `Stackoverflow` is anonymous, it is not possible to tie specific votes to specific individuals. `Stackoverflow`'s data covers the period from the website's launch until present, and therefore the measures of user activity relate to their full history of participation, as opposed to the one-month snapshot considered for `reddit`.

6.2.1 Users' level of activity

There are 756,694 user accounts in the `Stackoverflow` data. This section considers the distribution of various forms of activity between these users. Analysis of one month's data from `reddit` revealed the prevalence of power law type distributions of activity between users, and similar patterns seem to be in evidence on `Stackoverflow`. Figure 6.34 shows the distribution of answer submissions between users -figures showing the distribution of votes, questions and comments show a similar profile and have not been reproduced here.
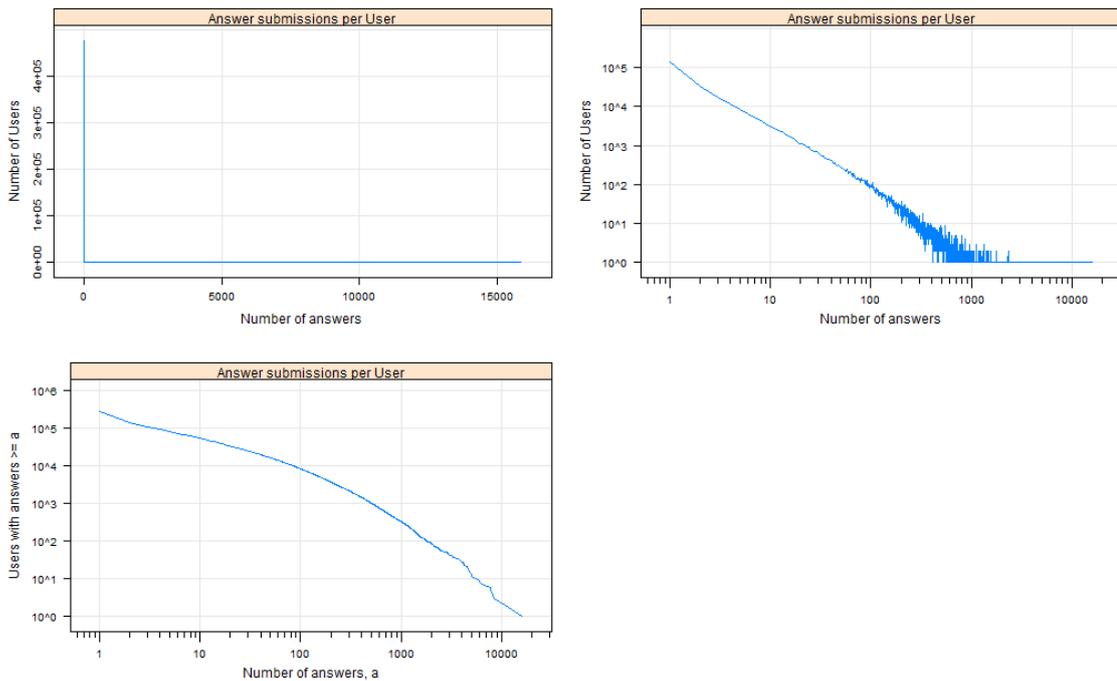
Figure 6.34: Showing the distribution of answer submissions between users on `Stackoverflow`

6.2.2 Users' types of activity

The analysis of `reddit`'s users revealed a high degree of specialisation in the ways which users participated. This section looks for similar patterns on `Stackoverflow`. However, `Stackoverflow` and `reddit` serve different purposes. Where voting is central to `reddit`'s functioning (because it determines what will be displayed on the Front page and these items are shown to the widest audience possible) this is unlikely to be the case on `Stackoverflow`.

`Stackoverflow` is geared towards generating good answers to questions and making these question/answer diads accessible to subsequent users (so that the same question does not need to be asked and answered every week). On navigating to `Stackoverflow`'s 'Front page' one is presented with new questions that do not have (many) answers yet. It is possible to sort questions by their voting score, but this is just one of the many criteria by which questions can be selected and

displayed. Answers are by default sorted by their voting scores, so votes on answers are likely to 'carry more weight' than votes on questions. However, an answer's voting score is still trumped by 'answer acceptance' - whichever answer the user who submitted the question accepts will be displayed at the top of the page regardless of score. On `reddit` we hypothesise that being involved in making a decision which is perceived as important is one of the factors which motivates users to vote - on `Stackoverflow` this motivation seems much less likely to come into play. Unfortunately user voting behaviour is one of the few aspects of `Stackoverflow` on which data are not readily available - so it is difficult to make any direct comparisons between voting behaviour on `Stackoverflow` and `reddit`.

On `Stackoverflow` users have a much greater range of posssible behaviours open to them than on `reddit`. `Stackoverflow` users can ask questions, submit answers, comment on questions or answers, vote on questions, answsers and comments, tag questions, edit questions/answers, offer 'bounties' for the answers to questions, approve or roll back other users' edits, among others. Some of these possibilities (e.g. adding tags to questions) are obviously geared towards making the knowledge produced on the website navigable for subsequent users.

Questions and Answers

As `Stackoverflow` is a website which deals primarily in questions and answers these are the first types of user behaviour considered here. Are the users who submit questions and answers the same users, or are there users who tend to submit more of one than the other?

Adamic et al. (2008) looked at one month of activity from `Yahoo Answers` and found several patterns of activity which were associated with different types of category. Of the categories they considered the 'programming' category is most similar to `Stackoverflow`. They found that users who asked questions and provided answers tended to operate in one capacity or the other but not both. In the month their analyses relate to 57% of users who asked a question did not submit any answers, while 51% of users who answered a question did not ask any.

In the data for `Stackoverflow` 62% of users who asked at least one question did not provide any answers, while 49% of users who provided at least one answer did not ask any questions. These percentages are similar to those reported by Adamic et al. (2008), but in the context of `Stackoverflow` they are even more striking because the measures reflect users' full history of participation as opposed to a one-month snapshot. Figure 6.35 shows the proportion of all users' question/answer submissions which are answers, and supports the idea that many of `Stackoverflow`s' submitting users submit questions or answers exclusively.

On `reddit` the likelihood of a user being a voter was much higher for more active users. On `Stackoverflow` we might expect those users who answer rather than ask questions to be more
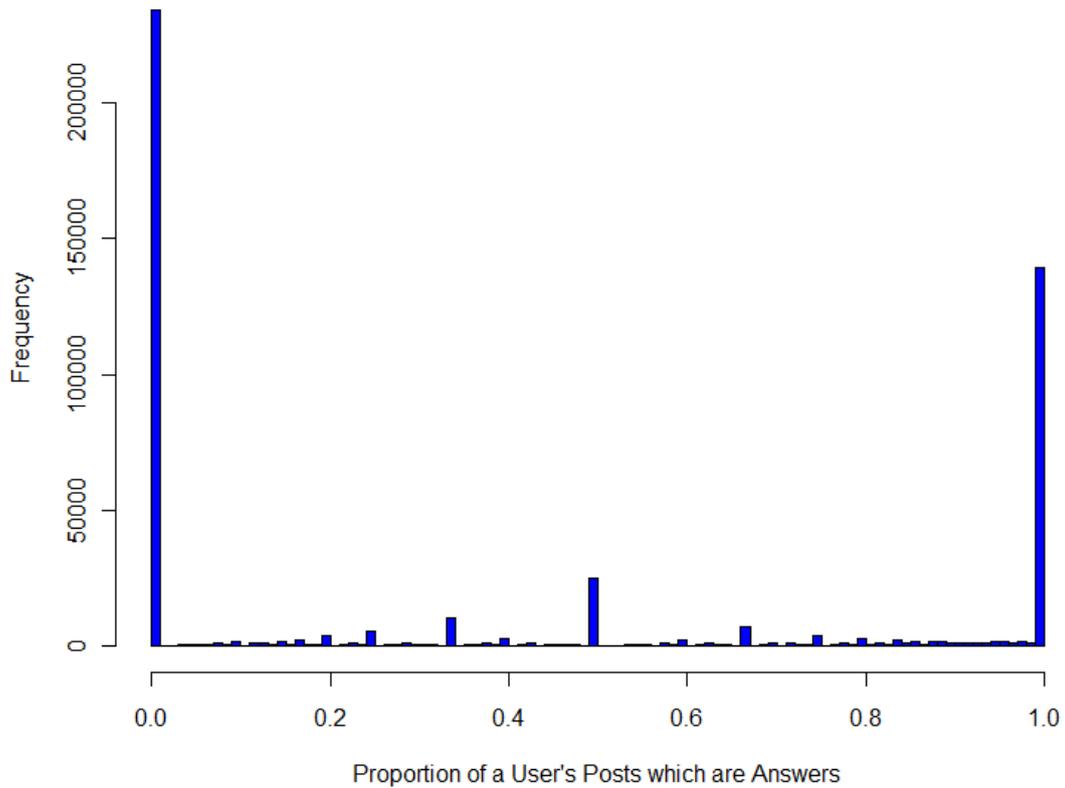
Figure 6.35: Showing the proportion of users' Question/Answer submissions which are Answers

experienced or active. Kernel density plots (Figures 6.36 and 6.37) suggest that users with between 11-50 question/answer submissions are equally likely to engage in asking questions or providing answers and often engage in both types of activity. When we look at more active users however, those with between 51-100 question/answer submissions, the pattern is quite different, with most of these users submitting many more answers than questions.

At this point users were binned according to their total number of question and answer submissions. Figure 6.38 reveals the expected relationship between activity level and proportion of activity which is questions/answers. Users with a higher number of question/answer posts are much more likely to be submitting answers than questions.

Where do `Stackoverflow`'s user contributions come from?

This section considers users as divided into fifteen bins according to their number of question and answer submissions. What proportion of the questions and answers on `Stackoverflow` originate with more or less active users? Figure 6.39 shows the percentage of all questions, answers and accepted answers which originate with users in each of the fifteen activity bins, and also the

User Question and Answer submissions (Users with 11-50 posts)
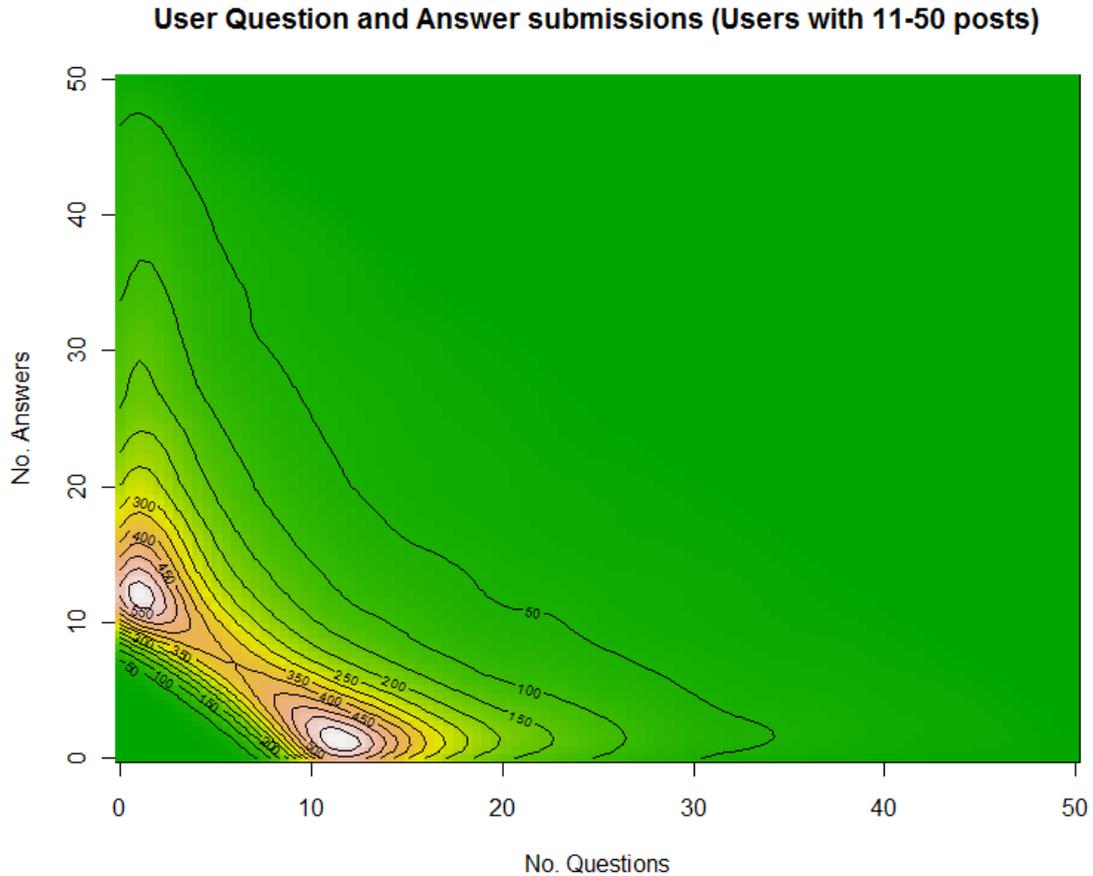
Figure 6.36: Showing the pattern of question/answer submissions for users with between 11 and 50 posts

proportion of `Stackoverflow`'s users who fall into each activity bin. Low-activity users account for a sizeable proportion of question submissions but these users have not provided many answers, and very few of the answers which were accepted by the questioning user came from these low-activity users. The bin representing 9,942 users with between 64 and 128 question/answer submissions appears to be the most balanced with regard to activity types - this group accounting for 14% of all questions, answers and accepted answers. The bins representing the 5200 users with 128-256 actions and the 2208 users with 256-512 actions account for around 15.5% and 14.5% of answer submissions respectively. Beyond the bin with 128-256 users two interesting trends emerge - the number of questions asked drops off sharply, and the bins account for a larger proportion of accepted answers. This suggests that users with more than 256 actions are submitting better answers in addition to a higher rate of answers per user. This trend is exemplified by the highest activity level bin representing just four users who between them have provided 2% of all the accepted answers on the website.

We now move to consider other forms of interaction which users engage in on `Stackoverflow`.

User Question and Answer submissions (Users with 51-100 posts)
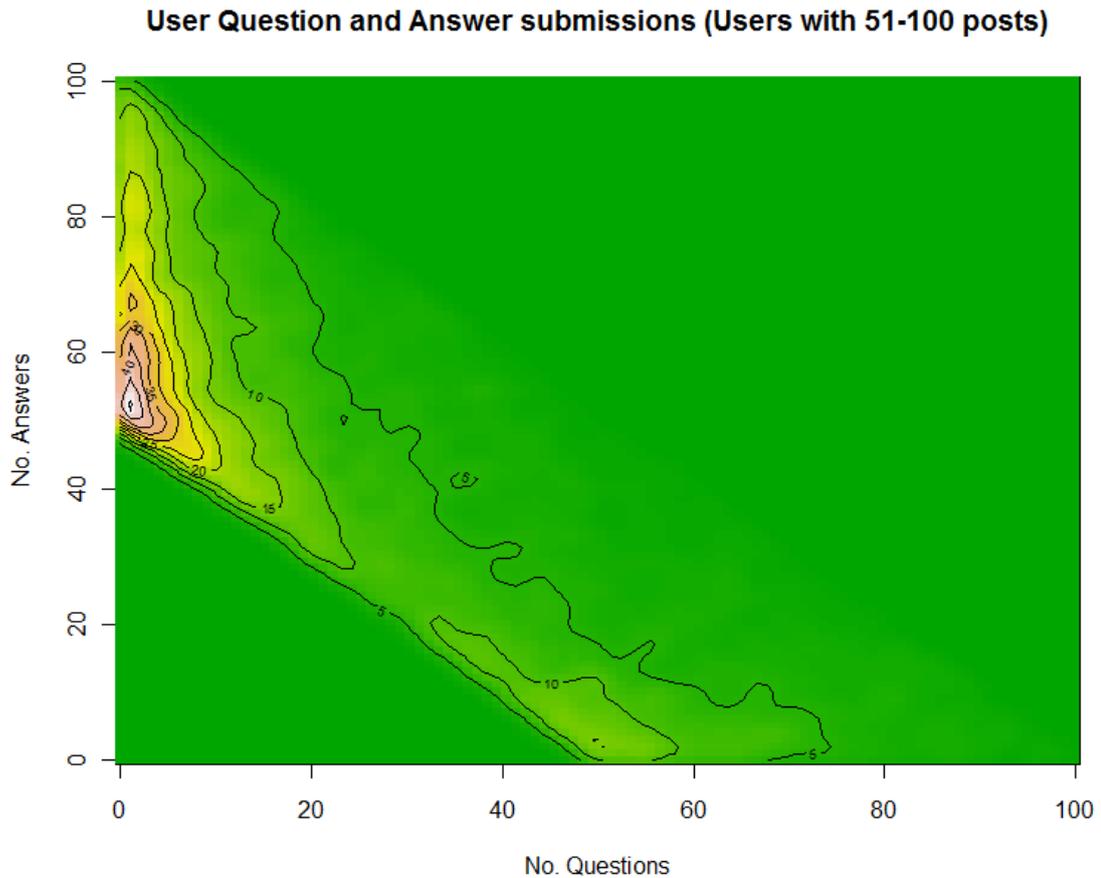
Figure 6.37: Showing the pattern of question/answer submissions for users with between 51 and 100 posts

Figure 6.40 shows the percentage of answers, comments, votes, edits and reversions which originate with users in each of the activity bins - answers are also included as a reference to the previous graph. The activity bin a user falls into is still determined by their total number of question/answer posts. The distribution of comments between activity bins closely follows that of answers, giving an initial indication that users who submit answers likely also submit comments. The distribution of votes follows an interesting trajectory. Users with low numbers of posts also make very few votes while the activity bins incorporating users with between 16 and 512 posts account for a disproportionate number of votes, however the prominence of voting drops off for users with more than 512 post submissions. Conversely, edits of another user's posts (and rollbacks of previous edits) come predominantly from users with higher numbers of question/answer submissions.
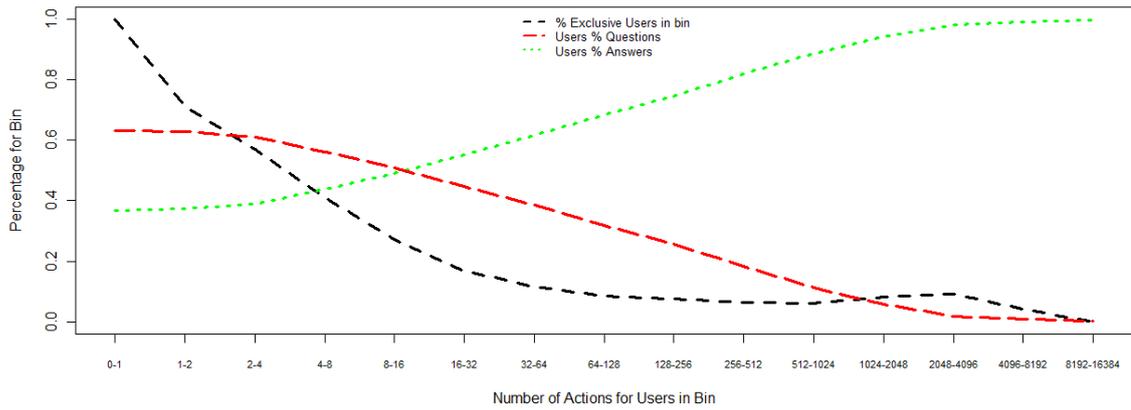
Figure 6.38: Showing the percentage of 'exclusive' users in each of the activity bins, and the mean proportion of question and answer submissions for users in each bin
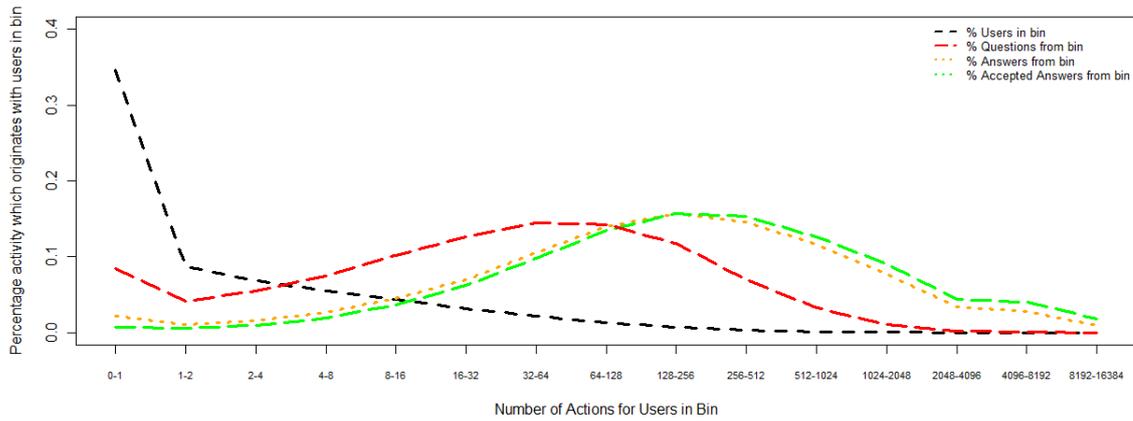
Figure 6.39: Showing the percentage of questions, answers and accepted answers which originate with users in each of the activity bins, and also the percentage of users who fall into each bin.
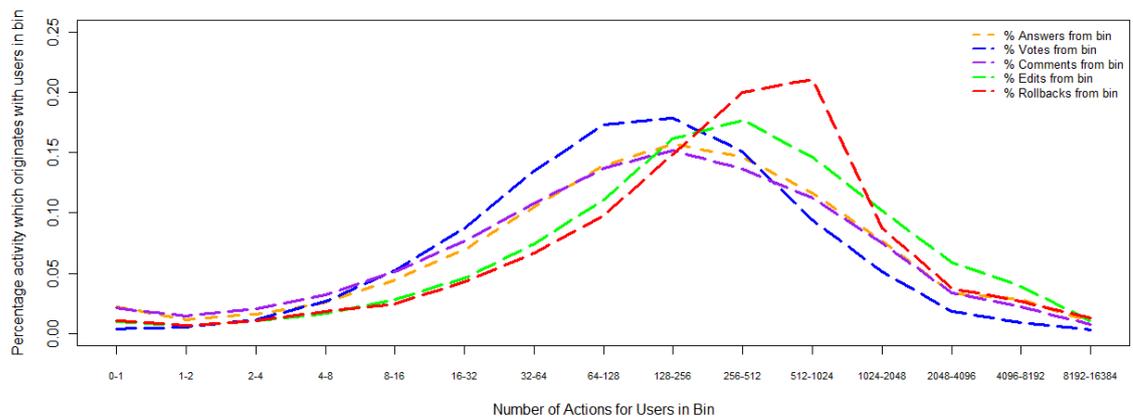
Figure 6.40: Showing the percentage of answers, comments, votes, edits and reversions which originate with users in each of the activity bins

6.3 Chapter Summary

There are some common patterns to user behaviour on `reddit` and `Stackoverflow`. Firstly, the distribution of activity between users is highly skewed - most users are casual and their accounts are associated with a low level of activity, but at the other end of the scale there are a small number of highly active users. This pattern is unsurprising and has been documented in a wide range of voluntary endeavours (Wilkinson, 2008).

Secondly, on both websites there appears to be a tendancy for users to specialise in a particular form of activity. On `reddit` voting can be thought of as the core activity, and there are users who specialise in voting - they vote very frequently and in ways which are distinct from their less active peers. On `Stackoverflow` the answering of questions can be thought of as the core activity, and `Stackoverflow`'s highly active users tend to specialise in this form of activity. In the same manner that `reddit`'s highly active users engage in 'expert' voting behaviours - there is evidence that the answers which highly active users on `Stackoverflow` provide tend to be of higher quality than their less active peers.

There is however a difference in the effect which highly active and specialised users are likely to have on these two websites. On `Stackoverflow` the presence of these users is most likely beneficial to the website as a whole - because they provide good answers and that is the website's primary purpose. On `reddit` the value of highly active users is not as clear - they undoubtedly perform useful tasks (e.g. voting on New posts and excluding the poorest submissions) but the manner in which they do so (e.g. quick voting) raises questions about their motivation. Whether these users are on the whole beneficial to `reddit` is unclear, and depends on the precise criteria on which the website's success is being measured.

If one values a DM system like `reddit`'s for its capacity to *accurately reflect the opinions of its users* then these highly active users would be seen as problematic - because by investing more time and voting in certain ways they are exerting a disproportionate level of influence on `reddit`'s voting system. If, on the other hand, the purpose ascribed to `reddit`'s voting system is to place 'good' content on the Front page, then the value of highly active users would be determined by whether they vote in ways which are consistent with the desires of `reddit`'s users and visitors more generally. From the perspective of `reddit` as a company 'fanatical' levels of voting are probably not seen as inherently problematic because it means that some proportion of users are spending a substantial amount of time on the website. If one is assessing `reddit` as a company this is a positive attribute because that company's success is intimately tied to the level of user attention and participation the website receives.

Relating this back to Distributed Moderation more generally - we might prefer to assess the DM system's performance based on its capacity to accurately reflect the opinions of its contributing

users. In this case it may be preferable to place limits on the frequency with which users can vote, sacrificing the system's capacity to attract very heavy users in favour of an end product which is a better reflection of the 'average user's' opinions.

Chapter 7

The Social Significance of Social News

Previous chapters have considered how Distributed Moderation works, who the participants are and how they behave. This chapter will build on this knowledge of how DM works and consider how the nature of the DM infrastructure influences the resulting product. Where previous chapters have considered large data-sets representing all of the activity for a given period of time this chapter will consider narrowly defined case studies in much greater detail.

This chapter is focused entirely on `reddit`, because for the `Stack Exchange` websites the picture is much clearer - these websites receive questions and provide a framework through which they can be answered and the resultant knowledge organised. The `Stack Exchange` websites seem to meet their purpose well, and their broader social significance comes from providing access to such expert knowledge to anyone for free - this will be discussed further in the final chapter.

The broader social significance of `reddit` is more difficult to pin down, because it is intimately related to what `reddit`'s users choose to put on the website's Front page. `reddit` is described here as a 'Social News' website, and `reddit`'s broader social impact is perhaps more likely to come from subreddits which deal with content that is more serious or 'news-worthy' - than subreddits which deal with content of a humerous or entertaining nature. With this in mind, the decision was taken to study the coverage of particular news stories on `reddit` - to find out whether `reddit` can be said to 'cover' a story in the same manner as a conventional media outlet, and what the specific characteristics of `reddit`'s 'coverage' might be.

The initial approach here was to study coverage of the story on `reddit` and through a selection of conventional news sources (i.e. newspapers and television news channels) - comparing the nature of coverage from each source. However, it quickly became clear that understanding how these stories

unfolded on `reddit` would require a significant investment of time - and that there was effectively no limit to the time which could be invested in exploring this. In analysing a newspaper's coverage of a story one is limited to analysing the materials which were actually published, but on `reddit` one has access to all of the posts which did not reach the Front page and one can also look at the process whereby some posts but not others were selected to appear on the Front page. It was therefore felt that research time would be better spent on `reddit` as DM websites are the focus of the research and the systems to gather data from `reddit` were already in place (whereas no procedures were in place to collect the text of newspaper articles).

Where comparisons are made between `reddit` and conventional news we therefore rely on a general understanding of how newspaper articles or television bulletins tend to be presented - e.g. Glasgow Media Group (1976). There are aspects of `reddit`'s 'coverage' (e.g. the prevalence of swearing) which are obviously distinct from what one would expect of a conventional news source. Such major discrepancies between coverage on `reddit` and in the conventional news media can be identified based on prior experience without the necessity of studying a specific newspaper's coverage of this particular story. In fact the predominance of major differences in style and content would make it difficult to conduct a direct textual comparison between `reddit` and newspapers. We must also bear in mind that much of `reddit`'s news coverage comes in the form of links to external sources (including newspapers' websites) and in this sense `reddit`'s coverage subsumes that of the conventional news media. A more fruitful approach here is to consider the types of source which are linked to as part of `reddit`'s coverage (be they newspapers, blogs, videos, etc.) and what these sources have to say about the story in a general sense - rather than becoming mired in a detailed textual analysis of the specific sources linked to by `reddit` posts as compared to some other set of sources which would be taken as representing the conventional news media.

It is also important to bear in mind that as `reddit`'s userbase and profile have grown the website will increasingly play a part in the broader 'hybrid media system' which characterises countries like the USA and UK where most of `reddit`'s users live (Chadwick, 2013). There is little or no published research on this, but building on previous Chapters it is safe to conclude that `reddit` already demonstrates back-and-forth interplay with other media. Many of the posts on `reddit` are themselves links to media on other platforms both older and newer (e.g. TV/Newspaper websites, youtube, twitter, blogs) and we can assume that when these posts fare well they boost the amount of attention the items will receive. Is there any evidence that elite actors from the older media are now paying attention to `reddit` in the same way that they monitor `twitter` and a selection of blogs? (Chadwick, 2013)

The Hybrid Media System as described by Chadwick (2013) leaves little room for non-elite actors ('members of the public') to influence the makeup or reporting of the news. On occasions where this does occur it tends to be through the remediation of elite actors from the older media, who

hence retain their role of 'gatekeepers' of the news. One example described involved a blogger who conducted their own investigation of a source for a major story, then tweeted about their blog, had their tweet re-tweeted such that it was seen by a Channel 4 News reporter and hence had an affect on that story through this avenue. Chadwick (2013) also reports an interview with a blogger who deliberately followed such a strategy, and who acknowledged that their blog's primary target audience was a small coterie of journalists who may pick up on the blogger's stories and channel attention back to the blog through a link in their own report. There was also some ambiguity expressed about whether newer media has the capacity to 'break' a story or whether this requires the involvement of an elite actor.

> "I don't think putting up a tweet or a blog post is breaking a story, except in some technical sense... The point at which it's broken is by the mainstream media, not the tweet but the whole story." - Will Straw, *Left Foot Forward* (Chadwick, 2013, p. 172)

Social News websites focus their users' attention on the Front page, and this is akin to a broadcasting mechanism. Can a website like `reddit` effectively break a news story by directing enough attention to it that it becomes impossible for older news media to ignore? In the US Election of 2008 Television news heavily influenced social media discourse (Chadwick, 2013) - does coverage of a story on `reddit` exert any detectable level of influence over discourse on services like `Twitter`?

Chadwick (2013) also explores new variations on collective action which the social side of the hybrid media system makes possible. One such organisation of particular interest, based in the UK, is *38 degrees* - and Chadwick (2013) describes this as following a similar structure to *MoveOn* in the USA. The organisation which runs *38 degrees* creates campaigns and requests 'actions' of its 'members' - with members indicating on the website that they intend to participate by performing these actions.

> "Members are asked to sign online petitions or send e-mails and make phone calls to their MPs. They are asked to show up physically at lunchtimes to protest in front of buildings around the country, as they did in 2010 against proposed cuts to the BBC. They are asked to organize flash mobs at parties' local constituency campaign gatherings, as they did in several targeted seats during the 2010 general election, to raise awareness of the lobbying industry." (Chadwick, 2013, p. 172)

38 degrees does not follow the 'peer production' model whereby all users are essentially equal - but rather has a clearly defined leadership who set the agenda and decide which actions are placed on the website. The relationship between *38degrees* leaders and members is an unusual one and could be described as analogous to a politican and their electorate or a marketer and their market. Leaders pay very close attention to the discourse of their members in determining which actions to attempt, and members 'vote' to determine the actions that will be selected by indicating their

intent to participate. Leaders 'test' ideas by emailing them to small samples of the member list - using early (non-)responses to determine whether a nascent action should be abandoned, and even sending out multiple versions of the email which frame the issue in different ways to determine which framings resonate with 'members'.

A relatively small staff monitors the hybrid media system and in particular members' attitudes towards the news stories dominating the 'political information cycle'. They formulate actions which members might take that could move the situation/story in a favourable direction. *38 degrees* does not attempt to set the agenda for the news - but takes its lead from the news as this is where the thoughts of its members are also likely to be (Chadwick, 2013). This suggests one framing through which the case studies presented here could be considered - does `reddit`'s News content follow trends in the mass media, or does `reddit` attempt to set its own agenda?

It should be noted that by late 2012 `reddit` appeared to be moving away from 'news' type content in favour of entertainment, a trend which will be probed further in chapter 8. If one looks at the default Front page for the first time now, in late 2012, it may be difficult to imagine this same default Front page playing host to the case studies outlined below. Furthermore, in late 2012 I would not argue that `reddit`'s broader social significance comes from its coverage of 'news' stories but rather the primacy of user-generated content.

After looking at the characteristics of 'Social News' through two case studies we will also consider the performance of `reddit`'s voting system in its task of allowing users to collectively determine which posts are 'the best' - this is achieved through a content-rating experiment.

Some of the research questions to be addressed in this chapter are as follows:

- Can `reddit` be said to 'cover' a particular news story in a similar manner to the conventional news media? What are the specific characteristics of `reddit`'s 'coverage' which distinguish it from coverage provided by conventional news organisations? Does `reddit` take its lead from the conventional news media or does it set its own agenda? How might `reddit` mesh with the wider Hybrid Media System (Chadwick, 2013)? Case studies of two stories (`Wikileaks`' release of US diplomatic cables in section 7.1 and the Stop Online Piracy Act in section 7.2) will be employed to address these questions.

- How does `reddit`'s up/down voting system cope with a divergence of opinions on a topic amongst its users? Does the 'majority rule', or is there room for conflicting opinions to be expressed and attended to? Sunstein (2002) puts forward a convincing argument for the importance of being exposed to a range of conflicting perspectives on a topic (see section 2.5.4), is this something that Social News websites can provide? Section 7.1 considers this question in the context of the `Wikileaks` story.

- How well does `reddit`'s voting system perform in its task of ranking items of content ac-

cording to the judgments of its users? Section 7.3 addresses this issue with an experimental approach.

- `Reddit`'s voting system allows users to collectivey rate each post and determine how visibly it will be displayed. Can this same voting system be utilised as an organisational tool? Are `reddit`'s users capable of collectively making higher-order decisions and implementing these? Section 7.4.1 investigates one case where high-scoring posts urge users to take a particular course of action which is amenable to empirical scrutiny.

In this Chapter a variety of methodological approaches are utilised - each case study was conducted in a different way to probe different questions and an experiment has been conducted to investigate the performance of `reddit`'s voting system. For the first case study, concerning `Wikileaks`, the principal approach was to categorise posts and analyse these classifications in order to address questions about the treatment of minority opinions on `reddit`. For the second case study, concerning the Stop Online Piracy Act, a different approach was adopted. This approach involves recounting the story's coverage as it unfolded, in much more detail than for the `Wikileaks` case study and with more emphasis on the timing of events.

7.1 Reddit's coverage of a story - `Wikileaks`' release of United States' diplomatic cables in November/December 2010

This section concerns a case study of `reddit`'s coverage of the release of United States' diplomatic cables by `Wikileaks` in November/December 2010. A study period of 4th November to 27th December 2010 was selected. This time period covers some time before the release of the first batch of cables, in which charges of rape or sexual misconduct were brought against `Wikileaks`' spokesperson Julian Assange. This time period also incorporates several weeks following the release of the first set of diplomatic cables, where reports about the cables constituted a considerable proportion of the output of the conventional News media. By the end of this study period coverage of this story in the conventional news media had subsided considerably, but isolated articles related to the story did continue to appear during January 2011.

Some of the themes to the conventional news media's coverage of this story are as follows:

- The rape allegations brought against Julian Assange on 18th November 2010 and subsequent extradition proceedings brought against Assange.

- The release of the first batch of diplomatic cables on 29th November, and a large number of articles about the content of individual cables.

- The reaction of governments to the release of the cables; most notably the United States' gov-

ernment's condemnation of the release of this sensitive diplomatic information; and discourse among American public figures about whether the release of this information represented (or should be treated as) a criminal offence.

- The severing of ties to the `Wikileaks` organisation by several large corporations; including the decisions of Paypal and Visa to stop processing donations for `Wikileaks` and the decison of Amazon to stop providing web hosting for `Wikileaks`.

- Distributed Denial of Service attacks against `Wikileaks`' website in the first instance, and against those companies which severed ties to `Wikileaks` in the second instance (in this case being orchestrated by the 'Anonymous' group).

Chadwick (2013) presents a much more detailed account how this story unfolded, the timing of events, and also insight into the evolving relationship between `Wikileaks` and its older media partners.

The key words 'wikileaks' and 'assange' were selected, and for a post or article to be included in the study its title or headline had to include at least one of these key words.

Data to be analysed from `reddit` comes from several front end sources. A complete record of all Front page posts is available; in the study period there were 249 posts with a key word in their title which appeared on `reddit`'s Front page - these posts received a combined total of 87,416 comments. During the study period the 'politics' subreddit was being monitored - so a record of almost every post submitted here is available, along with observations of its voting performance on `reddit`, its comments, and the voting performance of each of these. In total there were 3,093 posts submitted to the politics subreddit during the study period with at least one key word in their title. These politics posts received a combined total of 43,921 comments.

Four newspapers were selected for the purpose of comparison; *The Guardian*, *The New York Times*, *The Independent* and *The Times*. A total of 270 articles appeared in these newspapers during the study period with at least one key word in their headline. Analysis of the content of these newspaper articles has not been conducted, and only the timing of their publication is considered.

Figure 7.1 shows the number of articles per day of study which appeared on the `reddit` Front page and for the four newspapers which were selected. This graph reveals a strong correlation between sources in terms of the number of articles per day about the topic. This suggests that `reddit` users, through the voting system, have afforded this story a similar level of coverage to the newspapers being observed. It should be noted that many of the articles from the four newspapers being observed come from *The Guardian* and *The New York Times*, these newspapers were among a small group to be provided with the diplomatic cables in advance of their release, and as such the level of coverage afforded to the story by these newspapers is not typical of newspapers in general.
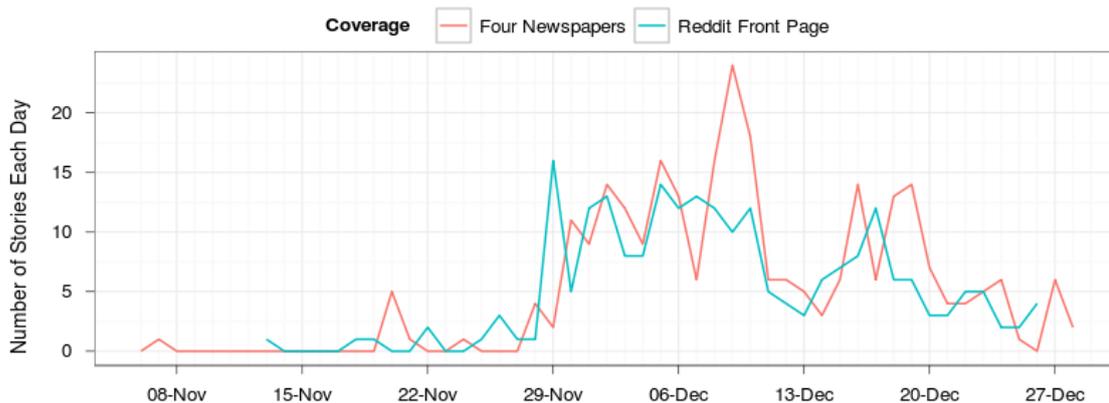
Figure 7.1: Showing the number of 'Wikileaks' posts per day which appeared on `reddit`'s Front page and the total number of stories per day in the four selected newspapers.

The majority of these `reddit` posts are link posts (they consist of a title which links to an external source). Figure 7.2 shows the types of source which were linked to by Front page posts, a good starting point for consideration of how `reddit` is integrating with the Hybrid Media System (Chadwick, 2013). 'Old' media sources such as the websites of newspapers and television news channels account for 50 and 21 Front page posts respectively. 'Web articles' (articles written by professional journalists for websites which do not have an off-line equivalent) are the most popular source being linked to by Front page posts (with 61 posts linking to these sources). Blogs account for 23 Front page posts, 17 link to `Twitter` streams and 6 link to `Youtube` videos.

19 Front page posts link to websites (i.e. the Main page of another website) - many of these are links to the `Wikileaks` website itself. These links were submitted when `Wikileaks`' domain names were revoked or it was forced to change servers - for example several are links to IP addresses where the website could be accessed when it did not have a domain. These posts can be thought of as a counter-measure against efforts to take the website off-line. When a domain was revoked or a hosting account cancelled, the `Wikileaks` organisation moved their website to another location, and through Social Media (e.g. `Twitter` and `reddit`) they could quickly spread word of this change. These cases represent one instance where `reddit`'s users, through the voting system, have determined that `reddit` will become directly involved in this story. In placing a link to the IP address or domain name where the `Wikileaks` website could be found *at that moment* on its Front page, `reddit`'s users were actively working against the various actors who sought to make the `Wikileaks` website inaccessible.

A similar series of events is described by Chadwick (2013, p.95) in relation to Wikileaks in 2008 - where their domain name was revoked after a legal ruling and the *New York Times* and CBS News published the IP address for `Wikileaks`' server. This could therefore be thought of as `reddit`
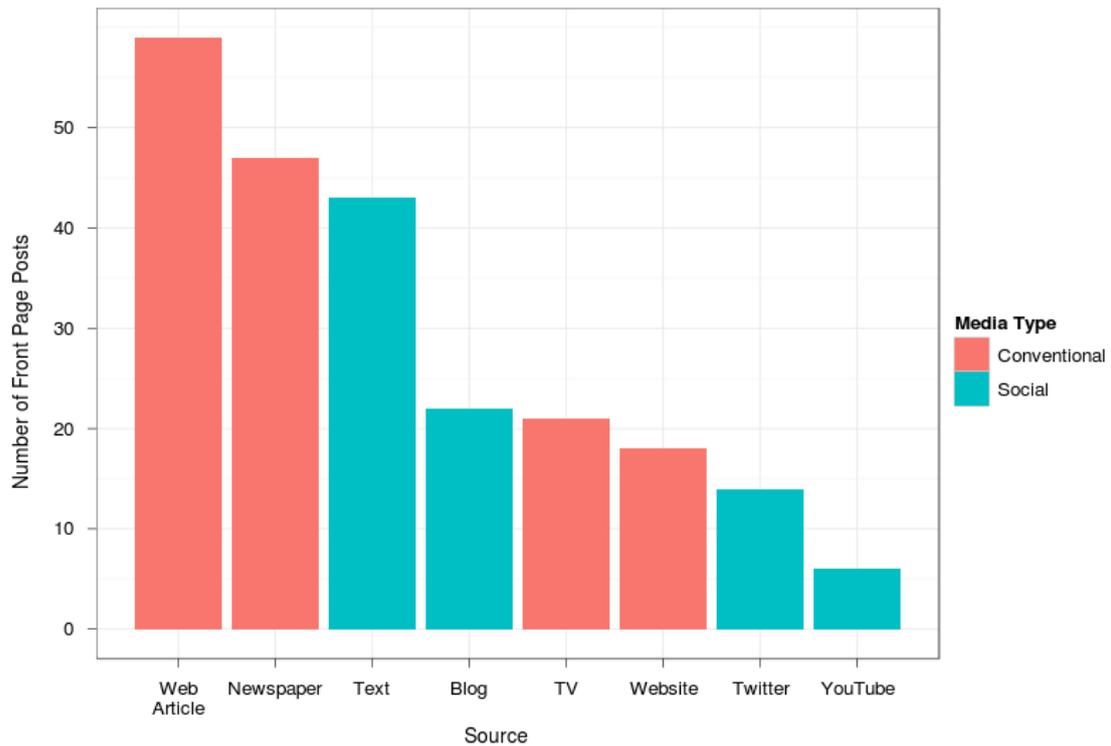
Figure 7.2: Showing the sources linked to by Front page posts on `reddit`

emulating past behaviour of an elite older media entity. In the face of Distributed Denial of Service attacks and pressure on companies to revoke hosting, the IP address for Wikileaks server was much more dynamic in this case - and perhaps `reddit`'s perpetually rolling Front page was well suited to this task, enabling fresh posts to appear on the Front page shortly after a new IP address was announced.

There were also popular posts which called on `reddit`'s users to participate in endeavours which aimed to make use of the leaked diplomatic cables, new subreddits were quickly established where work on these projects could be co-ordinated (/r/Leakspin and /r/WikileaksAnalysis) but these endeavours failed to acquire the critical mass of willing volunteers which would have been required to sustain them. `Reddit` also features posts about 'Operation Payback' - the series of Distributed Denial of Service (DDoS) attacks orchestrated by the hacktivist collective 'Anonymous'. Chadwick (2013) describes this operation as follows:

> "These actions were an essential aspect of the hybrid media system. They were a show of strength of sorts, by the members of an online anti-leader network who had become politicized and who were willing to take personal risks in order to demonstrate their support for transparancy, freedom of expression, and the principle of whistle-blower anonymity."(Chadwick, 2013, p. 108)

`Reddit` has in the past directly covered 'Anonymous' endeavours on a number of occasions, with

the nature of this coverage ranging from announcements or appeals for participation to reports on successful 'operations' and latterly the arrest of participants and the subsequent consequences.

This eagerness to become directly involved in the story as it develops is one way in which `reddit`'s coverage differed to that of the conventional news media - which tends to maintain a level of detachment from the story (with occasional exceptions such as that described above).

48 of the Front page posts relating to this story are 'self' posts - text entered directly on `reddit` by a `reddit` user. These represent an interesting category of content - they can be thought of as an equivalent to editorials or letters to the editor in newspapers. The difference is that in `reddit`'s case the users who submit these pieces can also be involved (through the voting system) in making the decision as to which opinions are 'published' on the Front page.

All Front page posts and a random sample of posts to the politics subreddit have been classified along a number of criteria: the subject of the post, its tone in relation to this subject, the source the post linked to, and the type of resource being linked to. Presented below is a breakdown of the subjects of posts which reached the Front page and the tone of the post with regard to this subject (tone is rated on a 5 point scale, from very negative to very positive, with a seperate 'direct' classification for posts which link directly to a web resource produced by the subject of the post - e.g. links to `Wikileaks`' current domain or IP address).

A random sample of 20 items were also rated by a colleague to generate a measure of reliability for these classifications. Only the subjects and tones of posts were rated by the second rater. Cohen's KAPPA (Kvalseth, 1989) was used to measure inter-rater reliability as this is a suitable statistic for both ordinal and nominal ratings. For the subjects of posts (a nominal variable) Cohen's KAPPA was calculated to be 0.61 (s.e. 0.12) and there was agreement on 13 of the 20 items. Cohen's KAPPA was also calculated for the tones of the 13 items on which the same subject had been selected (with the variable being weighted linearly because it is ordinal), on this variable the KAPPA statistic was 0.63 (s.e. 0.12) and the maximum distance between the two sets of ratings was 1 point on the scale. Both of these KAPPA values represent 'substantial' agreement, according to Landis and Koch (1977). Given that the second rater was not familiar with `reddit` or `reddit`'s coverage of this story this level of inter-rater reliability is sufficient evidence that these classifications are reliable.

Figure 7.3 suggests that many posts which reach the Front page are neutral with respect to their subject. It does appear however that posts which concern the governments' reaction to the release of the cables, or the reaction of large corporations (e.g. Paypal, Visa), or the coverage of the story by the mass media - are more likely to be negative than positive. Posts which appear on the Front page that concern Julian Assange or the `Wikileaks` organisation are more likely to have a positive tone. To take the posts which relate to the `Wikileaks` organisation as an example: 10%
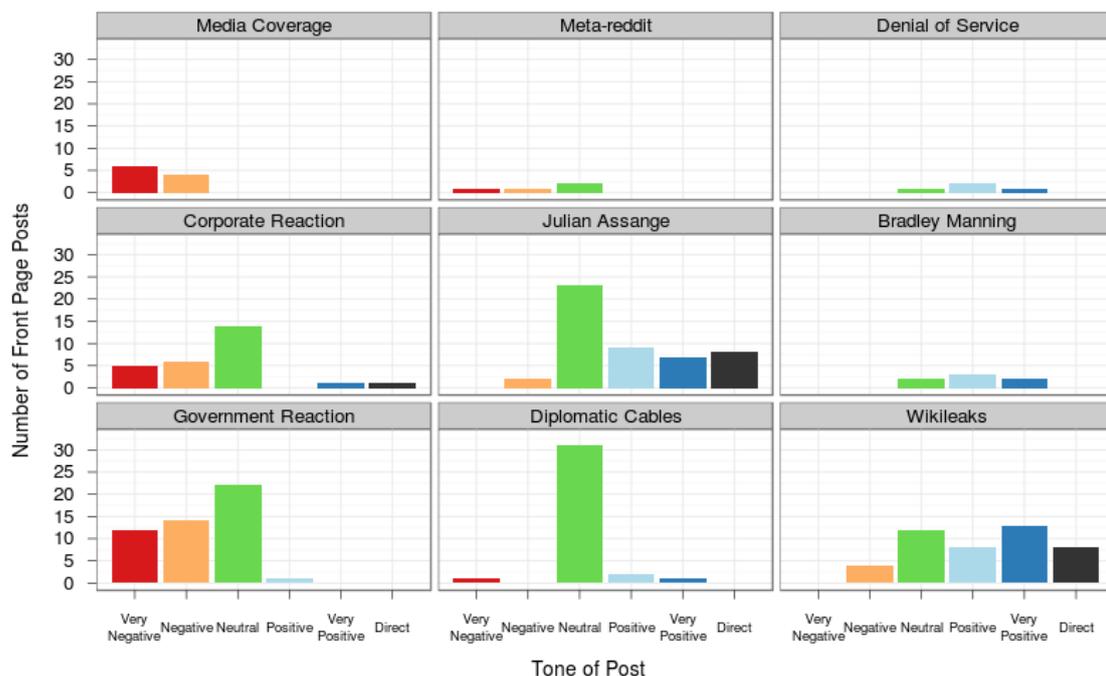
Figure 7.3: Showing the subject of Front page posts and the tone of the item being linked to with regard to this subject.

are negative, 32% are neutral, and 58% are positive. One important question to ask is whether these figures reflect the opinions of `reddit` users?

These measures suggest that `reddit`'s users are largely pro-`Wikileaks`, and critical of the government/business' reaction to the leaks. There are however a small number of posts which go against this trend, suggesting that `reddit`'s coverage of the story is not entirely one-sided. Posts which link directly to their subject are most likely to concern `Wikileaks` or Julian Assange. In the case of posts concerning `Wikileaks` these are links to the current domain/IP where the website could be accessed, or links to announcements on `Wikileaks`' twitter stream. In the case of posts concerning Julian Assange they are links to interviews conducted with him or articles/blog posts written by him.

The next question to ask is whether this bias in terms of Front page coverage is a reflection of a bias among `reddit` users, and if so how strong might this bias be? In the absence of reliable information about the opinions of `reddit` users the profile of text posts submitted to the politics subreddit will be taken as an indication of `reddit` users' opinions. These text posts come directly from `reddit` users and therefore should provide a good indication of the opinions of `reddit` users who submit posts and are active on the politics subreddit. They are also perhaps more likely to espouse a tone which is either positive or negative with regard to their subject. One must however bear in mind that in Chapter 6 we saw that users who submit posts and users who vote are to some degree seperate populations.
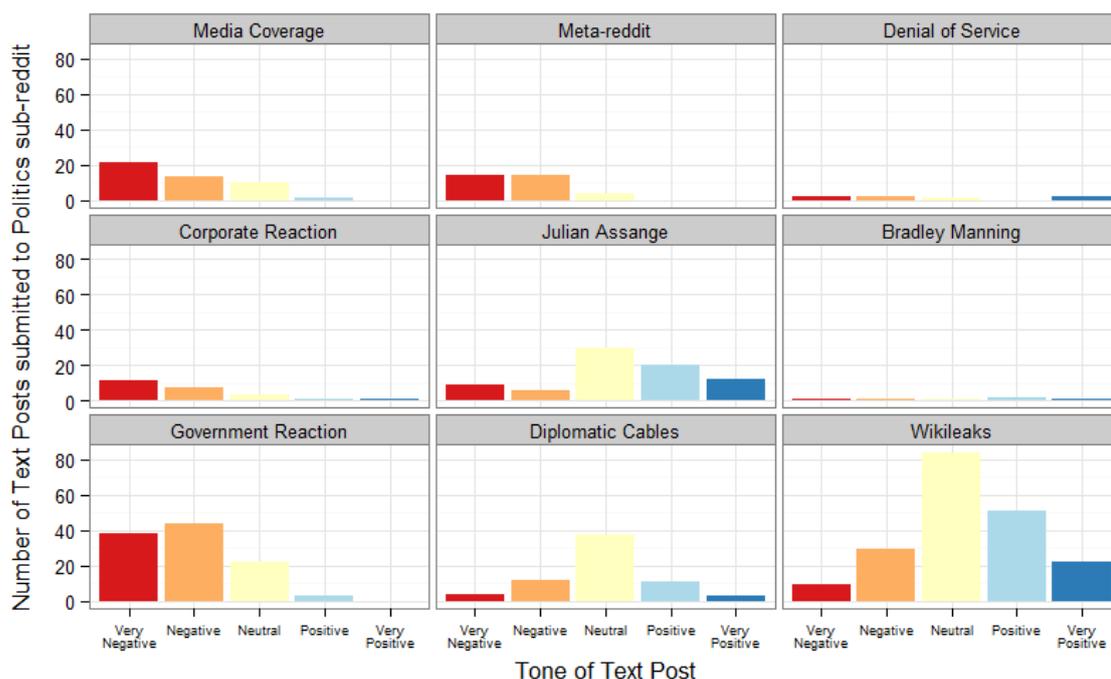
Figure 7.4: Showing the subject of text posts submitted to the 'politics' subreddit and the tone of these posts with regard to their subject.

Figure 7.4 suggests that text posts are indeed more likely to adopt a tone which is not neutral with regard to their subject, and that posts which are negative or critical of their subject are more common than those which are positive. In relation to `Wikileaks` and Julian Assange there is much greater variety in the tones espoused by these posts than on the Front page - suggesting greater diversity in the opinions of `reddit` users than the Front page would indicate. To return to the example of posts about the `Wikileaks` organisation; here 19% of posts are negative, 43% are neutral and 37% are positive. On the subjects of government/corporate reaction to the leaks and media coverage of the story, opinions seem to be more one-sided, with very few posts having a positive tone in relation to these subjects.

Posts which have been classified as 'meta-reddit' represent a particularly interesting type of content. These are posts which address coverage of the story *on `reddit`*, usually in a critical manner - the most common criticisms in this case being that there are too many prominent posts about the `Wikileaks` story, or that posts about the `Wikileaks` story are too one-sided and out of touch with the general population. These 'meta' posts which address some aspect of `reddit`'s collective behaviour are relatively common on the website. [1] Indeed it is quite common to see a number of posts on `reddit`'s Front page about a particular topic, alongside a post criticising the current popularity of said topic. For one example of this see redd.it/n0iaf, which appeared on the Front

---

[1] Over time this kind of post has become less common on `reddit`'s default Front page, by 2013 they have become rare.

263

page through /r/funny and referenced a trend on /r/atheism over the previous days whereby individuals would make posts offering to donate a certain amount to a charity for each up-vote they got (and would receive a lot of up-votes).

`Reddit` users actually devote quite a lot of time to discussions about the nature of the website; with many users citing the downsides of the `reddit` 'hive mind'. These users seem acutely aware of the possibility that the website's voting system has a particular effect on the discourse which takes place there, and caution against being swept along by the tide of popular opinion on the website. Whether these 'meta' posts actually affect user behaviour in the long run remains to be seen; but they are undoubtedly a fascinating aspect of the activity on `reddit`. The effectiveness of 'meta' posts as a form of self-regulation is considered further in section 7.4.
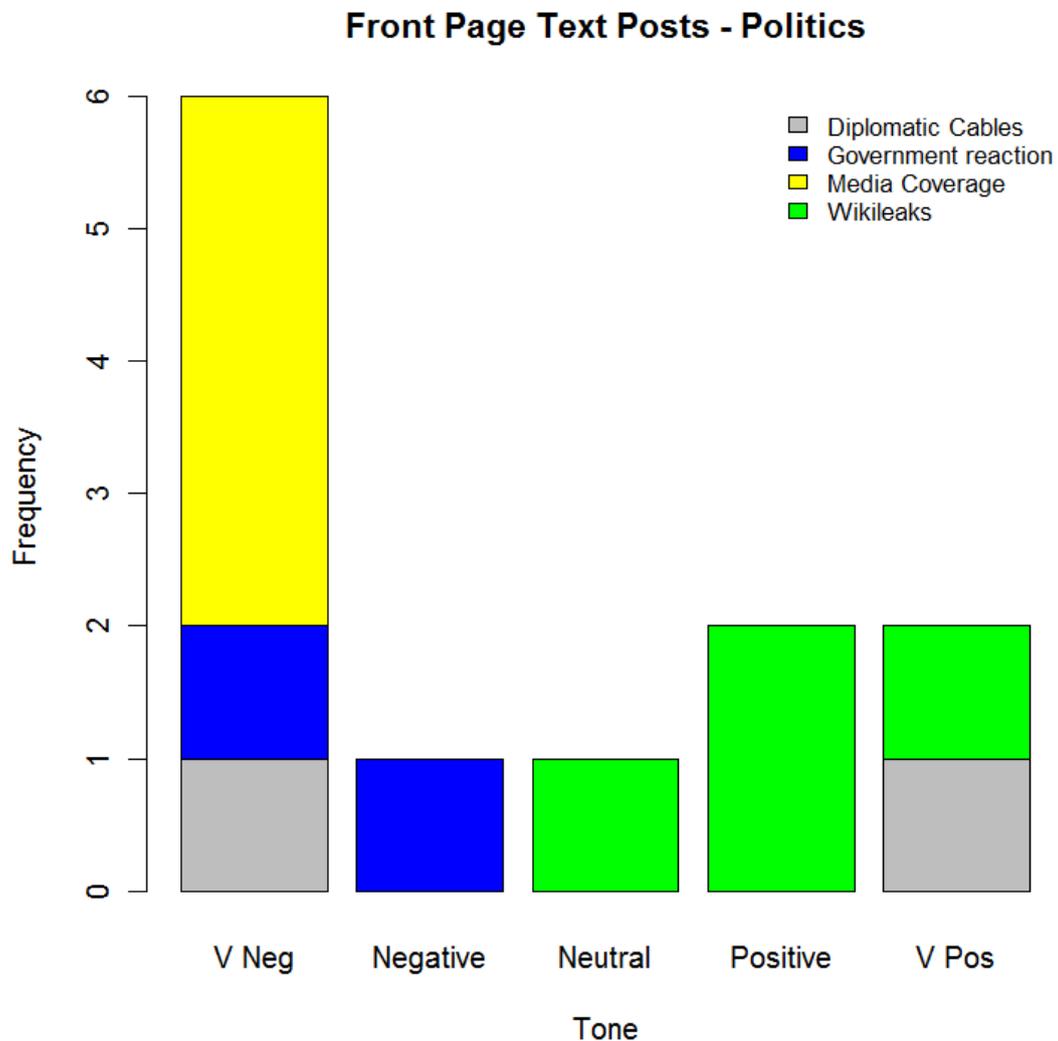
Figure 7.5: Showing the subjects and tones of the politics text posts which reached the Front page.

Figure 7.5 shows the subjects and tones of the twelve politics text posts which reached the Front page. Posts which are very critical of the mass media's coverage of the story seem to have been

received particularly well through the voting system. Most of these posts that reached the Front page fit with the general pro-Wikileaks tendency of reddit users; none of the text posts which espouse a negative opinion on Wikileaks or Julian Assange reached the Front page. There is one exception to this trend among these twelve posts, a post which was very critical of the release of the diplomatic cables. This was a long and well written post about the negative ramifications of the cables' release in terms of future diplomacy (redd.it/ed1ub). The presence of this post on the Front page suggests that discourse about the topic displayed on reddit's Front page was not completely one-sided. This post also highlights another trend in this data; text posts which are well written tend to receive more positive responses; poorly written posts are unlikely to attract up-votes regardless of the opinions they voice. It is likely that a measure of writing style or quality would prove useful in further analyses of text posts on reddit.

The final scores (converted to a four-level ordinal variable) of the 297 politics posts which have been classified are shown in figure 7.6. For this plot, tone has been collapsed to a 3-level ordinal variable and the 'direct' nominal category has also been retained. If reddit's users were deliberately excluding posts based on their subject and tone one would expect to see a much clearer trend in this plot. For example, if reddit was suppressing negative opinions about the Wikileaks organisation then we would expect most of the posts about Wikileaks to have received a negative score - but in this data posts which espoused a negative tone in relation to Wikileaks were just as likely to achieve a score of between 2 and 20 as a negative score, and one of these posts falls into the highest score band.

One must also bear in mind that the comments pages for popular reddit posts receive many comments - and within the ranks of high-scoring comments one will often see a range of opinions (and in turn counter-points to these in 2nd-level comments). A small sample of 189 comments from the Wikileaks data-set were coded according to their relationship with their parent post/comment - and of these 70 (33%) were classed as disagreeing with their parent, 13 expressed support (with argument/reference), 8 expressed agreement, 4 provided evidence counter to its parent and 3 sought to develop the ideas of their parent.

Reddit's voting system may marginalize minority opinions to some degree but in the case of Wikileaks coverage these do not seem to have been deliberately excluded from prominent locations. Reddit's voting system does not seem to preclude the possibility of this happening; it would therefore appear that reddit's users collectively do not wish to entirely exclude opinions they disagree with. Reddit's page of etiquette guidelines ('reddiquette') states that users should not 'Downvote opinions just because you disagree with them'. Perhaps reddit's users follow this guideline more than they give each other credit for - posts and comments which bemoan the reddit 'hive mind' are common, and often cite the exclusion of minority opinions as one of the community's failings.
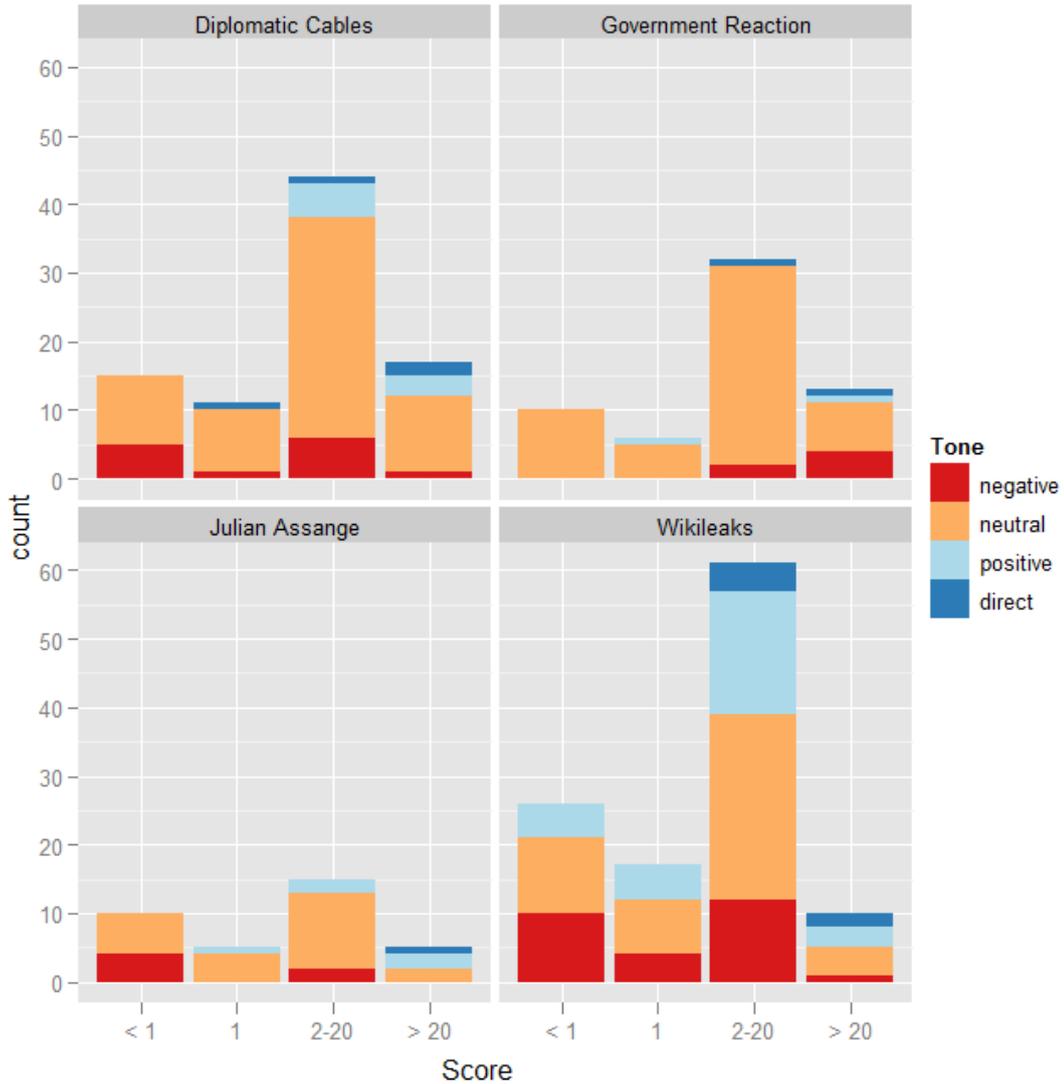
Figure 7.6: Showing the final scores of a sample of posts submitted to the politics subreddit - with each pane representing posts about a particular subject (only the four most common subjects are shown) and the bars being coloured to reflect the tone of these posts in relation to their subject

In relation to the `Wikileaks` story there were at least two popular posts where the author made reference to the 'hive mind' and explained why they disagreed with it (redd.it/eg7l4 and redd.it/emsym). The fact that this type of post can be up-voted to the Front page is further evidence against the dominance of a 'hive mind' on `reddit`; or at the very least an indication that the 'hive mind' values criticism which explicitly addresses it.

7.2   `Reddit` and the Stop Online Piracy Act

This section details a case study of `reddit`'s 'coverage' of the Stop Online Piracy Act (or SOPA); a piece of legislation which was being discussed by the United States House of Congress in late

2011 and early 2012. The Stop Online Piracy Act included several measures which, ultimately, proved distasteful to many Internet users and technology companies/organisations. Among its more controversial aspects SOPA would have partitioned the web into two categories, domestic and foreign websites; and would also have allowed for sites which are accused of facilitating copyright infringement to be censored in various ways (e.g. removed from DNS routing systems, removed from search engine results, cut off by payment providers). The legislation which came to be known as SOPA (having been formerly referred to as the E-PARASITE Act) had a similar counterpart in the United States Senate; called Protect Intellectual Property (or Protect IP or PIPA). A detailed analysis of this legislation is beyond the scope of this case study - which instead will focus on `reddit`'s 'coverage' of this story and consider the various 'Internet Campaigns' which occurred around these issues, and the role which `reddit` played in these campaigns (BBC, 2011).

The time period which this case study will consider runs from October 27th (when the first post about SOPA was observed on the `reddit` Front page) until February 3rd (a somewhat arbitrarily chosen date after SOPA itself was shelved and `reddit`'s focus began to shift away from Internet-regulating legislation). The focus here is on `reddit`'s default Front page, only considering those posts which have been observed on the first leaf (i.e. with a rank of 1-25) of the default Front page. For the most part the case study relies on observations of the default Front page which are recorded at 30-minute intervals. However, from January 3rd until January 18th the data-collection server was offline and this gap in Front page records has been filled through reference to a 3rd-party website which maintains a record of daily high-scoring posts on `reddit` (`www.redditarchive.com`).

The records of `reddit`'s Front page were first filtered to create a smaller data-set including only those posts which had a relevant key-word in their title. The key words selected are as follows:

- E-PARASITE (another name for SOPA used earlier on in the study period)

- SOPA

- Stop Online Piracy

- PIPA

- Protect IP

- GoDaddy (to include posts regarding a boycott of the GoDaddy domain name registration company - see section 7.2.4)

- Rob Zerban, Paul Ryan & Lamar Smith (to include posts regarding certain politicians who were referenced repeatedly in relation to this story - see section 7.2.5)

- Blackout (to include posts regarding a blackout of websites in protest of SOPA/PIPA - see section 7.2.6)

- ACTA (to include posts regarding the international Anti-Counterfeiting Trade Agreement - see section 7.2.7)

The initial set of keywords have been expanded to include posts regarding stories which are directly related to SOPA coverage on `reddit`. For example, GoDaddy (a large copmpany which provides domain name registration and website hosting services) has been included as a keyword because redditors were instrumental in organising a boycott of this company over its support of SOPA - there were subsequently many posts about the GoDaddy boycott and not all of these made explicit reference to the reason behind the boycott (SOPA support). See section 7.2.4.

Similarily, several US politicians rose to prominence on `reddit` through their support of or opposition to SOPA and were then the subject of posts which did not reference SOPA explicitly (see section 7.2.5). Also, while coverage of SOPA and PIPA is the main focus of this study, towards the end of the study period `reddit`'s attention branched out to include a range of other 'Internet regulation/censorship' topics, hence the inclusion of ACTA as a key word - see section 7.2.7.

After the removal of false-positive matches there were 259 posts which matched one of the keywords that were observed at least once on `reddit`'s Front page during the study period. Figure 7.7 shows the number of these posts which appeared on the default Front page on each day of the observation period. This figure suggests that coverage of the story saw a surge in the second half of November 2011, followed by a lull in early December and then a high number of SOPA-related posts on the Front page each day for the remainder of the observation period.
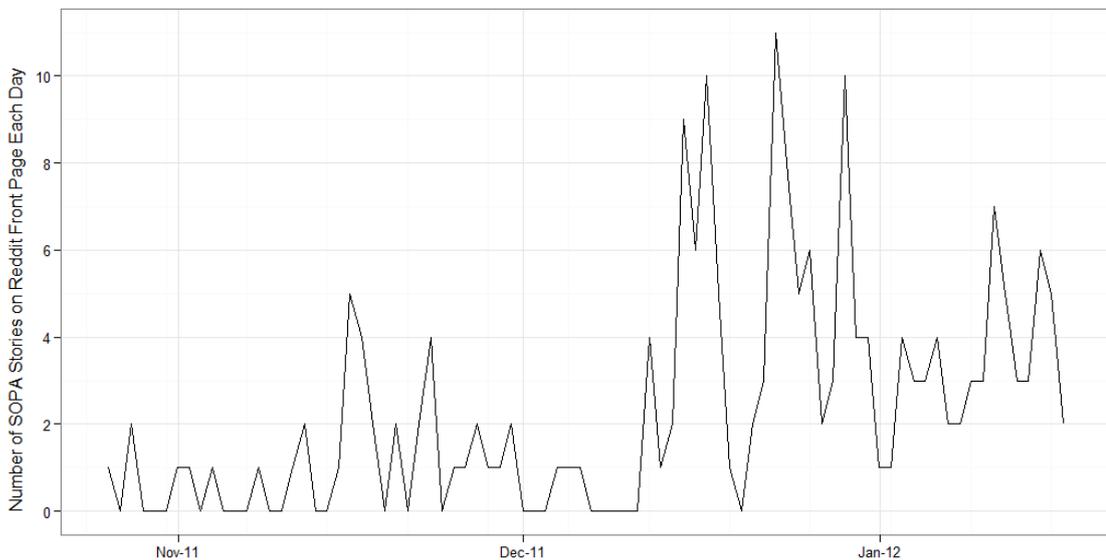
Figure 7.7: Showing the number of posts about this story appearing on `reddit`'s default Front page per day

The study proceeded by scrutinising each of these posts (and their comments page) in chronological

order - categorising these with regard to the nature, subject and tone of the post in relation to its subject, and also classifying the source which the post linked to. In describing `reddit`'s coverage of the story I will broadly follow the sequence of events as they unfolded while discussing posts of a particular type or following a particular theme within self-contained sub-sections.

7.2.1 Media coverage and public awareness of SOPA

It is important to first describe some key differences between this story and the previous '`Wikileaks`' case study. The `Wikileaks` story saw considerable coverage by the mainstream broadcast and print news media, and `reddit`'s level of coverage broadly parallelled that of the mainstream media (while providing a different slant on the story). As a result, the `Wikileaks` story was widely known to the public irrespective of `reddit`'s coverage.

In contrast, the US mainstream news media provided relatively little coverage of SOPA and this coverage came relatively late in the timeline we will consider. The first major US newspaper to formally address SOPA and voice its opposition was the *Los Angeles Times* on 25th November (Editorial, 2011a), quickly followed by the *New York Times* on 26th November (Editorial, 2011b). Before this the only reference to SOPA or PIPA in newspapers came from occasional mentions by columnists (Masnick, 2011a).

There is also a disparity here between print and broadcast news media, with newspapers and magazines providing much more coverage of this story than television news networks. There is a conspicuous lack of coverage of this story on the major US TV news networks (Dimiero, 2012a,b); with many of these networks being owned by large media companies who supported SOPA. In the case of the `Wikileaks` story `reddit` served to provide an alternative narrative to the mainstream media (with an emphasis on getting involved and helping `Wikileaks`). In the case of SOPA `reddit` had an additional role to fill, that of raising awareness of the fact that this legislation existed.

Figures 7.7, 7.8 and 7.9 suggest that `reddit`'s coverage of this story begins well before the story becomes the subject of widespread public awareness/interest. Figure 7.8 shows that searches for SOPA-related terms (top pane) and references to these terms in the news media (bottom pane) were rare in November, more common in December, and saw a large increase in January with a spike on January 18th when many websites (`reddit` included) 'blacked out' in protest of SOPA (see section 7.2.6). Mentions of these terms on `Twitter` follow a similar pattern, with small flurries of activity in mid-November and December followed by a spike at around the 18th of January. This suggests that the blackout was an effective means of raising awareness of SOPA - with `Wikipedia` choosing to black out and Google placing links to SOPA information on its front page probably having a much larger impact than `reddit` itself blacking out. We will re-visit the blackout specifically in section 7.2.6 below.
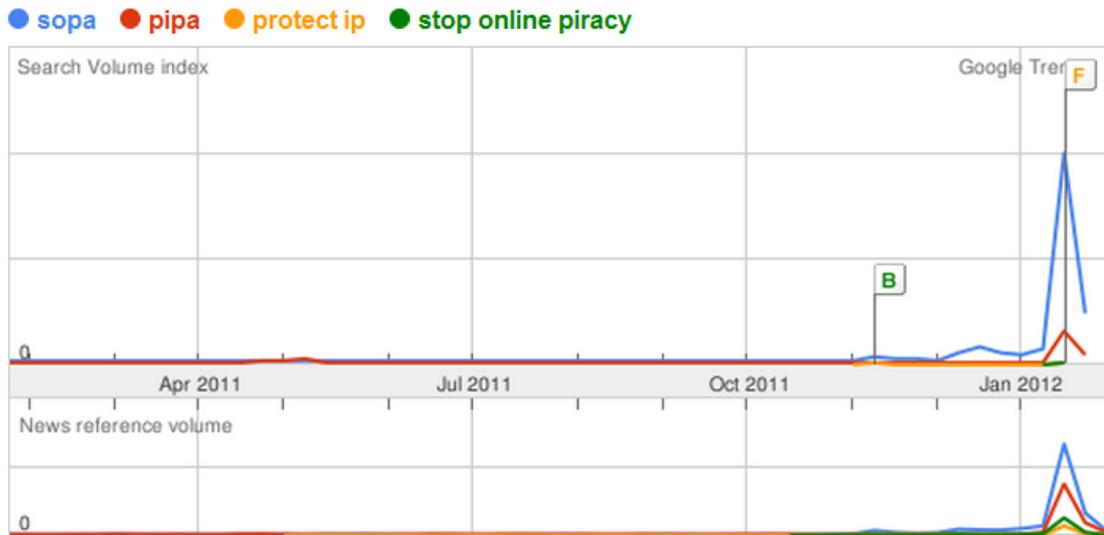
Figure 7.8: Prevalence of key words on google, retrieved from 'Google Trends'
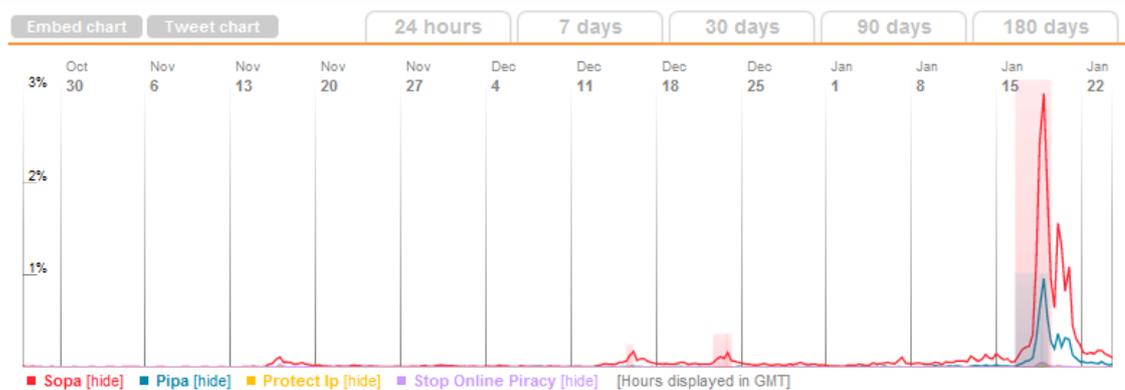
Figure 7.9: Prevalence of key words on `Twitter` - percentage of tweets which included them

How important was `reddit` in raising awareness of SOPA?

While `reddit`'s role in raising awareness of SOPA is of particular interest it is very difficult to address this question directly. The above graphs suggest that SOPA was a common topic on `reddit` much earlier than it was a common Google search or `Twitter` hashtag - it follows that SOPA was widely known among the `reddit` community before it was widely known among the general public.

One cannot however state with confidence that `reddit` was a major factor in raising public awareness of SOPA. There were undoubtedly many `reddit` users who first heard about SOPA through `reddit`. These individuals may then have spread that information further by communicating with their own networks of friends through Social Media, face to face interaction, email, or on their own website or blog. It would be very difficult to track the spread of information in this fashion, and even if relevant data could be obtained it is unlikely that users would always reference `reddit` as

their source or the place where they learned about SOPA. Even in a situation where an individual forwards a link to SOPA information which they found through `reddit` they are perhaps more likely to link directly to the article or piece which was linked to by the `reddit` post (and therefore their communication would make no reference to `reddit`).

In place of this kind of analysis I offer some numbers that give an idea of the scale of activity on `reddit` in this time period. A post on the `reddit` blog (Reddit-Blog, 2012) in January 2012 provides some traffic statistics for December 2011. `Reddit` had around 35 million unique visitors (by IP address) in this month and served 2 billion page views. The problems of equating 'unique visitors' to individual people have been discussed in previous chapters. In this instance it is perhaps better to look at 'visits', a measure of instances where an individual has browsed `reddit` for a single 'session'. A figure of 159 million 'visits' is quoted for December 2011 - with mean visit duration of 16 minutes in which an average of 13 pages were viewed. If these visits were spread evenly throughout the month this would equate to around 5.1 million visits each day. One can assume with some confidence that most of these users will have viewed the `reddit` Front page, but for users who were signed into an account this Front page would be populated by posts from the subreddits they subscribe to.

One can also consider recorded activity levels for the 259 Front page posts which matched key terms for this case study. These posts received a combined total of 161,222 comments - an average of 622.5 comments per Front page post. These posts also had a combined score of 493,537 upvotes less downvotes.

It is not possible to determine how many `reddit` users actually clicked through to read posts and articles about SOPA which appeared on the Front page - many users who saw the default Front page may have bypassed SOPA-related posts in favour of the more humerous or entertaining posts which dominate `reddit`'s Front page. A post on the `Techdirt` blog (a blog linked to by 16 of the Front page SOPA posts on `reddit`) may be helpful in this regard as it mentions `reddit` in its traffic review of 2011 (Yang, 2012). For the year of 2011 `reddit` 'referred' 2.1 million users to the `Techdirt` blog (when someone follows a link from one website to another this is recorded as a referral) - a 277% increase as compared to the previous year. Eight of the SOPA-related posts which appeared on `reddit`'s Front page and linked to `Techdirt` appeared in 2011, and three of these feature in the top 10 `Techdirt` articles by traffic for 2011.

This highlights a second characteristic of `reddit` with regard to raising awareness of SOPA. `Reddit` was definitely not the first website where SOPA was brought to light and discussed (i.e. the first post about SOPA on `reddit`'s Front page links to the `Techdirt` blog). There were likely many other technology-oriented blogs and forums, and some activist organisations (e.g. the Electronic Frontier Foundation), which were discussing SOPA before it appeared on `reddit`'s Front page. Within the social component of the Hybrid Media System (Chadwick, 2013), `reddit` likely performed the role

of spreading this awareness (which had been confined to smaller domain-specific communities) to a larger audience. `Reddit`'s users are thought to be quite technologically savvy, indeed many of the early posts about SOPA appeared on the Front page through the 'technology' subreddit. In time awareness of SOPA within the `reddit` community spread from users who are explicitly interested in technology to the `reddit` user-base more generally (see section 7.2.8) - and these individuals may have then spread awareness beyond `reddit` to their own circles of friends/contacts. This is one method by which `reddit` may have 'raised awareness' of SOPA - and it helps us to situate reddit in the hybrid media system as an 'amplifier' which can direct a considerable volume of human attention to items produced by older media elites and newer media upstarts alike - and which simultaneously facilitates a large-scale 'democratically mediated' discussion around these artefacts.

The second method by which `reddit` may have raised awareness of SOPA involves other media outlets reporting on events unfolding through `reddit`. For example, the boycott of GoDaddy (see section 7.2.4) which originated on `reddit` was reported on by a number of recognisable news outlets which have a much larger audience than `reddit`, e.g. BBC (2011). Similarly, 'the blackout' (see section 7.2.6) did a lot to raise awareness of SOPA and `reddit` was involved in this from an early stage.

Without comprehensive study of the hybrid media system (Chadwick, 2013) more broadly we cannot determine whether `reddit` was truly instrumental in 'breaking' this story - but it is safe to assume that `reddit` had introduced the legislation to many users through a number of posts on its Front page before this was 'a story' for the conventional news media. It is certainly a strong indication that `reddit` does not always have its agenda set for it by older media elites.

7.2.2 Early Posts about SOPA - October 27th to November 25th

This section considers the earliest SOPA-related posts to appear on `reddit`'s Front page, up until November 25th when the *Los Angeles Times* published an editorial in opposition to SOPA. In this four week period there were 30 posts about SOPA which appeared on `reddit`'s Front page.

`Reddit`'s first five Front page posts about SOPA

The first post to appear on `reddit`'s Front page (redd.it/lq2b1) did so on October 27th and linked to the '`techdirt.com`' technology-oriented blog. The blog post which was linked to referred to the legislation by its full title of 'Enforcing and Protecting American Rights Against Sites Intent on Theft and Exploitation Act' but noted that it may also be referred to by the shorter moniker of the 'Stopping Online Piracy Act'. This blog post described SOPA as an attempt to create a 'great firewall of America', and noted opposition to the legislation from the tech industry, legal

experts, investors and entrepreneurs. The blog post being linked to also provided the full text of the legislation.

This post was submitted to the /r/technology subreddit; it was observed at rank 9 on the Front page and had around 400 comments when it ceased appearing on the Front page. By the standard of posts which would appear later in the study period this one was moderately successful. We can think of this post as serving to inform `reddit` users about SOPA's existence. Later in the study period posts would shift towards trying to spread awareness/opposition to SOPA *beyond* `reddit`, but before this could happen it was first necessary for `reddit` to inform its own users.

This post also appears to have set the tone for `reddit`'s coverage; of the 259 posts considered here 91 were submitted to the technology subreddit (more than any other subreddit) and 16 linked to the `Techdirt` blog (more than any other individual content-providing website by a considerable margin). Popular comments on this post also echo sentiments which would become common in relation to posts about SOPA - voicing opinions about the legislation being frightening, opening the door to Internet censorship and the result of corporate lobbying/donations.

Two more posts regarding SOPA appeared on `reddit`'s Front page the following day, one of these was a post submitted to the videos subreddit which linked to an amateur `Youtube` video on the subject. The second post (redd.it/lrux7) linked to another article about SOPA but its title read "For those upset at the E-PARASITE (aka Blacklist Bill aka PROTECT-IP), Ron Wyden (D-OR) is holding up the bill. Let's show him we support his action. Write him, donate, do what you can. his website is http://www.standtallforamerica.com/". This post was indicative of one of the themes of `reddit`'s coverage which would emerge later in the observation period - namely the identification of pro and anti-SOPA politicians and subsequent criticism or praise of these individuals (see section 7.2.5).

On 1st November the fourth post about this story appeared on `reddit`'s Front page (redd.it/lvr2x), this post being a link to a `WhiteHouse.gov` petition which was created the day before (probably by the same user who submitted the post to `reddit`). This post was submitted to the technology subreddit and appeared on the default `reddit` Front page for 13 hours, reaching a maximum position of 3rd on the Front page. The petition being linked to received around 52,000 signatures, but the popular `reddit` comments on this post are skeptical about the value of a `WhiteHouse.gov` petition or critical of the language used in this particular petition. There are also several comments which urge `reddit` users to contact their representatives directly as opposed to signing the petition. This post is the first which appeared on `reddit`'s Front page to call for some form of action from `reddit` users on this subject; a type of post which would become common later in the observation period (discussed in section 7.2.6).

On 2nd November the fifth post about SOPA appeared on `reddit`'s Front page (redd.it/lvr2x);

this post concerned the submitting user's correspondence with their Member of Congress on the subject. The 'original poster' (OP to use `reddit`'s terminology) created an image of the reply they received from their Member of Congress. On this post's comments page the top comment is from a user who claims to be a 'former Capitol Hill staffer', this user provides detailed advice on how to pursue the matter further and more effectively. There are several further comments from other Capitol Hill staffers, and also a number of comments in which other users post the text of the correspondence they received from their representatives.

After reviewing just the first five posts about SOPA on `reddit`'s Front page we can already see some trends which are likely implicated in `reddit`'s effectiveness as a platform for mobilising collective action on the matter. First there are posts providing information about the legislation (information which comes from 'amateur' sources and not the conventional media) with detailed discussions on their comments pages. Then there is a post which calls for action (signing a petition) - this post is seen by enough people to garner more than 50k signatures (the threshold required to elicit a response from the White House). On this post's comments page there is further discussion, this time around the topic of how `reddit` users might productively work against the legislation (with many comments suggesting contact with representatives). The next day there is a post from a user who has taken the advised action and contacted their representative - the fact that this post received enough up-votes to place it on the Front page can be read as positive reinforcement for this course of action.

This point in particular marks a departure from how the story could or would have developed through the mass media. Suppose (1) a mainstream news outlet covered the story and (2) adopted a negative perspective on the legislation and (3) went on to urge their readers/viewers to take action. At this point the reader/viewer is on their own, whether they take the course of action advocated by the news organisation is up to them, they have no way to confer with other readers/viewers to find out if others are also taking this course of action. On `reddit` this post about an individual's efforts at taking action *is part of `reddit`'s coverage of the story*, it received the same level of recognition and was 'broadcast' in the same way as earlier posts about the topic. Furthermore, these users were able to hold a productive discussion of the action they were taking on the post's comments page, with many other users also reporting similar interchanges with their representatives. These factors likely served to increase `reddit` users' willingness to become actively involved in the story. Displayed on the Front page we have evidence of the recognition and support received from the community by one ordinary `reddit` user because they took the time to write an email to their representative.

The conventional news media do occasionally initiate campaigns for various causes - but in these cases the campaign is lead by the organisation's journalists or editorial staff, with readers/viewers' role being confined to taking whatever action is requested by the organisation (usually signing a

petition). Recently there have been cases where the news media have integrated these campaigns with Social Media (Vaccari, 2012) - thus circumventing the problem whereby readers are isolated from each other and do not know whether other readers are participating. However, the role of readers in these campaigns is still confined to one of 'backing up' the newspaper's position. The situation is one where the news organisation can mobilise its readers to show support for its position and Social Media is employed to enhance this ability. On `reddit` campaigns of collective action have *been initiated and lead* by ordinary users, this cannot occur for a newspaper's campaign because readers have no capacity to have their views published within the newspaper (with the possible exception of letters to the editor, but these are only publised at the editor's discretion).

Finally, we see one of `reddit`'s strengths come to the fore with this post, the strength of numbers. While there are millions of users browsing the `reddit` Front page on any given day there is a strong chance that some of these users hold some expertise on whatever stories are featured there. In this case former Capitol Hill staffers saw the post and contributed to the discussion with expertise on how users could increase the effectiveness of their efforts. Through the use of the voting system these valuable comments were promoted to the top of the comments page where they would be seen by many people - in the process providing some measure of recognition and encouragement for the expert user who provided them. This is a common occurance on `reddit`, there are frequently comments on a post which demonstrate a high degree of knowledge of the subject. The implementation of up/down voting for comments allows these high-value comments to be displayed prominently, without up/down voting for comments they would effectively be lost among the deluge of other comments. There are subreddits which rely on these characteristics of `reddit` to function. For example /r/askscience works because there are enough scientists from various disciplines reading the subreddit such that there is usually at least one user who can provide a good answer (or as good an answer as the scientific literature on the topic allows for) to questions which are often difficult to answer. The voting system works for this purpose as long as the other users reading the comments have enough familiarity with the subject that they can recognise high quality comments when they see them and up-vote accordingly.

In these early posts and their comments one can witness `reddit` developing its collective position on this emerging story, 'figuring out' which logics and tactics it will deploy to which ends, and even 'scouting' external entities which would later become an integral part of the *ad hoc* Hybrid Media assemblage (Chadwick, 2013) that evolved around `reddit`. Later in the observation period the nature of posts and their comments appearing on the Front page shifts such that questions about the legislation or `reddit`'s coverage become more involved and reflect an emerging general consensus on the key points.

Reddit's Front page coverage from 4th November until 25th November

Returning to the analysis of posts which appeared on reddit's Front page until November 25th - the remaining 25 posts largely fall along one of several themes. There are eight further posts with information and criticism of SOPA from a variety of sources. Some of these posts link to online news providers (e.g. redd.it/lrux7), some link to blogs (e.g. redd.it/mipae), others link to the websites of activist organisations (e.g. the Electronic Frontier Foundation, redd.it/m4byf and redd.it/m9r1p).

On 4th November there is a post which links to an article about Google's opposition to SOPA (redd.it/m0bii). Between November 15th-17th there are a further four posts which concern opposition to SOPA from various technology companies (mcnjm, me21k, meg8w, mer7w). These include a post on November 16th in which one of reddit's admins states that they are hoping to testify before congress in opposition to SOPA, and asks the reddit community for ideas on what they should say. Identifying corporations who oppose SOPA (and later some corporations who supported SOPA) became a common theme throughout reddit's coverage of this story, and these posts will be considered in greater detail in section 7.2.4.

There are also posts which concern political support for and opposition to the legislation. The first of these posts appears on November 16th (reddi.t/mdqnf) and highlights Ron Paul's opposition to the bill particularly; a similar post two days later also mentions Ron Paul specifically (redd.it/mgg28). Previous observations have suggested that there is a contingent of libertarian Ron Paul supporters on reddit, and other users have made reference to a bias in favour of Ron Paul on reddit. The third post to identify politicians opposing SOPA appeared on the Front page on November 22nd (redd.it/ml0ii) and concerned Senator Ron Wyden. This post can also be thought of as a call to act, it linked to a website created by Ron Wyden's camp where he spelled out his opposition to SOPA and voiced a plan to filibuster the bill should it reach the house floor. The website contained a petition of sorts, users could sign this petition and Ron Wyden would read aloud the list of signitories during his filibuster.

On November 23rd and 24th there are two posts which identify politicians who support SOPA. The first of these concerns Senator Al Franken (redd.it/mmsg5). The post's title mistakenly identifies him as a co-sponsor of SOPA, several top comments point out that he is actually a sponsor of the Senate equivalent bill PIPA or Protect IP. This post is interesting because reddit had previously held some favour towards Al Franken (redd.it/aiu2i, redd.it/c1fev, redd.it/cmuj8) and Franken had participated in an 'Ask Me Anything' interview on reddit (redd.it/g0hir). The posts about Franken's sponsorship of PIPA and the comments thereon indicate a sense of betrayal among reddit's users. The second post to identify a pro-SOPA politician singles out Lamar Smith (a US congressman for a district in Texas) as the 'founder' of SOPA. Section 7.2.5 below considers this

theme of `reddit` posts which identify political supporters and opponents of SOPA more thoroughly.

The remaining posts related to this story which appeared on `reddit`'s Front page before November 25th concern taking action against the legislation. On November 17th a follow-up post concerning the earlier Whitehouse petition post (redd.it/lvr2x) appears on the Front page (redd.it/mf6qm), in this post the OP thanks `reddit` for generating 30,000 signatures for their petition. The next day another post appears promoting another Whitehouse petition (redd.it/mgw7f). This time the top scoring comments are highly critical of the petition, criticising the author's grammar and also the fact that he has asked for things which are not within the Whitehouse's power to grant. This petition failed to reach the threshold of 50k signatures, but did appear on `reddit`'s Front page for a number of hours. This post serves as an example of a dissonance between post and comment voting/scores on `reddit` which is often observed (see section 5.1.10). The post achieved a score of 2,590 but all of the top comments are highly critical of the petition. If the same population of users that commented and voted on comments were voting on the post, one would expect that it would have been quickly down-voted off the Front page or would never have appeared in such a prominent location in the first place - but as section 5.1.10 showed `reddit` does not always have the capacity to remove 'poor' posts from the Front page once they have been shown to be 'poor'.

A further example of this dissonance can be seen in a post (redd.it/mfwkl) which appeared on the Front page on November 17th. In this post a `reddit` user has taken a screenshot of the 'Access Denied' message they were presented with when they tried to view `Wikipedia`'s article on SOPA from their school and posted this with the title "And so it begins, my school has conveniently blocked `Wikipedia`'s article about SOPA... this is just ridiculous". On the comments page many high-scoring comments voice skepticism about this being part of a government conspiracy and suggest that it may have been automatically blocked because the URL contained the word 'piracy'. It appears that this post was quickly down-voted from the Front page after its appearance there.

The remaining posts to appear on the `reddit` Front page during this period voice support for collective action against SOPA. On 20th November a post (redd.it/mizfo) links to an article praising the `Tumblr` community for offering a service which would 'train' its users to speak to their representative about SOPA and then connected them on such a call. This could be thought of as `reddit` paying attention to and promoting the efforts of another social media elite 'actor' - with later developments indicating that `reddit`'s users may have learned from this observation of a Social Media 'neighbour'. On November 17th there is a post citing international support for the anti-SOPA activism of Americans (redd.it/mfuiy).

The final post in this time period which has not yet been considered is an unusual one. On November 24th this post (redd.it/mn9pv) linked to a *Forbes* article with the title "Wary Of SOPA, Reddit Users Aim To Build A New, Censorship-Free Internet". This article concerned a subreddit (/r/darknetplan) which was established to facilitate work on a decentralised network of peers.

While this plan is geared towards producing a network which would be resistant to censorship, it pre-dates SOPA by some time and the direct link to SOPA seems to have been created by the article's author. This post does however serve as an example of a type of post one often sees on `reddit`. When a recognisable media outlet produces an article or other piece about `reddit` these tend to be submitted on `reddit` and to do reasonably well through the voting system. It seems that `reddit` users like to make press articles about their community visible, even when these articles are based on flawed or erroneous assumptions. [2]

To conclude this section we will consider `reddit`'s reaction when some traditional news organisations (The *LA* and *New York Times*) began to cover SOPA and PIPA. As stated earlier the *LA* and *New York Times* published editorials opposing the legislation on November 25th and 26th (Editorial (2011a), Editorial (2011b)). Perhaps surprisingly, there were no posts linking to these articles which appeared on `reddit`'s Front page. This is a departure from what was observed during the `Wikileaks` case study - where articles published by well-known news outlets were among the most highly up-voted. A post linking to the *New York Times* article was submitted on November 27th (redd.it/mqtly) but it received just a single down-vote. Perhaps the post's title ("Going after the pirates") was to blame for its failure, that title certainly does not capture the essence of the article. `Reddit` does not allow the same URL to be submitted more than once to a subreddit and this poorly titled post may have blocked subsequent attempts to submit a direct link to the editorial. The fact that the *New York* and *LA Times* published anti-SOPA editorials *was* the subject of a post which linked to the `Techdirt` blog and did appear on the Front page on November 28th (redd.it/mrzie). At this stage this was the 3rd post linking to `Techdirt.com` about SOPA which appeared on the Front page, including the first post about SOPA to appear on `reddit`'s Front page.

It is possible that by the time large recognisable news outlets like the *New York Times* started to address SOPA, `reddit` had already settled on a way of covering the story (with links to blog posts, activist organisations, articles by online-only publications, and plenty of ideas and reports directly from `reddit` users). The hybrid media assemblage which `reddit` was central to was already in evidence and may have already begun to solidify at this stage.

It is worth noting that 22 of the 30 posts concerning SOPA which appeared on `reddit`'s Front page up until 25th November were submitted to the /r/technology subreddit.

---

[2]This is another example of a trend which has become less prominent on `reddit` over time. There are two possible causes for this - 1) the number of articles written about `reddit` and published by recognisable news organisations has increased substantially, making it impractical to feature all of these on the Front page and reducing the 'novelty value' of articles written about `reddit`. 2) As `reddit`'s user-base has grown it has become more fragmented (see Chapter 8), raising the possibility that many users no longer see news *about* Reddit as relevant to them.

7.2.3 Information about SOPA

Throughout the observation period `reddit`'s Front page displayed many posts with information on SOPA and PIPA, too many to describe each one individually. These posts all share a negative outlook on SOPA and usually offer a critique of the legislation. The sources of these posts are varied, ranging from a single sentence summary written by a `reddit` user (redd.it/nbepe) to a 23-page critique written by a Harvard law professor (redd.it/na3z8), with many links to blog posts and some articles from the conventional news media.

`Reddit`'s coverage of a rapidly developing story

This case study also allows us to look at `reddit`'s coverage of a story which changes rapidly, and the publication of erroneous information - a potential misinformation 'cyber-cascade' (Sunstein, 2002). On December 16th the Congress House Judiciary committee held a meeting to review SOPA, discuss amendments and vote on whether the bill would be passed to Congress proper for a vote. This was to be the last day on which congress worked before breaking for holidays, and therefore when the committee decided to postpone their vote until the 'next available date' it was assumed that this meant a postponement until 2012. Two posts appeared on `reddit`'s Front page announcing this news. One was submitted to the technology subreddit (redd.it/nfk28) and the other to the politics subreddit (redd.it/nfj28) and they were first observed on the Front page at 2130 and 2200 GMT respectively. This news was greeted positively by `reddit`, with one of the posts even prefixing its title with 'VICTORY!'.

However, by 23:00 there were comments on both of these posts stating that the vote had not in fact been postponed until 2012 but that a special session had been arranged for the following Wednesday. In the context of `reddit` at the time, where there was much talk about collective action against the bill, this was an important update. A situation where `reddit`'s Front page is telling people they have until 2012 to campaign against the bill when it is potentially moving forward in less than a week would be damaging to `reddit`'s capacity to work against the bill. Within three hours of this update emerging there was a post on `reddit`'s Front page through the technology subreddit stating 'SOPA has NOT been postponed to 2012!' and linking to a `Techdirt` article with details of what the current state of affairs was (redd.it/nfrru). Within the next six hours posts with the same title and linking to the same `Techdirt` article also appeared on `reddit`'s default Front page through the politics (redd.it/nfxhj) and WTF (redd.it/ng1wp) subreddits.

Taking `reddit`'s Distributed Moderation system into account (i.e. to appear on the Front page a post must acquire enough upvotes to outscore its rivals) this represents quite a timely response to quickly changing information about the story. Word first spread through the comments pages of posts which were already on the Front page. A text post submitted to the /r/sopa subreddit

(redd.it/nff7k) had its body of text edited to reflect the new information, and in fact directed readers to the post with updated information on the politics subreddit.

It is `reddit`'s community of users who were responsible for using the voting system to disseminate updated information quickly. `Reddit`'s software actually hindered this process somewhat; it is not possible to edit the title of any post on `reddit`, and for posts which link to a URL it is not possible to change this URL. The users who made posts about a postponement until 2012 were presumably made aware of the change in the situation quite quickly if they were reading the comments on their posts, but were powerless to change the post itself to reflect the new information. Users had to work collectively to downvote the misleading posts and upvote the newer posts with accurate information.

This episode speaks to the concern voiced by Sunstein (2002) and others about 'cyber-cascades' of mis-information - wherein false information is released and spreads rapidly but news of its falsehood does not achieve the same penetration in a timely manner. Information which is false/flawed/outdated does sometimes appear on `reddit`'s Front page, but this case shows that users can collectively 'shut down' the dissemination of an item once it has been shown to be wrong and 'print a retraction' by Front paging an item which corrects the mistake, and all in a timely manner.

7.2.4 Corporate Allies and Enemies

One of the themes of `reddit`'s coverage of SOPA is the identification of corporations which opposed or supported SOPA. Between the start of the study period and December 5th there were eight posts featured on `reddit`'s Front page whose main point was to identify corporations who had voiced opposition to SOPA. SOPA opponents identified in this way include Google (redd.it/m0bii), `Facebook` (redd.it/me21k), Microsoft (redd.it/mmokt) and `reddit` itself (redd.it/meg8w). During this time there is also one post appearing on the Front page which links to an article stating that NBC had threatened their partners into signing a SOPA 'grassroots support' document (redd.it/mtmib).

On December 13th the tone of these posts shifted from merely identifying SOPA opponents and supporters to discussions of action which could be taken with or against these corporations. The main thrust of these discussions is that SOPA opponents should be encouraged to 'black out' their website to raise awareness of SOPA (e.g. redd.it/ng2cz), while SOPA supporters should be threatened with boycotts in an effort to change their stance. This change in focus from identification to action coincides with a surge in the number of SOPA stories appearing on `reddit`'s Front page.

The GoDaddy boycott

The first post concerning a boycott to appear on `reddit`'s Front page did so on December 21st (redd.it/nl9e0). This was a text post made to the AskReddit subreddit, initially this post contained a link to a list of companies in the Business Software Alliance, a group which supported SOPA. This post was the centre of a large-scale discussion incorporating more than 2,500 comments. In the course of this discussion it was established that many of the companies in the Business Software Alliance actually did not support SOPA, and the Original Poster edited the text of their post nine times to first correct this error and then add links to other more accurate lists of SOPA supporters.

The next day a post appeared on the Front page in which a `reddit` user stated that they were transferring their own small business's domain names away from GoDaddy over the company's support of SOPA (redd.it/nmnie). They suggested that other `reddit` users do likewise, nominating December 29th as 'move your domain away from GoDaddy day'. GoDaddy is one of the largest and most well known domain name registration and web hosting companies. People use GoDaddy to buy (or 'register') domain names (web addresses) and there is a yearly fee for retaining ownership of a domain. This post was asking users to transfer their domain names to an alternative domain name registrar; as a result GoDaddy would lose the yearly income from these domain names and this business would go to one of their competitors.

Once this post appeared on the `reddit` Front page it received considerable attention and ultimately more than 3,000 comments. Many popular comments were from other `reddit` users stating that they would be moving their own domain names away from GoDaddy. Some of these users claimed to own or administer hundreds or thousands of domains. There are also comments from users who claim to be former GoDaddy employees, one advises `reddit` users on how to identify domain name registrars which are owned by GoDaddy while another talks about suspect business practices at GoDaddy.

As commonly occurs with text posts, this post's OP edited the text of their post with updates. In this case some of GoDaddy's competitors offered special discount codes for users who wished to transfer their domain away from GoDaddy and these codes were edited into the text of the post. The OP also edited in links to articles which were written about this developing story (Masnick, 2011b).

Over the next two days (December 23rd and 24th) there were no fewer than 13 posts on `reddit`'s Front page about the GoDaddy boycott. There follows a brief description of posts about the GoDaddy boycott in the order in which they were observed on the Front page.

- redd.it/nmsiu - a post to the technology subreddit which links directly to the post which sparked the boycott (redd.it/nmnie) - the earlier post was submitted to /r/politics, this post serves to bring it to the attention of /r/technology readers.

- redd.it/nn4j5 - this post links to an article published by the `Ars Technica` website about the boycott and its origins on `reddit`.

- redd.it/nnc8r - an article about the `Cheezburger network` moving its 1,000 domains from GoDaddy.

- redd.it/nntxp - titled "GoDaddy's Response to the Boycott:"Go Daddy has received some emails that appear to stem from the boycott prompt, but we have not seen any impact to our business." Reddit, Lets make them feel the impact and move your domains! Spread the word!" This post concerns GoDaddy's initial response to the boycott, a response which seems to have only increased the enthusiasm of `reddit` users for the boycott, see also Masnick (2011b).

- redd.it/nnv9l and redd.it/nnuwt - these posts link to the WHOIS page for `Wikipedia`, showing that GoDaddy is their domain name registrar, with the implication that `Wikipedia` should leave GoDaddy.

- redd.it/nnynm - the WHOIS page for `Imgur`, `reddit`'s favoured image hosting website, again showing that GoDaddy is their domain name registrar with the idea that they should transfer their domains away from GoDaddy.

- redd.it/no6pr - an article on `Ars Technica` proclaiming "Victory! Boycott forces GoDaddy to drop its support for SOPA", popular comments all state that their boycott should/will still go ahead.

- redd.it/nokde - a text post urging users to ignore GoDaddy's 'flip-flop' and proceed with the boycott.

- redd.it/no7rr and redd.it/np5cp - posts linking to the `Twitter` stream of Jimmy Wales (`Wikipedia` co-founder) stating that `Wikipedia` would transfer their domains from GoDaddy.

- redd.it/no1ux - a post asking for advice on how to transfer domains from GoDaddy and who to transfer them to.

- redd.it/nohs8 - "GoDaddy mass migration day STILL ON for December 29th - do NOT stop now!"

- redd.it/npair - "GoDaddy has NOT withdrawn its official congressional support for SOPA", the text of this post contains an interview statement from the GoDaddy CEO saying that they had not withdrawn their official support for SOPA.

Over the next five days (December 25th - December 30th) there were a further 11 posts about the GoDaddy boycott on `reddit`'s Front page.

- redd.it/npj2q - an article about GoDaddy losing customers (21,000 domains the previous

day) over SOPA support.

- redd.it/npmav - another article about the effectiveness of the boycott, this time quoting a loss of 72,000 domains over the previous week.

- redd.it/nq4g2 - a post about Google using GoDaddy as their domain name registrar partner (with the idea being that Google change provider)

- redd.it/nrgij and redd.it/nr7xc - two posts linking to an article which says that GoDaddy were involved in drafting the SOPA bill and are also granted an exemption should it be implemted.

- redd.it/nrptu - a GoDaddy employee does an Ask Me Anything interview on `reddit`.

- redd.it/nrd49 - a post about GoDaddy obstructing people from transferring their domains away.

- redd.it/ntlzq - `Imgur` to join GoDaddy boycott.

- redd.it/nujea - a post about the GoDaddy CEO, stating that they started out selling bible software and contrasting this with the raunchy advertisements produced by GoDaddy.

- redd.it/nuimq - a post reminding people that it is now December 29th, the appointed day to transfer domains from GoDaddy, the post links to a guide to doing so.

- redd.it/nvdf4 and redd.it/nvg18 - these posts link to an article detailing a press release from GoDaddy in which they state that they now actively oppose SOPA.

In just one week a post to `reddit` about one user's decision to transfer their domain names away from GoDaddy sparked a boycott which appears to have forced this large corporation to change its stance on SOPA. Reviewing the posts about this which appeared on `reddit`'s Front page (and there were also many posts about this which did not reach the Front page) offers some insight into how this came to pass. On the original post's comments page it quickly becomes clear that there is support for a boycott among `reddit` users, and that some users have control of a large enough number of domains to make an impact on GoDaddy's business. Very quickly a post appears at the top of the technology subreddit drawing /r/technology readers attention to what's happening on /r/politics, spreading the word to a greater portion of `reddit`'s userbase.

At this point (23rd December) something pivotal happens. External websites (`Techdirt` and `Ars Technica`) publish articles about what's happening on `reddit`; links to these articles are then disseminated widely through `reddit`. This feedback tells `reddit` users that what they are doing is being noticed, and presumably adds to their enthusiasm to make the boycott happen. At this stage there are also posts linking to articles or announcements about recognisable names (`Cheezburger network`) joining `reddit`'s boycott; and posts in which the `reddit` community reach out to other

entities that they see as being on their side (`Wikipedia`, `Imgur`, Google) to ask them to also join the boycott. There are further posts announcing and celebrating the fact that some of these external entities did join the boycott. Then we have GoDaddy releasing statements addressing the boycott, further evidence for the `reddit` community that what they are doing is having a noticable impact.

GoDaddy then releases a statement in which they back away from their support of SOPA. If `reddit` users were operating in isolation then presumably some portion of would-be boycotters would have taken this as a victory and an end to the boycott. Instead we have a number of posts on the `reddit` Front page condemming this half-hearted 'flip-flop' and urging redditors to continue the boycott. We then have articles which put a number on the effectiveness of the boycott (21,000 domains lost in a day, 72,000 in a week), again reinfocing the notion that the boycott is happening and the results are being felt by GoDaddy.

Mixed in among these we also have posts and comments where `reddit` users offer advice on how to go about transferring domain names from GoDaddy, and who these domain names might be transferred to. There are also posts and comments which criticise GoDaddy, based on investigations of the CEO's past or some immoral business practices employed by GoDaddy.

On December 29th there is a post on the Front page reminding people that 'today is the day to transfer domains away from GoDaddy'. This is another indication of `reddit` working as a collective entity. A post by an ordinary `reddit` user puts forward an idea (boycotting GoDaddy) and picks an arbitrary date for its execution. The `reddit` community adopt this idea as a plan that they will attempt to execute. When the appointed date arrives another ordinary `reddit` user makes a post to remind people and the `reddit` community collectively up-vote this to the Front page.

A short time later GoDaddy released another statement in which they announce that they are now opposed to SOPA. After this there are very few posts appearing on `reddit`'s Front page about the GoDaddy boycott.

Boycott fever

While the boycott of GoDaddy was frequently the subject Front page posts a number of posts also appeared on the Front page which advocated or discussed boycotting as a more general strategy. Between December 24th-27th (during the week-long peak of 'GoDaddy boycott' posts) there were six posts appearing on the Front page which discussed or advocated further boycotts.

The first of these posts was submitted to /r/AskReddit and suggested a boycott of movie theatres (redd.it/nokhw), with the reason being that Hollywood was one of SOPA's biggest supporters. There is a lack of enthusiasm for this boycott among the popular comments, with one insightful comment suggesting that Hollywood is too large a target and that the success of the GoDaddy boycott was due to `reddit` having 'connections' in their domain. Another comment by a projec-

tionist in an independant cinema argues that companies like theirs would suffer most despite not being SOPA supporters. On the same day a similar post was submitted to /r/politics and also reached the Front page (redd.it/nob8i). This post was titled "Next SOPA boycott (we'll do one at a time) "Until Time Warner drops all financial support for SOPA voting congressmen we will buy no music nor any film ticket from their subsidiaries - We'll just choose any another label" ". Again many of the popular comments for this post suggest that it would not work because Hollywood is too large and `reddit` cannot reach a large enough segment of their market.

The next day saw a post to /r/AskReddit appearing on the Front page asking redditors what they would do if Conde Nast (`reddit`'s parent company) supported SOPA (redd.it/npob1). Popular comments on this post were mostly humerous in nature or concerned the difficulty in giving up `reddit`.

There are also two posts appearing on the Front page at this time which question the focus on GoDaddy as the only target of a boycott (redd.it/nq7cy and redd.it/nqumv). The popular comments on these posts re-iterate the point that GoDaddy was a good target for `reddit` because of the large overlap between their customers and people who `reddit` can reach (with `reddit` users assumed to be more technologically savvy than the general population). The second of these posts suggested a boycott of EA Games over their support of SOPA. Another post on December 26th suggests a boycott of Nintento (redd.it/nr2m3).

These posts are interesting for two reasons. The first of these concerns the expression of dissenting opinions (Sunstein, 2002). If we consider only posts about the GoDaddy boycott this idea seems to have emerged largely from a single post and very quickly received widespread support from `reddit` users. The posts considered here reveal that not all of `reddit`'s users jumped on the 'GoDaddy boycott' bandwagon, there are posts which question the focus on a single target (redd.it/nqumv in particular is quite critical in tone - "How can `reddit` take itself seriously if we only boycott websites that we don't use, or don't have good replacements for(like GoDaddy). EA also supports SOPA. Stop playing BF3 and let's boycott it."). These posts which question the strategy of only boycotting GoDaddy appeared on the Front page alongside posts which were very supportive of the GoDaddy boycott. This can be taken as a further indication that it is possible for opinions which disagree with the majority to be broadcast through the Front page. `Reddit` users, through their use of the voting system, collectively decide to broadcast opinions on the Front page which go against the perceived majority opinion.

The second point of interest here concerns the importance of comments. Although these posts reached the Front page their comments pages hosted discussions which tended to argue against the suggestion made by the post itself. Many popular comments on these posts explain why it is prudent to focus on GoDaddy or criticise the suggested target of the post. It is probably no coincidence that, despite appearing on the Front page, these posts failed to establish the momentum

on `reddit` which was critical to the success of the GoDaddy boycott. This trend could also be viewed as further evidence that voting on posts tends to be different in nature (parhaps more shallow) to voting on comments.

Following the failure of these posts to launch new boycotts, posts in this vein revert back to merely identifying corporations who support SOPA (e.g. redd.it/nsfma) without explicitly calling for a boycott.

7.2.5 Political Allies and Enemies

We have seen that the identification of corporations who supported or opposed SOPA was a strong theme to `reddit`'s coverage of the story. Another strong theme to the coverage concerned the identification of politicians who supported or opposed SOPA, and later ways to take action against/for these politicians.

In the early stages of this story, that is November 2011, there are five posts which identify politicians who support (redd.it/mmsg5 - Al Franken, redd.it/mnafo - Lamar Smith) or oppose (redd.it/mdqnf - Ron Paul, redd.it/mgg28 - Nancy Pelosi and Ron Paul, redd.it/ml0ii - Ron Wyden) SOPA.

SOPA supporters	SOPA Opponents
Lamar Smith (Texas Congressman - Republican - SOPA sponsor)	Sheriff Richard Mack (Opponent in Republican Primary)
Al Franken (Minnesota Senator - Democrat - PIPA co-sponsor)	
	Ron Paul (Texas Congressman - Republican)
	Ron Wyden (Oregon Senator - Democrat)
Lindsey Graham (South Carolina Senator - Republican)	
Paul Ryan (Wisconsin Congressman - Republican)	Rob Zerban (Ryan's opponent in upcoming election)

Table 7.1: Showing politicians who supported and opposed SOPA - individuals on the same row are direct competitors

The first post to suggest taking action in relation to politicians appears on the Front page on 15th December (redd.it/ntfzw). The suggestion put forward here is that a list of politicians who vote in favour of SOPA should be created and maintained, then the next time these politicians are up for re-election `reddit` users would not support them. On the next day there is a post in which a `reddit` user states they are willing to run against Lamar Smith in the next election (redd.it/nehr6). This post elicits over 1,000 comments, with many popular comments offering support or advice. However, over the course of the discussion it is discovered that the potential candidate is one day too late to submit the paperwork required for them to stand.

Following this there is a lull in posts concerning politicians until December 28th when a surge in posts of this nature begins. The first post in this surge (redd.it/ntfzw) is titled "Let's pick ONE Senator of voted for NDAA/SOPA and destroy him like we're doing for GoDaddy. Relentlessly investigate and find skeletons in his closet, money bomb his opponents, etc." (NDAA is another piece of legslation which was receiving a lot of negative coverage from /r/politics). The success of

this post in reaching the Front page and receiving over 1,500 comments is likely due in part to a feeling of empowerment within the `reddit` community originating with the concurrent GoDaddy boycott. In these comments users set out their proposed targets and detail their reasoning, or offer advice/help with making the effort a success.

Operations 'Graham Cracker' and 'Pull Ryan'

The top scoring comment nominates Senator Lindsay Graham as the target and names the initiative "Operation Graham Cracker". Within a few hours a professional web developer replies to say that they have bought the domain `www.operationgrahamcracker.com` and intend to produce a website to aid coordination, and have already started a `Facebook` page. Another post is submitted announcing 'Operation Graham Cracker' (redd.it/ntpl7) and directing to a newly formed subreddit (reddit.com/r/grahamcracker) which more than 1,700 users subscribe to. This is a pattern which has been observed many times on `reddit` when some new idea for collective action is put forward and appears popular (e.g. /r/LeakSpin and /r/WikileaksAnalysis from the `Wikileaks` case study); a website and subreddit is usually created within hours and often some method of communication external to `reddit` is established (e.g. IRC chat-rooms, `Facebook` pages).

These initiatives often fizzle out after a short period of time, and Operation Graham Cracker is no exception, although the manner of its demise is unusual. In this case the `reddit` community switched to a different target (Paul Ryan) very quickly. The process whereby the target changed is worthy of scrutiny. On the /r/grahamcracker subreddit a number of posts were submitted which first express concerns about the choice of target (redd.it/ntr3k) then suggest other targets (redd.it/ntpl7, redd.it/ntulb). A poll is held (redd.it/ntve0) using `reddit`'s comment voting system (top level comments are names of candidates, users vote to state their preferred candidate) and Paul Ryan wins by some margin (592 points as compared to second place Lamar Smith on 54 points - figures observed several months later). A post is created on /r/grahamcracker announcing the change of target (redd.it/nu5th) and this post receives up-votes from a large proportion of interested users (the subreddit had around 1,700 subscribers and this post has a score of 1,392). This concerted up-voting leaves the post re-directing to the new campaign (operation 'Pull Ryan') at the very top of the Graham Cracker subreddit for months afterwards, when the community up-rooted and re-located to a new subreddit they left a signpost behind for anyone that subsequently finds their way unto the Graham Cracker subreddit.

The salient characteristic of this endeavour is its speed. The post which suggested picking a political target was submitted at 0937 GMT on December 28th, this post received more than 1,500 comments and later in the day another post which announced 'Operation Graham Cracker' also received more than 1,000 comments. Within 14 hours of this original post's submission `reddit` had picked a target, created a website and subreddit to facilitate further organisation, discussed

alternative targets, held a poll to select an alternative target, and then created a subreddit and web page to facilitate the new campaign (Operation Pull Ryan). While the speed at which all of this unfolded is impressive it is perhaps also problematic and greater care may have been beneficial. The reasons cited for a switch from Lindsay Graham as the target were that he was too well entrenched in his seat and that he was not up for re-election until 2014. These objections were raised in comments to the post which started all of this, but it appears that by the time these comments appeared in prominent locations a group of users had already set the wheels of 'Operation Graham Cracker' in motion.

This activity could also be described as chaotic, but perhaps this is to be expected given the openness of the process. `Reddit` allows anyone to submit or vote on any post or comment they wish, and the website is designed to produce a ranked list of posts which changes from minute to minute. `Reddit` has not been designed for the purposes its voting system is being used for in this instance, namely deciding on the target of a political campaign, discussing and deciding upon a strategy and then facilitating the execution of the agreed-upon strategy. This is further illustrated by a post which appeared on the Front page on December 29th in addendum to the posts about picking a political target of the previous day. This post (redd.it/nuk78) argues in favour of picking two targets (one Republican and one Democrat) instead of one, and receives 890 comments (some of which argue against the proposal). This post also spawns its own subreddit for further discussion (/r/1red1bluekick2) but in this case the endeavour appears to run out of momentum quite quickly.

'Operation Pull Ryan' persisted beyond the point where Paul Ryan announced that he was not supporting SOPA. Much of the activity related to this happened on the /r/OperationPullRyan subreddit and will not be considered here, instead we will consider the effect this campaign had through the `reddit` Front page. One beneficiary of 'Operation Pull Ryan' was Paul Ryan's opponent in the upcoming election, Rob Zerban. Within one day of the announcement of 'Operation Pull Ryan' there are several posts appearing on `reddit`'s Front page which favour Rob Zerban. The first of these (redd.it/ntxr8) is a post by a `reddit` user who had contacted Rob Zerban about SOPA. This post consists of a screenshoted email reply from Rob Zerban stating that he did not like SOPA.

Within hours Rob Zerban himself (or at least someone from his campaign) enters the fray, creating an Ask Me Anything (AMA) post in response to 'calls and emails from redditors'. This AMA post (redd.it/nu501) is quite successful, seeing more than 2,000 comments and with Rob Zerban voicing many responses which are in line with popular opinion on `reddit`. It appears on the default Front page and an /r/bestof post linking to the AMA post also appears on the Front page (this /r/bestof post highlights the fact that Rob Zerban's account thanked every person who volunteered through direct replies to their comments). While the responses from redditors on the AMA post are generally positive, comments on the /r/bestof post are more critical. Some of these dismiss

the thanking of individual users as an example of electioneering and suggest that this part of the AMA was probably conducted by an unpaid intern. Some users criticise his responses as being vague and non-committal, others celebrate his willingness to come and speak to `reddit` users on their own territory.

The next two days see another two 'thank you' posts from Rob Zerban's `reddit` account appearing on the Front page. The first of these (redd.it/nwez5) thanks `reddit` for their support in the form of $5,000 in donations, the 'great press' his AMA generated, and thousands of emails offering support and signing up to volunteer. The second of these posts (redd.it/nxlgc) updates the figure to $15,000. The top comment on this second post suggests a possible line of attack that Zerban might consider employing against Paul Ryan.

While this is unfolding, a post linking to a *The Nation* article about the possibility of Paul Ryan being defeated (redd.it/nttdl) is also up-voted to the Front page. The next day a post appears on the Front page which links to a *The Atlantic Wire* article about 'Operation Pull Ryan' (redd.it/nv6zr). This article states that `reddit` forced Paul Ryan to 'play defence on SOPA', but also notes that it is not clear that Paul Ryan had ever actually identified himself as being a supporter of SOPA.

This sequence of events draws to a close (in terms of posts appearing on the Front page) ten days later when Rob Zerban's account creates a post titled "Reddit successfully pressures Rep. Paul Ryan (R-WI) to back off support of SOPA." (redd.it/o9gq7). Some popular comments on this post congratulate or endorse Rob Zerban, while others dismiss the post as an attempt by Zerban to engratiate himself with the `reddit` community, questioning whether Paul Ryan was ever actually a SOPA supporter or whether `reddit` had anything to do with his latest statement.

This story is interesting from the perspective of a politician engaging with Social Media (and social news in particular). Although this is not the first time that an American politician had conducted an AMA interview on `reddit`, it is the first instance I have observed of sustained `reddit` activity on behalf of the politician. This case also differs from previous instances in that Rob Zerban was relatively unknown to `reddit` before 'Operation Pull Ryan' began but seems to have quickly earned `reddit`'s good graces (in the form of donations, volunteering and a positive voting response to future submissions on `reddit`). In previous cases where a political figure has conducted an AMA interview they tended to be well known (and usually liked) by `reddit` before conducting the interview.

There are a number of facets of Rob Zerban's interactions with `reddit` which suggest a familiarity with `reddit` on the part of himself or (perhaps more likely) a member of his campaign staff. In his first post (the post on which he conducted the AMA, redd.it/nu501) he demonstrates a familiarity with some `reddit`-specific norms. For example, in the first sentance of this post ("Hello `reddit`. I've been getting calls and emails from redditors today and have been encouraged to participate in

an AMA session.") he addresses `reddit` users collectively with 'Hello `reddit`' and uses 'redditors' as the plural. Zerban also demonstrates a willingness to engage with `reddit` beyond the AMA. One comment on his AMA post expresses dismay that `RobZerban.com` is registered with GoDaddy, the next day Zerban makes a post to the 'Operation Pull Ryan' subreddit containing an open letter to GoDaddy in which he explains that he will be transferring his domain due to their support of SOPA.

On the whole it appears that Rob Zerban's approach to `reddit` has been quite successful. He takes the opportunity granted by a highly visible post to direct redditors towards his website, `Facebook` and `Twitter` accounts - avenues through which they might have a more sustained interaction with his campaign, perhaps as a supporter. This is a common occurance, AMA posts by recongisable individuals frequently serve as a non-too-thinly veiled marketing ploy and take the opportunity to direct `reddit` users elsewhere (for an example which takes this too far and backfires see the Woody Harrelson AMA, where questions were only answered if an endorsement of 'Rampart' could be incorporated - redd.it/p9a1v). However, Zerban also demonstrates a willingness to meet `reddit` part-way, by conforming to `reddit`-specific norms and responding to suggestions with positive action. Without these gestures `reddit` users may have felt that the AMA interview was a one-off event that served simply as part of Zerban's Social Media strategy. As things unfolded Zerban seems to have made a positive impression on `reddit`, as attested to by a positive voting response to his subsequent submissions and a steady stream of donations which lasted for several days at least.

The positive response Rob Zerban received through `reddit` may have helped to establish social news websites as a part of the newly emerging Social Media arena around politics. In fact, between the time when Rob Zerban's AMA post was submitted (December 28th) until the end of the obsevration period (February 2nd) there are another two AMA-style posts from politicians which appeared on `reddit`'s Front page (redd.it/oicwg and redd.it/opgol).

Some time later (August 2012) Barrack Obama also participated in an AMA interview (redd.it/z1c9z) as part of his campaign for re-election.

Lambasting Lamar Smith

Of all the posts about SOPA which concerned specified politicians and appeared on `reddit`'s Front page, Lamar Smith is mentioned by name in the titles of fourteen (all of these having a negative disposition towards him), much more than for any other individual. What is it about Lamar Smith which raised the ire of `reddit` to this degree?

Lamar Smith is a US Congressman representing a district in Texas, he is also the principle author of SOPA and this is the attribute which first brought him to `reddit`'s attention in a post which

appeared on the Front page on November 24th (redd.it/mnafo). The fact that Smith was the author and driving force behind SOPA is no doubt inplicated in the lambasting he received on `reddit`. However, posts concerning Smith specifically tended to occur later in the observation period. Between the time when Smith was identified as the author of SOPA (on November 24th) and January 2nd there was just one other post which mentioned him by name appearing on the Front page (redd.it/nehr6).

On January 3rd, posts concerning Lamar Smith began to appear on the Front page much more frequently. The first two of these posts link to articles which quote Smith making comments which are dismissive of the mounting public opposition to SOPA (redd.it/o19qi on January 3rd - "Lamar Smith Out Of Touch With The Internet: Still Thinks It's Just Google That Opposes SOPA"; redd.it/o326l on January 4th - "Smith Says `reddit` SOPA Protestors are 'Not Legitimate or Large in Number' "). The following four days see a further two posts about Lamar Smith in a similar vein appearing on the Front page (redd.it/o5bi9 - "Rep. Lamar Smith Decides Lying About, Insulting And Dismissing Opposition To SOPA Is A Winning Strategy", and redd.it/o5bi9 - "SOPA sponsor Rep. Lamar Smith to SOPA opponents: You dont matter"). Popular comments on each of these four posts range from the critical to the vitriolic, and seem to establish Smith as a legitimate target for `reddit`.

The next post about Lamar Smith (appearing on the Front page January 12th) links to a *Vice* article which reveals that Smith's website violates a photographer's copyright (redd.it/oe17z). The same message appears on the Front page one week later (redd.it/omffp) with the title "Scumbag SOPA Author", this time a `reddit` user had re-worked the content of the *Vice* article into a single image, this post's top comment directs redditors to the original *Vice* article.

The remaining posts about Lamar Smith appeared after SOPA was officially shelved on January 20th, suggesting that `reddit`'s anger towards Smith persisted beyond SOPA. One of these posts links to a picture of Lamar Smith which has been doctored (presumably by a `reddit` user) with a super-imposed 'head crab' (monster from the popular computer game Half-Life), the implication being that Smith is a zombified puppet of the entertainment industries.

The next post concerning Smith appears on the Front page on January 27th and attempts to start a campaign to thwart his re-election (redd.it/oyd16 - "Lamar Smith, the scumbag behind SOPA, is up for re-election this year. Let's make sure his opponent wins!"). In this case the proposed strategy is to support his opponent, Sheriff Richard Mack, an alternative Republican candidate who is described as '[not] exactly the kind of candidate that `reddit` would normally support' but the lesser of two evils. This idea appears to have some support in the comments, and a new subreddit is created (/r/SheriffMack4Congress/) for the cause. As part of this effort Sheriff Mack was invited to conduct an AMA interview on `reddit`, and when he consented this post was up-voted to the Front page (redd.it/pap79) on February 4th, one day after the observation period

for this case study. The AMA appears to have been quite successful with over 1,500 comments, and in the process Sheriff Mack and `reddit` appear to have discovered some common ground. As of May 2012 the SheriffMack4Congress subreddit is still active with 428 subscribers.

On February 1st another post appears on the Front page which references Lamar Smith (redd.it/p67jz), this time calling attention to a new piece of legislation (the 'Protecting Children from Internet Pornographers Act') which Smith is associated with. This post will be considered further in section 7.2.7 about coverage after SOPA was shelved.

The final post about Lamar Smith appearing on the Front page within the observation period called for an Operation geared towards his impeachment (redd.it/p8i1n). The top comment directs users towards the Sheriff Mack for Congress subreddit/campaign. Other popular comments point out that seeking impeachment is not a wise approach as it requires that a felony has been committed, and suggest instead that efforts to work against his re-election might be more fruitful.

Reviewing the timing and nature of posts on the Front page concerning Lamar Smith reveals a strong similarity with the GoDaddy boycott posts. In both instances some of the early posts at the beginning of a flurry of activity linked to *articles about the subject making comments which are dismissive of a campaign that `reddit` is involved in.* If we consider `reddit` as a single social entity these occurances suggest that it may be possible to provoke this entity, with the result being a subsequent tendancy for negative posts about the provoking party to appear on the Front page.

In the case of Lamar Smith the series of events likely unfolded as follows:

- `reddit` users are engaged in a campaign against SOPA

- SOPA author Lamar Smith makes comments about this campaign which dismiss it as a vocal minority

- An article is written about these comments

- A link to the article is submitted to `reddit`, as it concerns a 'hot topic' it is up-voted to the Front page

- Many `reddit` users see the article about these comments

- These redditors are then more likely to submit or upvote posts which are critical of Lamar Smith - with the result that more of this type of post will appear on the Front page

If this series of events is accurate it suggests that when posts about Lamar Smith dismissing the Internet campaign against SOPA appeared on the Front page - redditors (or some portion thereof) may have taken this as a personal slight against themselves. If it really is possible to 'piss off `reddit`' in this manner by making comments about a campaign that `reddit` users are involved in, this suggests that being 'a redditor' may be an important part of these individuals' identity

(Tajfel, 2010).

If we conceive of `reddit` as an unconnected mass of disorganised individuals who happen to use the same Distributed Moderation based website, it is difficult to account for patterns such as the one described here. This is taken as further evidence that to understand what's happening on and through `reddit` we need to conceive of `reddit` as a unitary social entity in some cases.

In Lamar Smith's case `reddit` also demonstrates a surprising level of persistence. In the months following the observation period posts which made reference to Lamar Smith appeared occasionally on `reddit`'s Front page. There was an AMA post by another would-be competitor in March (redd.it/rkyav) and a post about 'Don't mess with the Internet!' billboards being put up in Smith's district. On May 28th a post appeared on `reddit`'s Front page which served to remind redditors in Texas that the Reuplican primary (or "vote to oust SOPA posterboy Lamar Smith") was the following day. The text of this post contained links to major posts and events which `reddit` was involved in - links to the AMA posts with two of Smith's opponents and links relating to the billboard and TV advertisement which `reddit`'s 'Test PAC' paid for (see sectiion 7.2.7).

7.2.6 Taking action

This section considers collective action posts on `reddit` in relation to SOPA. Previous sections have discussed some of the ways in which `reddit` users took or attempted to take action (GoDaddy boycott, campaigns for/against politicians), this section will consider some of the other ways in which `reddit` attempted to oppose SOPA. Early posts which called on `reddit` users to act most often advocated forms of low-intensity activism (or 'slacktivism' as this is sometimes termed) like signing a petition or sending a letter/email to their representatives. As the story progressed calls for collective action become more creative, with the GoDaddy boycott and the blackout being two of the most notable endeavours.

Petitions

In total there were eight posts appearing on `reddit`'s Front page which called on users to 'sign' petitions expressing opposition to SOPA. Signing a petition requires minimal effort and `reddit`'s capacity to broadcast a link to hundreds of thousands of individuals quickly means that it is often involved in generating many signitures for petitions which resonate with `reddit`'s users. When `reddit` begins to rally around some cause often the first posts calling for users to act involve petitions, and the SOPA case is no exception. Four of the posts which linked to petitions appeared on the Front page in November while the other four appeared there in December. Three of these petitions were hosted on `WhiteHouse.gov` while the other five were created by various activist organisations.

While the comments on 'petition posts' often bemoan the ineffectiveness of petitions as a method of action these posts still regularly appear on the Front page and generate many signitures. These petition posts may serve as a way to establish that `reddit` users are in agreement on a cause and willing to act. The effort required to sign a petition is minimal so perhaps this is a good starting point for a campaign of collective action. Users who agree with the cause can sign the petition in a matter of seconds, if the petition does not pick up momentum these users have lost very little. While 50,000 signitures on a `WhiteHouse.gov` petition are not going to bring about a change in policy they may have more significance *within the `reddit` community*. A successful petition demonstrates that the `reddit` community is behind a cause in principle and in numbers, and this may be a useful base from which to attempt more ambitious and effortful forms of action.

The Blackout

On January 18th many high-profile websites 'blacked out' in a protest over SOPA - `reddit` and `Wikipedia` effectively shut their sites down while Google positioned prominent links to SOPA information on their front page (for US users). This was the highest profile event in the 'Internet campaign' against SOPA and `reddit` was one of the first entities to commit to a blackout and set a date. The decision to take `reddit` offline in protest was ultimately made by `reddit`'s administrators, but for several weeks before they made this decision there were high-profile posts in which ordinary users put forward and discussed the possibility of a blackout.

In total there were seventeen posts about a blackout which appeared on `reddit`'s Front page before the blackout actually took place, and several afterwards concerning its effectiveness. Posts about a blackout can be grouped into two temporal intervals.

The first post on `reddit`'s Front page to mention a blackout appeared there on 13th December and linked to an article about `Wikipedia` co-founder Jimmy Wales considering a `Wikipedia` blackout (redd.it/na2fm). Two days later on the 15th of December there were a further three posts about a blackout appearing on the Front page. The first of these called for online pornography websites to shut down for six hours that night in protest over SOPA (redd.it/nd67r), some of the popular comments on this post suggest that a Google blackout would be more effective or link to the previous article about `Wikipedia` considering a blackout. Next to hit the Front page is a post which again links to the article about a possible `Wikipedia` blackout but this time with the title "Why not black out `reddit` in protest of SOPA? ..." (redd.it/nddq7). It is noteworthy here that the top comment on this post suggests that blacking out `reddit` would be 'preaching to the choir'. This comment is replied to by a `reddit` administrator who agrees and says that `reddit` can be useful as a hub of anti-SOPA activism, this comment also states that a link to the SOPA subreddit has been added to the Front page. In this manner `reddit`'s adminisrators seem to be nudging users towards SOPA activism. Also on December 15th a post appears on the Front page

in which one user says that they have taken their own company's website offline in protest over SOPA (redd.it/ndz56).

Following this there is a lull in posts about a Blackout which lasts until January 3rd, during this gap the GoDaddy boycott took place and it appears that `reddit`'s focus shifted from blackouts to boycotts. The post on January 3rd links to an article about Google, Amazon, `Facebook` and `Twitter` considering a blackout.

Talk of a blackout sees a resurgence on January 7th when there were two posts on the Front page about this topic. The first of these links to an article in *Time* about the significance of a possible blackout by Google, `Facebook` and `Twitter` (redd.it/o6g4k). Later that day there is another post by an 'ordinary' `reddit` user concerning a blackout of `reddit` (redd.it/o6ymp), this time asking how `reddit` users would feel if `reddit` blacked out for a day. This time popular comments are more supportive of the idea, provided links to SOPA information are visible during the blackout. The next day a well-known moderator makes a post stating that the moderators of major subreddits are discussing the possibility of blacking out their subreddits for a day and asking `reddit` users how they would feel about it. This post saw very high levels of activity, more than 4,000 comments. Again popular comments are supportive of the idea as long as links to information about SOPA are provided prominently during the blackout. This post is interesting because it suggests that had `reddit`'s administrators not decided to blackout the whole website moderators may have attempted to achieve something similar independantly.

On January 10th `reddit`'s administrators made a blog post announcing that `reddit` would be taken offline on January 18th in protest over SOPA (redd.it/obg8v). Popular comments are supportive and hopeful that other entities (Google, `Wikipedia`, `Facebook`) will also blackout on the same day. Shortly after this blog post hits the Front page there are posts linking to articles about `reddit`'s decision to blackout (redd.it/oc06f, redd.it/oce5g). This is another example of the feedback loop whereby something that happens on `reddit` is echoed back through external reporting.

On January 11th a post reaches the Front page which asks redditors to "Tell `Facebook` to go dark on January 18 to spread awareness of SOPA/PROTECT IP" and links to the `Facebook` contact page (redd.it/ocqxi). Popular comments on this post contain suggested templates that people might use when they contact `Facebook`. On January 15th, a similar post appears this time concerning Google, this post does not initially call on redditors to contact google but the suggestion that they should send messages to Google on `Twitter` is edited into the original post in response to the discussion on the post's comments page.

In addition to these posts calling on external entities to join a blackout, there are posts celebrating when some of these entities announce their own blackout. On January 12th one such post concerns the *Cheezeburger network* (redd.it/oeaxa). On January 16th there are two posts about `Wikipedia`'s

announcement that they will join a blackout on January 18th (redd.it/ojh20 and redd.it/ojnkk). On the day of the blackout itself there are a further two posts (appearing before `reddit` was taken offline) pointing out that Google has placed a SOPA awareness doodle on its Main page (redd.it/olznp and redd.it/olzo5).

On the day after the blackout (January 19th) there are four posts appearing on `reddit`'s Front page about political support for the legislation collapsing (redd.it/omc0e, redd.it/omclw, redd.it/on6q8 and redd.it/oo64j). The mood in the popular comments on these posts is that `reddit` has been involved in achieving something important, but that this is just the beginning of a longer struggle to protect freedoms which exist on the Internet. On the next day SOPA is officially withdrawn (redd.it/op5su) but again popular comments do not adopt a tone of self-congratulation but instead urge vigilance in the long term as the objectionble ideas in SOPA are likely to re-emerge in other bills.

7.2.7 Maintaining momentum

On the day when SOPA was officially withdrawn (January 20th) there are two posts appearing on the Front page which urge `reddit` users to now focus on new Internet regulating legislation (redd.it/op9z0 - "Threatening New Bill - Worse Than SOPA/PIPA (Make This Our Next Target `reddit`" and redd.it/opc2u - "H.R.1981, the real bill by lamar smith that will cripple the internet regardless of our efforts. SOPA and PIPA were a smoke screen."). There appears to be a sense at this point that `reddit` has some momentum in its crusade to protect Internet freedoms and that to keep this momentum alive new targets are needed.

The next day sees the first post which attempts to channel this momentum towards ACTA (the Anti-Counterfeiting Trade Agreement). This post (redd.it/oplxh) is titled "Americans, we helped you and spread awareness about SOPA. Now it's time to help us. Europe has their own SOPA, called ACTA.". In what would become a common occurance, this post likens ACTA to SOPA whereas the similarities between the two are minimal, and some popular comments point this out. This is another sign that these posts constitute an effort to maintain momentum, SOPA is the hot topic that `reddit` has been rallying around and it is repeatedly invoked to draw attention to other pieces of legislation that in reality bear very little similarity to SOPA.

In the 13 days between 21st January and the end of the observation period there were 20 posts pertaining to ACTA which were observed on `reddit`'s Front page, it appears that ACTA was chosen as the 'next target' for `reddit`. One of these posts (redd.it/oqdxw) emphasises `reddit`'s international nature ("European Redditors: What is ACTA, and what can U.S. Redditors do about it?").

In addition to the identification of new targets there are two posts which offer thoughts on the

long-term implications of the 'Internet campaign' against SOPA. The first of these (redd.it/orj14) is a text post about "why SOPA might be the best thing ever..." suggesting that "The younger, online generation just found out that if they pay attention to politics, participate, and speak up about issues they disagree with, they can actually impact legislation.". The second (redd.it/p4iod) links to an article about January 18th being the largest digital protest ever.

Throughout this period there are numerous popular comments and posts bemoaning the fact that legislators will continue to push the ideas in SOPA in future legislation, often accompanied by the conclusion that the legislators will win by out-lasting 'the Internet's' capacity to protest vociferously. Some of the ideas voiced in response to this are that people need to be more pro-active in calling for legislation which will protect the Internet, and that `reddit` needs to have a way to marshal its apparant political will and enthusiasm on a longer-term and more organised basis.

On January 19th a post (redd.it/onb6f) appears on `reddit`'s Front page announcing the formation of a `reddit` Political Action Committee (or PAC). One week later this PAC is featured on the Front page again (redd.it/oz8le) announcing the launch of its own website (`www.testpacpleaseignore.org`) and subreddit (/r/rpac). The name of the PAC (Test Pac, Please Ignore) is a reference to the highest-scoring post on `reddit` of all time (redd.it/92dd8). As of May 2012 the /r/rpac subreddit is still active with more than 5,000 subscribers, $20,000 dollars have been raised for the PAC, and this is currently being spent on the PAC's (democratically determined) first goal - unseating Lamar Smith from congress.

7.2.8 SOPA coverage and subreddits

This section considers the subreddits which SOPA posts that reached the Front page were submitted to. `Reddit` is comprised of thousands of subreddits which operate independantly and are aggregated (selectively, based on a user's subscriptions) to produce the Front page.

Between October 27th and November 30th there were 37 posts about SOPA on the Front page, 26 of these posts were submitted to the technology subreddit. The /r/technlogy subreddit was the first default subreddit to show an interest in SOPA. This is somewhat expected as the technology subreddit often features posts about legislation related to technology - and SOPA most certainly falls into this category. During this time the /r/politics subreddit, another subreddit which one might have expected to 'cover' SOPA, was devoting a lot of attention to the National Defence Authorization Act for 2012 - specifically its provisions for the detention of American citizens.

In December SOPA coverage penetrated a much broader range of subreddits, with almost as many posts coming from /r/politics as /r/technology. There is considerable overlap between the Front page posts coming from these subreddits, with at least 18 incidents observed where a very similar post (often linking to exactly the same article) appeared on the Front page through both of these

subreddit	October/November	December	January/February
total	37	100	122
technology	26	38	33
politics	5	34	30
worldnews	1	2	14
AskReddit	2	10	11
IAmA	0	3	5
gaming	1	4	6
movies	0	0	1
music	0	0	1
videos	1	0	1
pics	0	1	1
funny	0	3	4
AdviceAnimals	0	1	7
WTF	0	1	4
todayilearned	0	2	1
bestof	0	1	1
blog	1	0	2

Table 7.2: The number of posts per month appearing on the default Front page from individual subreddits

subreddits on the same day. This suggests a similarity in the opinions of users who vote on these subreddits and possibly a large overlap in users themselves. However it also suggests that redditors conceive of subreddits as seperate entities, the fact that a post has appeared on the Front page through /r/technology does not appear to be reason to downvote it when it is encountered on /r/politics. In one observed incident a post reached the Front page through /r/technology with the specific intention of highlightling a previous post to /r/politics (the post which instigated the GoDaddy boycott).

Posts reaching the Front page through the AskReddit subreddit (2 in November, 10 in December, 11 in January) also merit a mention here. These are posts where reddit users have asked each other about their opinions on SOPA, about the effectiveness of some course of action or about how to take action effectively. These posts are significant for two reasons. Firstly, they serve as places which are explicitly devoted to discussion and they are often used to directly gauge redditors' response to a suggestion. Secondly, unlike the politics and technology subreddits, /r/AskReddit is not dedicated to content which is political, serious or important in nature. It seems likely that there is less overlap between the readers of AskReddit and politics/technology than there is between those two subreddits. AskReddit may therefore have served as a place where reddit users

who are not usually interested in this kind of content would have been exposed to messages about SOPA. A similar argument could be made about the gaming subreddit.

The role of humour

Of the 259 Front page posts considered as part of this case study 19 have been classified as humerous in nature. Humerous posts about SOPA tend to either speculate on courses of action which could be taken if SOPA passed or (later) satirise individuals who are perceivd as SOPA supporters.

These posts begin appearing on the Front page on December 15th, co-inciding with a surge in `reddit`'s coverage of SOPA. The appearance of these posts on the Front page through subreddits like /r/funny and /r/AdviceAnimals likely marks the point where awareness of SOPA had reached `reddit`'s user-base as a whole. At this stage there had been many posts about SOPA appearing on the Front page through /r/technology and /r/politics. We can speculate that some users who were aware of SOPA began to make humerous posts and submit these to humour-oriented subreddits, and that there were enough subscribers to these subreddits who were aware of SOPA and could appreciate the humour to upvote the posts to the Front page. When these posts appeared on the Front page through these humour-oriented subreddits we can assume that 'SOPA awareness' had penetrated the majority of `reddit`'s user-base. If we conceive of default subreddits as being 'serious' or 'humorous' in nature then /r/politics and /r/adviceanimals would probably lie on opposite ends of this spectrum, and most users who subscribe to any of the default subreddits likely subscribe to a subreddit which at this stage had hosted SOPA-related content in a prominent location.

Many of these humorous posts were image posts that were hosted on `Imgur.com`. This provides us with some extra information on the number of people who viewed SOPA-related content through `reddit` - as `Imgur.com` displays view counts for images hosted there. The humorous SOPA-related posts hosted on `Imgur` which had reached `reddit`'s Front page were viewed between 250,000 and 700,000 times each (view counts observed May 2012).

7.2.9 SOPA Summary

If we consider `reddit`'s coverage of the SOPA story in comparison to the way the conventional news media covers stories there are similarities and differences. Early posts about SOPA on `reddit`, and posts which are informative in nature, are most similar to the type of coverage one would expect to observe in the conventional media. For example, the first post about SOPA to appear on the `reddit` Front page linked to a `Techdirt` article about the legislation. This article provides a comprehensive and straightforward introduction to the legislation but it is also overwhelmingly critical in tone, whereas articles published by the conventional news media tend to be more neutral in tone. This

article also embeds the full text of the legislation within the same page, again a departure from the conventional news media.

If we consider the SOPA case study along with the earlier `Wikileaks` case study this suggests that when `reddit` begins to cover a story it will do so by linking to comprehensive external articles. If such articles are available in recognisable news outlets (as was the case in the `Wikileaks` study) these will be linked to by Front page `reddit` posts. The SOPA case study suggests that `reddit` can decide to cover a story which is *not* being covered by the conventional news media - and in this case Front page posts will link to wherever information about the story is available. This is important because it indicates that `reddit` is capable of setting its own agenda and is not limited to providing an alternative perspective on stories which are popularised by the conventional news media. This distinguishes it from the collective action organisation *38 degrees* as described by Chadwick (2013). In the case of SOPA `reddit` demonstrated persistence in its intent to cover this story for a period of 3 weeks before any elite older media entities showed an interest in picking it up, and there is a case to be made that `reddit` was instrumental in 'breaking' this story.

`Reddit` frequently used `Techdirt` as a source of information for this story, whereas it was not linked to by coverage of the `Wikileaks` story. In the former case this kind of foundational information was provided through links to the conventional news media, but with SOPA these sources were not available during the early phases of the coverage. This can be read as `reddit`'s users collectively realising the need for solid information about the story and recognising `techdirt` as a reliable provider for this. In a broader sense it is part of an emerging social media 'assemblage' (Chadwick, 2013) around coverage of this story, with a particularly strong interplay between `reddit` and the `Techdirt` blog - `reddit` provided attention for `Techdirt` and perhaps `Techdirt` reciprocated to some degree by continuing to cover the SOPA story which was drawing in so much traffic from `reddit`.

The SOPA case study also re-affirms the importance of comments pages in `reddit`'s coverage. Through these democratically ranked comments `reddit` users can hold a large-scale discussion of a post's contents - often criticising the post or bringing additional information to bear. It is quite common to see comments from a user who has some expertise in the relevant area, and for these comments to be highly upvoted. This is in some senses a similarity to the conventional news media, who will often invite an expert to give comment on a particular story. The difference is that a conventional news organisation will invite (and presumably pay) their chosen expert to contribute. On `reddit` these experts come from within the ranks of `reddit`'s users, they choose to make their comment of their own volition and it is up to the users who vote on comments to decide whether these comments warrant prominent positions.

During this case study there was also a post on `reddit`'s Front page which can be interpreted as `reddit` 'having a correspondent on location' - a capacity which is more often associated with

conventional media coverage . In this post (redd.it/ne9zn), titled "I am at the SOPA House Judiciary Committee hearing. AMA", a user who was present at the House Judiciary Committee provided updates and answered questions from other users. This capacity results from the combination of a large and growing user base and a voting system which is based on contribution quality. As `reddit`'s user base continues to grow so does the likelihood that some of the `reddit`'s users will be in a position to provide expert commentary or coverage - the voting system allows these contributions to be recognised and made visible even when they come from a `reddit` user who does not have an established history on `reddit`. The result is that `reddit` can provide expert comment or on-the-spot reporting on a convenience basis for free. These are services which the conventional news media pay to provide, and one might assume that they are aspects of news coverage which *require* the organisational capacity of a conventional, hierarchical, news company. This case study has shown that `reddit` can provide an analogous service through crowd-sourcing and 'citizen journalists'.

This case study also allows us to consider `reddit`'s response to a rapidly developing story. In comparing `reddit` to the conventional news media one would expect the conventional media to have an advantage here. When there is a development in a story which is being covered the journalist who has been assigned to cover the story need only consult their editor before publishing an update. On `reddit` there are no journalists who have been assigned to particular stories, and the role of an editor is being performed by thousands of uncoordinated individuals. Bearing this in mind, `reddit` demonstrated a surprising capacity to quickly 'publish' updates on a rapdily developing story. This episode offers hope that `reddit` has the capacity to correct mis-information cascades (Sunstein, 2002) in a timely manner.

We have thus far considered `reddit`'s information-oriented coverage, where similarities and analogies to the conventional news media are rife. However there is another side to `reddit`'s coverage of this story - posts which are action-oriented. This is where `reddit`'s coverage diverges from that of the conventional news media. The conventional news media, as a rule, do not serve as a platform on which grassroots collective action can emerge and become established. The conventional news media is operated on a one-to-many basis; professionals produce 'the news' and readers consume it, there is very little room for feedback from the readers and for readers to talk amongst themselves (although recently this is changing with the introduction of comments sections to the online offerings of many news producers) (Keen, 2007).

On `reddit` a post about one user's letter to their representative or decision to boycott a company is 'processed' in the same manner as a post linking to expert analysis - with placement on the Front page being the highest degree of visibility any post can achieve. Any `reddit` user with an idea or something to report has in principle the same chance of reaching the Front page as a professional who writes articles about a subject - this equality may be one of the characteristics which leads to

`reddit` serving as a platform for collective action.

There are feedback loops of attention on `reddit` which are also heavily implicated in its capacity to foster collective action. When a post concerning some idea for collective action appears on `reddit`'s Front page it has already passed the first hurdle towards becoming a reality - `reddit`'s voting users have approved it. Observations suggest that appearing on the Front page indicates only a shallow level of approval, perhaps an agreement with the sentiment behind the idea expressed in the post's title. Nonetheless, when a post appears on the Front page it is being put before a large audience of `reddit` users and it will then generate a large-scale discussion. Some ideas falter at this stage, with the most popular comments criticising the effectiveness of the proposed action. When the outcome of the discussion is positive there will be high-scoring comments endorsing it and others offering suggestions on how to increase its effectiveness. This also highlights the fact that `reddit` is a fundamentally public entity - the social machinations which determined the character of reddit's involvement in the story have all taken place in public and can therefore be consulted and referenced by users to understand *why* `reddit` is behaving in a certain way.

Where a post about some proposed form of action is successful at this stage it will have first been shown to many users on the Front page - if these users view its comments page they will be presented with supporting arguments which have been endorsed by the `reddit` community (i.e. comments with large scores). At this stage there are many `reddit` users who are both familiar with and supportive of the idea. Subsequent posts which concern the idea are more likely to perform well through the voting system (as many voters will have already seen the idea and been persuaded of its merits) - this knowledge and attitude 'cascades' through the user-base through repeated exposures on the Front page.

If the proposed collective action acquires momentum and people begin to take action this may result in an external entity publishing an article about the action taking place through `reddit`. When such an article is published it often appears on the `reddit` Front page. The appearance of this kind of article on the `reddit` Front page likely reinforces the sense among `reddit` users that the action they have embarked upon is being noticed and having an effect. This would in turn increase the chances that subsequent posts about the collective action will be upvoted to the Front page.

When an idea for collective action takes root on `reddit` it does so in a manner which is both fast and chaotic. The salient example from this case study is 'Operation Graham Cracker'. A user submitted a post calling for the targeting of one pro-SOPA politician. Within 14 hours of this post's submission `reddit` had picked a target, created a website and subreddit to facilitate further organisation, discussed alternative targets, held a poll to select an alternative target, and then created a subreddit and web page to facilitate the new campaign. This speed and chaos is due to `reddit`'s anarchic streak and lack of defined leaders. When an idea for action begins to take off

on `reddit` it will need a website and spaces in which people can communicate about it. There will usually be many people who see the idea and have enthusiasm for it and also the skills/time to produce something to aid with the idea's implementation (i.e. a website, an IRC chat-room). There is no recognised leader in the endeavour to consult, therefore the first person to take action in providing for some need becomes the de facto provider of that service. This anarchy allows redditors to get something going very quickly but it can also result in users undertaking work which is pointless (e.g. a `reddit` user now owns `operationgrahamcracker.com`, a domain for a political campaign which was dropped within hours of its conception).

The speed with which collective action proposals evolved and with which the a mis-information cascade was corrected also serve to highlight the importance of 'high temporal resolution' in data for this kind of research - a commonality for any research which deals with the contemporary Hybrid Media System (Elmer, 2013; Chadwick, 2013). Where once it was sufficient to know the day/edition of an article or the time slot for a news bulletin it is now necessary to have a timestamp which is as accurate as possible - a further indication of the more frantic pace at which media now operates. This fact emerges quite quickly once one begins to embark upon research of this nature - and the availability of data which allowed for the story to be tracked in such temporal detail offers some vindication for the data collection regime designed and implemented more than a year before the SOPA case study's inception.

The decision to collect data on a continuing basis has also been provident here. Without this approach the case study would only have been possible if I had realised SOPA coverage was worthy of study before posts about it began to appear on `reddit`, and even then there would be ambiguity about whether I had captured the first posts about it to appear on the Front page or whether it had appeared there before without my knowledge. There is no resource which allows one to re-construct the contents of `reddit`'s Front page at a given moment in time, aside from the data collected for this research. [3]

If we consider the collective action in this case study in comparison to campaigns organised by an organisation like *38 degrees* or *MoveOn* there are a lot of similarities in the types of 'action' which are requested of users, drawn mostly from a well established set of collective action repetoires (Bimber et al., 2012) - all three platforms have successfully executed 'campaigns' involving signing petitions, contacting representatives, donating money to purchase an advertisement, attending a protest, etc. There are also similarities in the strategic approaches of these campaigns - with trying to gain conventional media attention for a perspective, and targeting the decision-makers directly, being two common approaches. A further similarity is the speed with which collective action proposals can be put forward and executed (Chadwick, 2013, p.191).

---

[3]There are however services which allow one to get a rough idea of which posts were popular on a given day, and these have been used to fill gaps when my data collection server was down.

The major disparities between `reddit` and *38 degrees/MoveOn* are in how these strategies are developed and executed - chiefly the absence of pre-determined leaders on `reddit`. On `reddit` the voting system is the key mediator between participants who are fundamentally equal, and it is through this mediation that every aspect of the collective action endeavour is determined and reported. *38 degrees* does not follow the 'peer production' model whereby all users are essentially equal - but rather has a clearly defined leadership who set the agenda and decide which actions are placed on the website.

> "The leaders acknowledge that the decision to call those on its e-mail list 'members' was a deliberate attempt to encourage a sense of shared identity in the absence of organizational mechanisms..." (Chadwick, 2013, p. 189)

The relationship between *38degrees* leaders and members is an unusual one and could be described as analogous to a politican and their electorate or a marketer and their market. Leaders pay very close attention to the discourse of their members and also poll members directly to aid in the determination of which actions to attempt, and members 'vote' to determine the actions that will be 'selected' by indicating their intent to participate. Leaders 'test' ideas by emailing them to small samples of the member list - using early (non-)responses to determine whether a nascent action should be abandoned, and even sending out multiple versions of the email which frame the issue in different ways or altering the format of presentation to fine-tune the action's presentation. (Chadwick, 2013)

Collective action endeavours on `reddit` showed considerable similarity to those orchestrated by *38 degrees* - one can even see how the process of finding a workable plan of action played out on `reddit`, with plans that had a mixed response in comments being 'abandoned', and the 'doubling down' on approaches which bear fruit (chiefly boycotts in this case). It is perhaps remarkable that 'higher-order', almost strategic, 'decision-making' like this can emerge naturally through the voting-mediated interaction of a large number of equal participants - without recourse to the explicit opinion-monitoring and campaign testing/refinement carried out by *38 degrees*' leaders. *MoveOn* campaigns have been described as "not organising without organisations but organising with different organisations" (Karpf, 2012). In `reddit`'s case much of the 'organisation' is supplied server-side, omnipotent yet invisible aside from the voting arrows which accompany every item of content - and users exist in free-floating pseudonymity.

The fact that `reddit` is fulfilling the dual roles of 'news broadcaster' and 'collective action platform' concurrently may also be a significant advantage for the collective action endeavours which originate there - akin to having an 'in-house' magazine with a large readership. The capacity `reddit` has to 'set its own agenda' in terms of 'news' content likely extends to collective action, and therefore it may not be restricted to action on topics which are currently 'in the news' (the 'Rally to Restore Sanity and/or Fear' is another good example of this) as *38 degrees* restricts itself.

To re-visit the framing of 'Common Pool Resources' (Ostrom, 1990) - `reddit` has during both case studies demonstrated the capacity to deploy its CPR (attention) to a specific end (by promoting a particular perspective on `Wikileaks` and serving as a 'counter-measure' for actions against it, then in raising awareness of SOPA and organising collective action against it). In the SOPA case `reddit` went beyond consuming its own CPR and actively sought to expand this resource by deliberately attempting to garner attention from outside its own domain - both the GoDaddy boycott and the blackout were successful in doing this and generated a considerable level of coverage from both older and newer media elites.

The final point to take from this case study is that it offers further evidence that to fully understand `reddit` it is sometimes necessary to conceive of it as a collective social entity. In one sense this is a similarity between `reddit` and a conventional news organisation. A newspaper will often be associated with a particular stance on an issue, person or political party and the case studies recounted here demonstrate that `reddit` also has these characteristics. The difference is that this sense of identity comes about in a fundamentally different way - on `reddit` this occurs in a botton-up fashion but in a newspaper it is a top-down process (i.e. the newspaper's owner or high-level management determine the paper's stance while journalists adopt this decision as their own for the purpose of their writing). `Reddit`'s coverage may in fact have a more personal or humanistic tone; it appears that it is possible to 'piss off `reddit`' (GoDaddy and Lamar Smith both appear to have achieved this), and in another example `reddit`'s coverage of Al Franken in relation to SOPA belied a sense of betrayal - with many comments expressing disappoinement that a politician who redditors thought was 'on their side' would back such a reprehensible piece of legislation.

This idea that it is sometimes necessary to conceive of `reddit` as a collective social entity is pursued further in section 7.4.

7.3 Does `reddit`'s up/down voting system put the 'best' content in the most prominent locations? An experimental approach.

The principal task of a Social News voting system is to sort content such that the best items are ranked most highly and as a result are most widely seen. Where previous sections have considered the nature of posts that are highly ranked and the social impact of their appearance on the Front page - this section asks *how well* `reddit` performs in one of its most fundamental tasks.

While it is possible to track the performance of posts on a Social News website and observe their voting trajectories - to determine whether the 'best' posts are reaching the Front page some external measure of post quality is required. An experiment has been conducted which aimed to provide

such a measure and address a number of other questions.

The basic principle behind this experiment was to take a small sample of posts which had been tracked on `reddit` and present these to participants who would then rate them - providing a measure of 'post quality' which could then be compared to the posts' voting performance when they were submitted to `reddit`. The analysis of post voting in Chapter 5 suggests that many posts to large subreddits are quickly 'discarded' because they do not progress to the Rising or Main pages, a 'collective decision' which is effectively being made by relatively small numbers of users - and Chapter 6 presented evidence that there are certain users who account for a disproportionate level of voting activity on these early stage pages. This experiment presents all of the sampled items to all of the participants, thereby 'levelling the playing field' in this respect. When all of these posts are presented to a large enough number of participants this should generate a more reliable measure of how `reddit`'s population of users appraise these posts - this measure can then be compared to the posts' voting performance on `reddit`. If there is a strong correlation between post scores and the ratings generated by this experiment then `reddit`'s voting system can be said to be performing well.

Reddit users are the ideal participants for such an experiment because it aims to determine whether the voting performance of each post on `reddit` matches the perception of post quality *among reddit's population of users*. If participants in this experiment were drawn from some other population then a discrepancy between post scores and experimental ratings could be due to a difference in opinions on what constitutes a 'good post' between the `reddit` population and whatever population participants were drawn from. For similar reasons, the experiment's instructions did not specify exactly what criteria participants should rate items on - instead asking that participants apply the same criteria they would when they vote on `reddit` posts.

The 'pics' subreddit was chosen as the source for items to be rated in this experiment. It was felt that pictures could be rated much more quickly than other stimulus types (like articles or videos) - allowing for a greater number of items to be presented to each participant. Also, pictures were deemed less likely to suffer from being 'dated' than other possible stimulus types (e.g. a news article might be deemed worthy of attention on the day it was written, but of much less interest two weeks later).

250 image posts were randomly selected from the data recorded for the 'pics' subreddit in January 2011. 50 images for each of five post types were selected, these post types being based on the voting response receievd by the posts as follows:

- Posts which reached the default Front page (leaf 1).

- Posts which reached the Main page (leaf 1) for the pics subreddit.

- Posts which received a positive score but did not reach the subreddit's Main page.

- Posts which received no votes.

- Posts which received a negative score.

The instructions screen for the experiment has been included in figure 7.10.

Figure 7.10: Showing the instructions screen for the content-rating experiment.

The procedure for an experimental trial was as follows. First the post's title appears, the participant clicks on the title to load the image it links to, and this image appears with a rating scale; the participant makes their rating and the next trial begins. The decision was taken to display the post's title before the image being linked to because some of these posts have a kind of joke-punchline structure; where the image being linked to is effectively the punchline of a joke. Items were presented to each participant in a randomised order. Figures 7.11 and 7.12 show a sample trial from the experiment.

Each participant in the experiment was shown the same 50 items (10 of each type) from a core set of 50, and asked to rate these on a scale from -5 to +5. Participants were instructed to give a negative rating if they would down-vote the post, a rating of 0 if they would not vote on it, and a positive rating if they would up-vote the post. The decision to employ an 11-point scale was taken because this allows for a measure of how strongly participants felt about each item. Given that the instructions specified that negative ratings were equivalent to down-votes and positive ratings

Figure 7.11: Showing a sample item from the 'false information' condition of the experiment, the participant clicks the title to proceed.

equivalent to up-votes, it is also possible to collapse the rating scores so that they are more easily mapped onto `reddit`'s up/down voting system. With enough participants contrasting these two versions of the rating measure may yield interesting results. The experiment concluded with a set of survey questions concerning the participants demographics and their use of `reddit`. After this each participant could choose to continue rating /r/pics posts (with these posts being drawn from an expanded set of items distinct from the core set of 50).

The main difficulty in designing this experiment was a lack of control over how many participants could be recruited. The research project lacked funds for the payment of participants or advertising of the experiment, and therefore submitting a post to `reddit` which linked to the experiment was the only viable option for recruiting suitable participants. Having observed prior to this user-generated surveys which appeared on the Front page and received upwards of 20,000 responses (see section 6.1) - the decision was taken to design the experiment in such a way that it would benefit from having many participants. If the experiment was to appear in a prominent location once it was submitted to `reddit` this would happen only once - and therefore the experiment had to be designed from the outset in such a way that it would benefit from a large number of participants in the eventuality that this is what occurred.

A random sample of just 10 posts to represent each of the post types is sub-optimal, but it was felt that one could not expect each participant to rate more than 50 items. One solution was to allow participants to rate items from the expanded set of 200 items not in the core experiment once they had completed the core experiment. The experiment was also designed in such a way that the set of 50 core items could be switched to one of four other sets manually once the first set of core items had been rated by enough participants that additional data would be of no benefit.

Figure 7.12: Showing a sample item from the experiment after the participant has clicked the title. Once the participant selects their rating and clicks 'vote' the next trial begins.

The experiment would also have a further objective beyond establishing an 'external' measure of post quality - to investigate whether the perception of `reddit` users is influenced by the aggregate scores presented alongside a post on `reddit`. To this end three experimental conditions were incorporated in a between-subjects design (an individual participant is allocated to a condition randomly and completes the whole experiment in this condition). The first of these is a baseline condition - no score information is provided. In the second condition the post is presented with its voting score as observed four hours after the post was submitted on `reddit` (items have been formatted so that they look like `reddit` posts and the score appears in the appropriate location). In the third condition the post is presented with false score information (the score information for posts was switched randomly between posts). The hypotheses related to these conditions were that:

- In the false information condition ratings would be different to the true information condition *in the direction of the 'mislead'* - where the presented score was lower than the item's actual score this is a 'negative mislead' and ratings are expected to be lower than in the true information condition, the converse is true for items which were presented with a score that was higher than their actual score. If this is the case it suggests that the judgments of `reddit` users may be influenced by the previous judgments of other users, encapsulated by the item's

score. Another way to put this is that `reddit`'s users have learned that there is a relationship between `reddit` scores and item quality.

- There might be a temporal aspect to this effect whereby the effect weakens as the experiment proceeds. This is hypothesised because participants in the 'false information' condition may realise at some point that the scores being presented in the experiment had no utility or meaning. If this occurred it would allow us to consider how quickly `reddit`'s users can 'unlearn' the relationship between post scores and item quality when the relationship ceases to exist.

In May 2011 a post was submitted to `reddit` which invited `reddit` users to participate in the experiment and linked to the experiment's start page on the researcher's own web server. The post was submitted to the /r/pics subreddit to achieve the closest possible match between the population the experiment's participants would be drawn from and the population who had voted on the items' posts when they were originally submitted to `reddit` (both populations are /r/pics subscribers). Other subreddits may have been more suitable to submit an experiment to (e.g. /r/reddit.com, /r/TheoryOfReddit), but individuals who saw the experiment through a different subreddit would include some proportion of users who did not subscribe to /r/pics and had no interest in the type of content /r/pics caters to - and would therefore be unsuitable participants.

Unfortunately the post which was submitted to `reddit` did not fare very well, only achieving a modest score of around 10 and appearing briefly on the Rising page but not the Main page. Over the course of the weeks which followed the post was re-submitted twice but on no occasion did the post reach the number of potential participants that was hoped for. On each occasion the previous post had to be deleted because one cannot submit a post to a specific subreddit if a post linking to that URL has already been submitted. The experiment was initially conceived of at a time when posts which sought to understand `reddit` in some way were not uncommon on the Front page (e.g. the user-generated surveys in section 6.1 - but by the time it could be designed and implemented in software this trend had disappeared from the Front page. Chadwick (2013) remarks upon the importance of timing in the Hybrid Media System, and in this case sub-optimal timing has worked against the possibility of the experiment being broadcast on `reddit`'s Front page.

In total only 65 participants completed the whole experiment; 33 in the baseline condition, 17 in the true information condition and 15 in the false information condition. The experiment had been configured to randomly assign 50% of participants to the baseline condition and 25% to each of the experimental conditions because in the eventuality of a low number of participants it was felt that having a large enough number of participants completing the baseline condition should take priority.

A total number of 65 participants is much lower than was hoped for and means that there is too

little data to pursue all of the planned avenues of enquiry.

7.3.1 Do ratings obtained through the experiment reflect the performance of the items on `reddit`?

First the core question which the experiment sought to answer was addressed - is there a relationship between the items' voting performance as observed on `reddit` (recorded as one of five categorical levels) and the ratings obtained through this experiment? Here only the 33 participants who completed the baseline condition where the items were displayed without score information are considered. Given the low number of participants and ambiguity about what the ratings meant (beyond their being negative, neutral or positive), it was deemed preferable to treat these ratings as 'up/down scores'. The use of an 11-point scale to record ratings is related to the second aspect of the experiment, which was designed to look for a subtle influence on perception (more likely to be detected with a more sensitive measurement).

As this part of the experiment compares ratings of the items in the experiment to their voting performance on `reddit` it is preferable to transform the ratings such that they more closely match `reddit`'s voting system. The 11-point ratings were first converted to a 3-level ordinal variable with all negative ratings becoming -1 and all positive ratings becoming +1. For each item the 'up-votes' were summed and the 'down-votes' subtracted to produce a score. As there were 33 participants in the baseline condition of the experiment a post can have a maximum score of 33 (and minimum of -33).

This part of the experiment deals with the baseline (no score information presented) between-subjects condition of the experiment only. The only relevant independent variable here is the type of item presented and this has five levels (with 10 items from each level being presented). Order of presentation may also have influenced ratings (possibly through a fatigue effect) but this has been randomised between participants and therefore will not have a systematic effect on aggregated ratings.
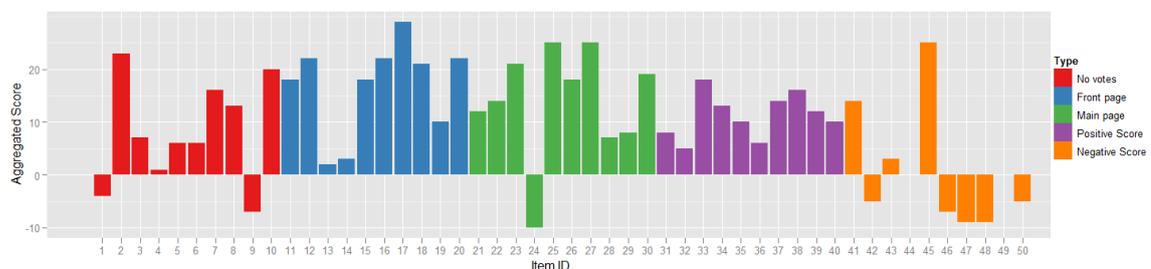
Figure 7.13: Showing Scores generated by the experiment (positive ratings less negative ratings) for the set of 50 core items. Each item has been coloured to reflect its type - determined by the voting response it received when submitted to /r/pics

Figure 7.13 shows the 'up-down' scores for the core set of 50 items. There does appear to be a trend whereby the experimental scores are related to the item's type (as defined by its performance on reddit), but for each type of post there are obvious exceptions to this trend. Posts which appeared on the Front or Main pages tended to receive the highest scores in the experiment, but there were two low-scoring Front page posts and one Main page post that received a negative score in the experiment. Of the ten posts which had a negative score on reddit five received a negative score in the experiment and two more had a score of zero - but one of the other posts from this category was among the highest-scoring items in the experiment. Similarly, some of the posts which received no votes on reddit received high scores in the experiment.

On the whole, figure 7.13 indicates that there is a relationship between how posts performed on reddit and how the same items performed in this experiment - but suggests that there is also a random component whereby good posts (as rated by participants in the experiment) can be ignored or down-voted and mediocre posts can appear in high-visibility locations. A linear model assuming a normal distribution was fitted to the items' 'up/down' scores with type as the explanatory variable, details of this model are presented in table 7.3. Posts which had a negative score have been chosen as the reference category, and while the coefficient for this category is not significantly different from zero the coefficients for all other levels are significantly different to this reference group. Furthermore, the coefficients for these levels follow the expected pattern: Front page > Main page > Positive Score > No Votes > Negative Score.

	Coefficient	Std. Error	p-value
Intercept (Negative Score reference group)	0.07	1.3	0.95
Front page	9.03	1.85	< 0.001
Main page	7.87	1.85	< 0.001
Positive Score	6.3	1.85	< 0.001
No Votes	5.73	1.85	< 0.01

Table 7.3: A linear model of Up/Down score in the experiment with post type as the explanatory variable - posts with a negative score are the reference category

There is evidence here that reddit's voting system as used by /r/pics subscribers is performing the task which it has been set, albeit not 'perfectly'. It is however not clear what standard the performance of reddit's voting system should be held to. Every month thousands of posts are submitted to this subreddit and an unorganised group of hundreds or thousands of individuals vote to filter and rank these posts - this part of the experiment shows that the collective judgments arrived at through the voting system stand up when a random sample of these items are presented to individuals from the population of /r/pics users several weeks later. Should this be regarded as a triumph - or should the fact that some 'good' items (that were highly rated by participants in

the experiment) did not progress far when they were submitted to `reddit` be counted as a serious flaw? The answer will depend on how well one *thinks* `reddit`'s voting system performs, because there are no equivalent studies on other websites that `reddit`'s performance might be compared to.

If one takes the analyses presented in Chapter 5 into account (particularly the low voting rates on early stage pages and the strong filtering effect for a large subreddit like /r/pics), `reddit`'s performance in this experiment is probably about as good as could be expected. If there was a strong monotonic difference in experimental ratings by post type, or if there was no evidence of a relationship between scores in the experiment and posts' performance on `reddit` - then we would be in a stronger position to state that `reddit`'s voting system works well or does not work at all.

One must also be careful not to place too much weight on the results of this experiment - the low number of posts of each type to be rated is a problem. There is a second issue also, Chapter 6 showed that users who vote on early-stage pages are in some ways different to the general user population and *may* hold more influence in the ranking process. The post advertising this experiment only appeared on the New and Rising pages of /r/pics, so all of the participants are likely to be users who are themselves active on early stage pages. If we assume that these users *do* have more influence in the ranking process then it is *possible* they are more likely to reproduce similar rankings in the experiment than users who are active elsewhere - but for this to occur there would need to be a systematic difference in opinions between the two groups of users. For /r/pics content it is difficult to imagine what this difference might be or how it would come about (perhaps 'expert' voters might tend to be younger or older?) - for a subreddit like /r/politics this seems like a more feasible possibility (e.g. users of a particular political persuasion might 'patrol' the New and Rising pages and down-vote content which goes against their own attitudes).

7.3.2 Does the score which an item was presented with influence users' perception of its quality?

With regard to the experiment's secondary aims the low volume of data is a more serious concern - only 17 and 15 participants respectively completed the two experimental conditions ('true information' and 'false information'). The effect of experimental condition on ratings was expected to be relatively small, and so a much larger number of participants would have been required to reliably determine if it was present. Here the 11-point scale is useful because the anticipated effect is likely to be weak, it is not necessarily expected that the presence of score information will change someone's opinion on a post from 'bad' to 'good' but perhaps more likely from 'bad' to 'terrible' or from 'good' to 'great'.

In analysing this part of the experiment only data from the two experimental conditions ('true

information' and 'false information') is considered. The mean rating of an item in the 'true information' condition is taken as representing its 'true rating'. Responses from the 'false information' condition are then compared to this 'true rating' to determine if a) these are different to the 'true rating' and b) whether the difference is in the direction of the 'mislead' effect. Where an item was presented, in the 'false information' condition, with a score that was greater than its post actually achieved on `reddit` (recorded four hours after submission) - this is a 'positive mislead'. When the presented score is lower than the post's score as observed on `reddit` after four hours, this is a 'negative mislead'. This part of the experiment does not consider an item's 'type', instead controlling for the fact that items were of variable quality by comparing ratings to the mean for that item from the 'true information' condition.

To address the question of whether presented score information influenced participants perception of quality a data-set consisting of 750 rows was analysed, each row representing a response from one of the participants in the 'false information' condition. The dependant variable is 'distance' - this was calculated by subtracting the mean rating for an item (taken from the 'true information' condition) from the rating which the participant gave this item. The distance variable therefore represents whether an individual participant appraised the item as better (positive distance measure) or worse (negative distance measure) than the average rating from the 'true information' condition. For example, a distance of 2 for a trial would mean that the participant rated the item 2 points higher than that same item had been rated on average by participants in the 'true information' condition. The independant variable is the direction of the 'mislead' ($+1$ where the presented false score was greater than the true score, -1 where the false score was less than the true score). It is expected that 'mislead' will modulate the direction of the distance scores, with a positive 'mislead' leading to positive distance measures and a negative 'mislead' leading to negative distance measures.

	Coefficient	Std. Error	p-value
Intercept	0.117	0.088	0.184
Mislead	0.235	0.089	< 0.01

Table 7.4: A linear model of 'distance' with 'mislead' as the explanatory variable - fitted to 750 responses from participants in the 'false information' condition.

A linear regression model assuming a normal distribution was fitted to this data-set (see Table 7.4). This model shows an effect of the 'mislead' variable which is significant at the 0.01 level. The intercept for this model relates to one item which was randomly assigned score information that matched its true score (and therefore has no 'mislead' associated with it). The coefficient for the 'mislead' variable suggests that ratings recorded in the 'false information' condition were different to the mean rating from the 'true information' condition by an average of 0.235 points *in the*

direction of the 'mislead' effect. This provides some evidence that the scores presented alongside `reddit` posts influence users' perception of their quality. However, one must bear in mind the low number of participants who completed the experiment.

A secondary hypothesis related to the presentation of false information was that the effect of the misleading scores would become weaker as the participant completed the experiment's 50 trials. Some of the item/score pairings are implausible (i.e. very poor images are presented as having high scores) and it was anticipated that during the experiment participants in the 'false information' condition would realise that the information was false or uninformative and subsequently disregard it. As presentation order was randomised for each participant, this anticipated effect would cause problems for any analysis of ratings for specific items or item types which are drawn from the participants who completed the 'false information' condition.

Instead the focus is on the presentation order variable. For each trial the previously created 'distance' variable was multiplied by the 'mislead' variable (+1 where the presented false score was greater than the true score, -1 where the false score was less than the true score) to produce a 'uni-directional distance' variable - which reflected whether the distance from the mean was in the direction of the mislead effect (positive values) or in the opposite direction (negative values).

The uni-directional distance variable was then plotted against presentation order with a smoother based on the normal distribution, the resulting graph is displayed in Figure 7.14. This Figure suggests a small effect in the expected direction which is stable until around the 30th item is presented and then begins to decline, having reversed slighlty by the final trials of the experiment. The implication here is that `reddit`'s users could quite quickly 'un-learn' the association between scores and quality when it became apparant that the scores contained no useful information. The evidence for this effect is however weak, due to the small sample size the 95% confidence interval for the effect includes zero over most of the range of presentation order.

7.3.3 Experiment Summary

This experiment offers support for each of the hypotheses to varying degrees.

As expected, there is a relationship between ratings collected during the experiment and the performance of posts on `reddit`, and efforts have been made to quantify this relationship for experimental items. Two assertions from previous Chapters are also supported in this regard. 1) The difference in 'quality' (ratings in the experiment) between posts which appear on the Front page and the subreddit's Main page is much smaller than the difference in their `reddit` scores would suggest - this is taken as confirmation that it is `reddit`'s voting system which produces a highly skewed distribution of scores rather than any attributes of the items themselves. 2) There is a random component to this process whereby 'good' posts can receive low scores and 'bad' posts
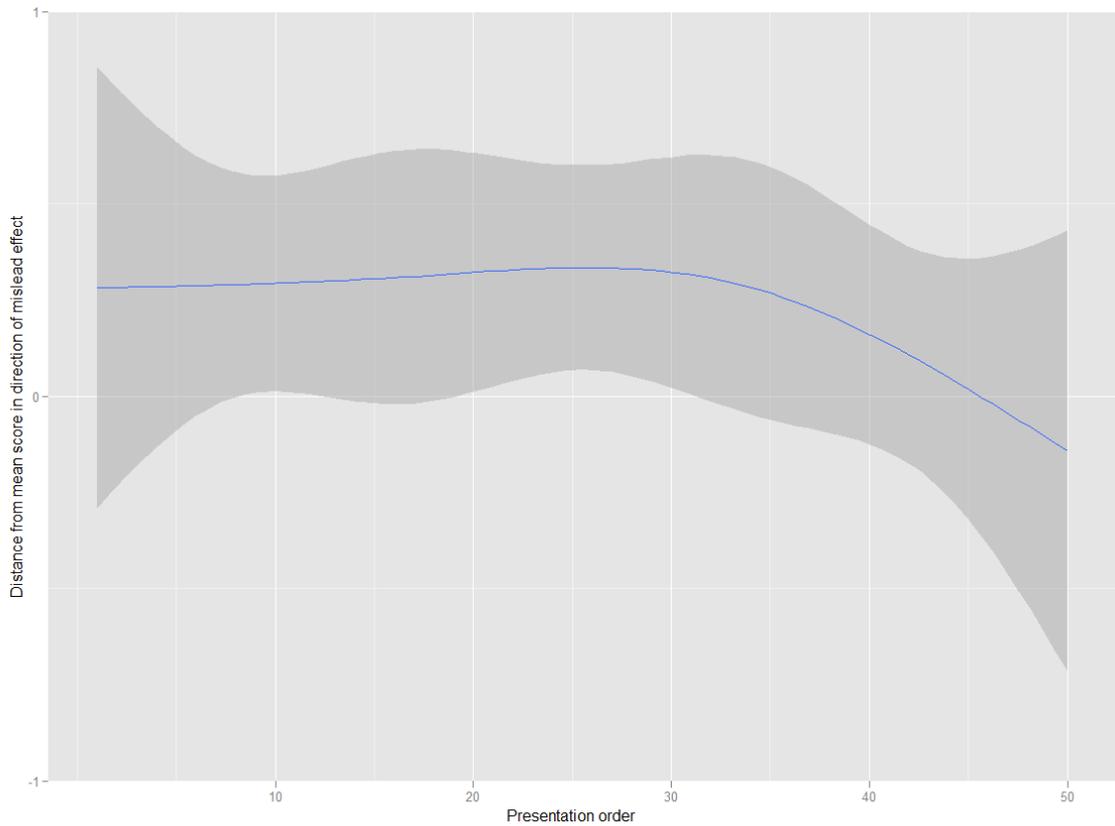
Figure 7.14: Showing the distance of ratings in the 'false information' condition from the mean in the 'true information' condition multiplied by the direction of the 'mislead' and displayed according to order of presentation.

can appear in high-visibility locations.

The second component of the experiment showed that the scores presented alongside items influenced participants' perception of their quality. This suggests that `reddit` users have learned that there is an association between the scores of `reddit` posts and their quality. Analysis of presentation order revealed that this effect had dissipated by the end of the experiment.

7.4 'Hive Mind' or 'Collective Consciousness'?

As established in Chapter 5, reddit's voting system involveas a number of positive feedback loops. Firstly, positive votes for a post are associated with increased visibility and in turn further votes - leading to the emergence of a small number of highly visible posts. These posts are displayed on `reddit`'s Front page for the whole community to see, and serve as shining examples of the kind of content which is successful on `reddit`. If a user wished to submit content which the `reddit` community would appreciate, a potentially fruitful strategy would be to mimic the types of post which they see on the Front page. It would seem that many users adopt such a strategy; it is quite

common for a subreddit to be flooded with new submissions which emulate a post that is currently performing well in that subreddit.

Redditors' penchant for up-voting certain types of post is often discussed by users. The tendency for users to submit posts of a type which have previously been received well is known in the community as 'circlejerking' and a subreddit exists for the purpose of highlighting and satirising these trends (the 'circlejerk' subreddit). This, and the perception that posts which go against certain commonly held beliefs or mores are always down-voted, are two of the main characteristics of what `reddit` users often refer to as the '`reddit` hive mind'.

In section 7.1 I argued that the evidence from the `Wikileaks` case study goes against this second component of the 'hive mind' - namely that posts which disagree with the 'hive mind' are always down-voted to oblivion. In this section I would like to introduce another term for this phenomenon which is more neutral than the overwhelmingly negative 'hive mind', and to look for some evidence that feedback loops which exist on `reddit` can influence the community's development in a positive way (beyond that which has been presented in section 7.2 relating to collective action against SOPA). The term I would apply to this phenomenon is 'Collective Consciousness'; originally used by Durkheim to refer to the shared beliefs and moral attitudes which operate as a unifying force within society. In the context of `reddit` I use this term to refer to complexity emerging from the collective behavior of hundreds of thousands of geographically disparate individuals as facilitated by `reddit`'s voting system.

Section 7.1 made reference to posts which I classified as 'meta-reddit'; text posts where a `reddit` user highlights or criticises some aspect of `reddit`'s behaviour. In the context of the `Wikileaks` case study these are predominantly posts which criticise `reddit`'s coverage of the story - the most common criticisms being that there are too many prominent posts about the `Wikileaks` story, or that posts about the `Wikileaks` story are too one-sided and out of touch with the general population. Indeed it is quite common to see a number of posts on `reddit`'s Front page about a particular topic or following a particular trend, alongside a post criticising the current popularity of said topic/trend.

When such a post appears on `reddit`'s Front page it has effectively been endorsed by the community (through the voting system) and is broadcast to all of the community's members who visit `reddit` while the post is visible on the Front page. Do these successful 'meta' posts influence the collective behaviour of the website's users? If they do, then this would constitute evidence that the voting system allows the community to self-regulate at a high level - or that `reddit` plays host to a rapidly developing 'collective consciousness'. If they do not, this suggests that the 'hive mind' suffers from cognitive dissonance or poor discipline/memory.

The nature of these 'meta' posts is such that it is usually difficult to gauge whether they have had

an effect on `reddit`'s behaviour as a social entity. The following section presents one example of a 'meta post' suggestion which is more amenable to empirical scrutiny.

7.4.1 Don't link to Fox News!

Reddit's users often criticise mass media news outlets; and it is not uncommon for such criticisms to reach high-visibility areas of the website (see section 7.1). Fox News is regularly the subject of such criticism and posts will often link to the Fox News website for the purpose of highlighting a particular flaw or faux pas. On 14th April 2011 two 'meta' posts concerning Fox News (redd.it/gpwzp and redd.it/gpy9s, referred to in this section as 'the meta posts') were submitted to the politics subreddit and were up-voted to the top of this subreddit and the default Front page. Both of these posts were submitted in response to a post which was then at the top of the politics subreddit (redd.it/gplqq); and which linked directly to the Fox News website for the purpose of criticising its content. The message of these two 'meta' posts was similar; that `reddit` users should stop linking directly to Fox News because incoming traffic from `reddit` was actually helping Fox by increasing their usage metrics and advertising revenue. These posts received scores of 1,500 and 800 respectively, indicating that there were many `reddit` users who agreed with the point they made. These posts also appeared at the top of the politics subreddit and on the default Front page for a combined total of 19 hours, meaning that many users of `reddit` (who subscribe to the politics subreddit) will have seen them during this time.

The comments on these 'meta' posts are quite mixed however. For post redd.it/gpwzp the highest-scoring comment by some margin suggests that people take screen captures and host these on `Imgur` if they want to submit something related to Fox News, but other high-scoring comments are either off-topic or take exception to being told how they should use `reddit`. For post redd.it/gpy9s the top comment is from someone who works in 'online ad sales', saying that Fox News doesn't make much money from `reddit` traffic, some other high-scoring comments adopt an argumentative tone towards the OP, which is perhaps to be expected as the post's title addresses `reddit` users as 'you fucking morons'.

The nature of this particular suggestion ('stop linking to Fox News') is such that it is possible to investigate whether it has influenced the behaviour of `reddit` users. There are two ways in which the users of /r/politics could respond to these suggestions. Firstly, individual users could stop submitting posts which link directly to Fox News, as was suggested by the meta posts. Secondly, users who vote on /r/politics could down-vote any submissions they see which link to Fox News, achieving the same end as suggested by the meta posts (reddit would no longer channel substantial traffic to Fox News if links to Fox did not appear in high-visibility locations). Of these two possibilities a more negative voting response seems the more likely outcome. An individual user

could single-handedly counter-balance any decline in the submission rate of Fox posts - whereas it is through the voting system that the collective will of the users is expressed.

Did the users of /r/politics respond to the community-endorsed and broadcast suggestion that they should stop channeling traffic to Fox News websites? If so, in what manner did they respond?
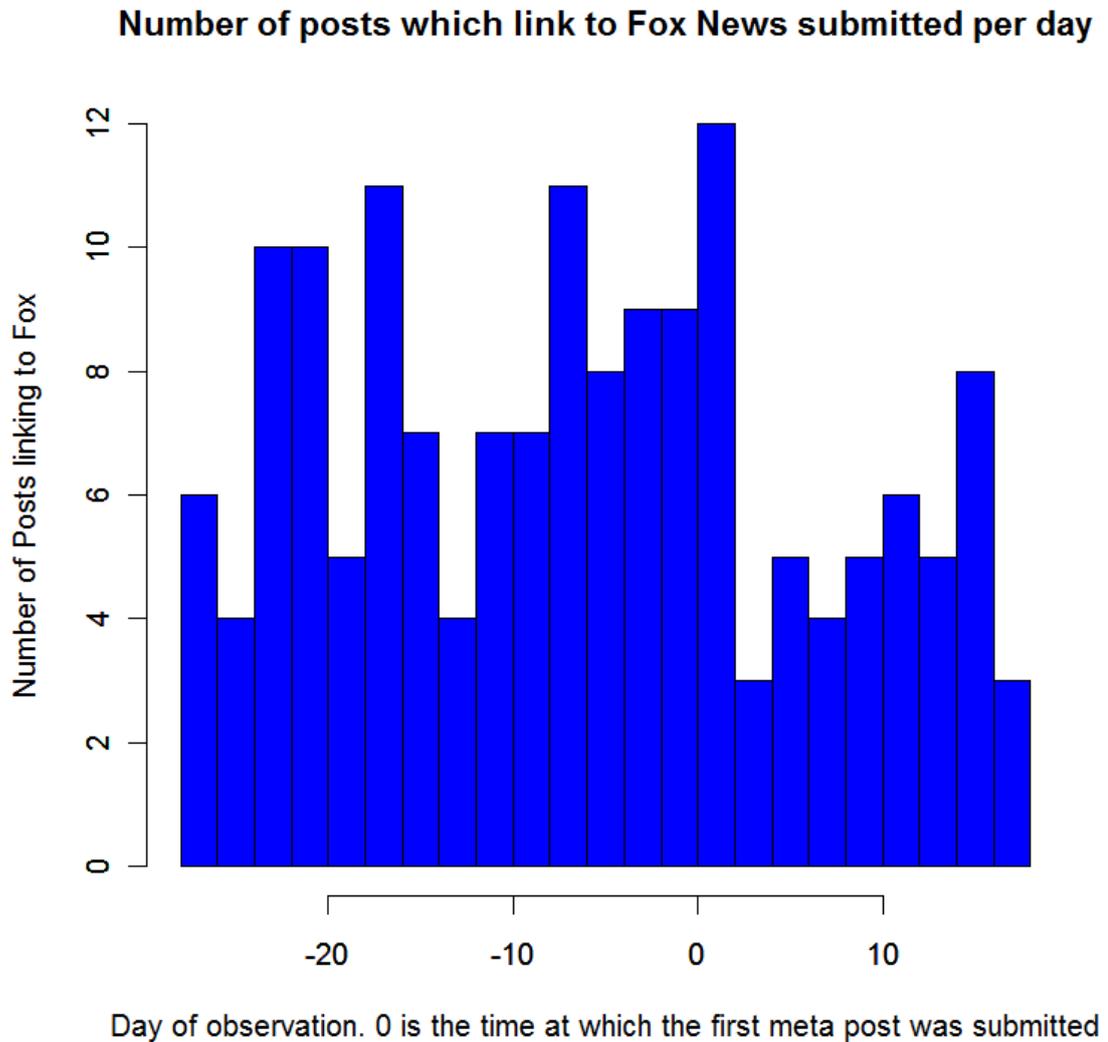
Figure 7.15: A histogram showing the number /r/politics posts submitted per day which linked to a Fox News domain. The 'meta' posts were submitted on day 0.

The politics subreddit was being monitored for 27 days before the 'meta' posts were submitted and 16 days afterwards. Figure 7.15 shows the number of Fox posts submitted to the politics subreddit per day; it can be seen that there is a slight drop in the number of Fox links submitted per day shortly after the meta posts were submitted. However it is certainly not the case that `reddit` users stopped submitting links to Fox News in response to the meta posts urging them to do so. In total there were 159 posts which linked to Fox News ('Fox Posts') in the 27 days before the meta posts were submitted and 16 days afterwards. An average of 4 Fox posts were submitted per day before

319

the 'meta' posts, dropping slightly to 3.2 per day afterwards. The immediate cessation of posts linking to Fox News which was called for did not occur.
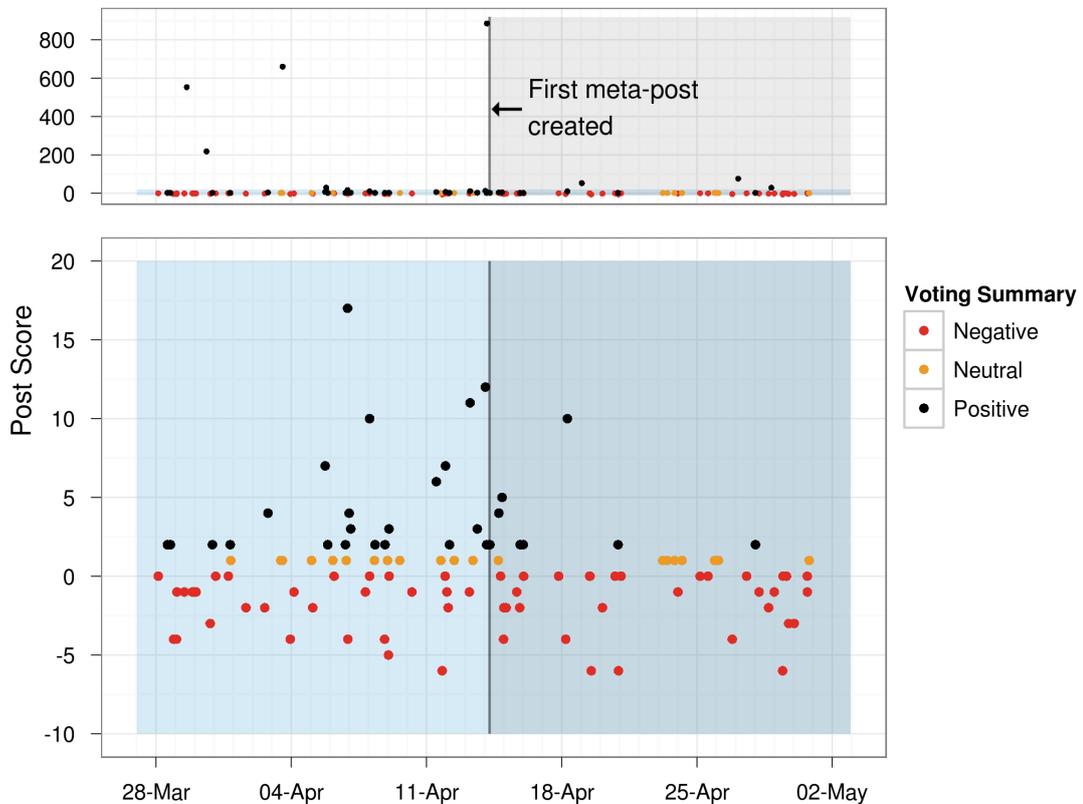
Figure 7.16: Showing the final scores of /r/politics posts which linked to a Fox News domain. The top pane shows the full range of scores, the bottom pane shows scores of between -10 and 20; most of the data lies within this range.

Figure 7.16 shows the scores achieved by all of the posts submitted during the observation period. In the 27 days before the meta posts there were four posts which linked to Fox News and achieved a score of greater than 200; the highest score achieved by a Fox post after the 'meta' posts was 76. In the bottom pane of this figure one can see a trend whereby the scores of Fox posts declined slightly after the meta posts were created. When we consider that a score of 1 has special significance on reddit (representing a neutral score or no votes) and look at whether the scores of these posts increased or decreased from this starting point, this constitutes more compelling evidence that the meta posts have influenced voting behaviour. There appears to be a clear drop in the number of Fox posts which achieved a positive score after the meta posts appeared on the Front page.

A logistic regression was performed to determine whether a Fox post's chances of receiving a positive score changed after the submission of the meta posts. The results obtained from this model are included in table 7.5. This model shows a significant effect of the meta posts on whether

	Coefficient	Std. Error	p-value
Intercept	-0.4132	0.1966	< 0.05
Submitted after Meta posts	-0.9978	0.4038	< 0.05

Table 7.5: Showing a binary logistic regression on whether Fox posts would receive a positive score.

subsequent Fox posts would receive a positive score. Fox posts submitted after the meta posts had 61% lower odds of receiving a positive score.

It appears that these two 'meta' posts - which came from ordinary `reddit` users with no special status but which were endorsed by the community through the voting system and consequently broadcast to that community, have indeed influenced the collective behaviour of users on the politics subreddit.

7.5 Chapter summary

Most of the posts submitted to `reddit` are seen by very few people, while a small minority can reach a large audience. `Reddit`'s Front page acts as a hub for the website's community of users. The posts which appear here go on to form the basis of discussions between users, and can be thought of as defining the website's identity (in the same manner that a newspaper can have an identity based on the articles it publishes). Section 7.3 showed that there is a relationship between a post's score on `reddit` and its 'quality' as rated by participants in an experiment, but that this relationship is less than 'perfect' and has considerable room for error. This experiment also suggested that `reddit` users' perception is influenced by the scores derived from the voting system, although the evidence for this is weak because of the low number of participants who completed experimental conditions.

The Front page showcases posts which have been endorsed by `reddit`'s users through the voting system, and these posts may serve as an example for users who wish to submit content that the community will appreciate - thus one popular Front page post may inspire a deluge of copycat submissions. This is one aspect of what `reddit`'s users refer to as the 'hive mind', the other being the marginalisation of posts or comments which go against certain norms or values which are held by a majority of `reddit` users.

Analysis of `reddit`'s coverage of the `Wikileaks` story suggests that the marginalisation of minority opinions may not be as severe as `reddit`'s users often speculate. The website was found to have a strong 'pro-`Wikileaks`' bias, but many of the posts which reached the Front page were neutral in tone, and several posts criticising `Wikileaks` or the release of the cables also managed to reach the Front page. Thus `reddit` does not appear to be an 'echo chamber' but a venue where readers

will be exposed to more than one perspective on an issue - an important finding in relation to the social implications of DM (Sunstein, 2002).

The `Wikileaks` case study also revealed a similarity in the level of coverage on `reddit` as compared to the mainstream news media - but some interesting differences in the nature of this coverage. `Reddit`'s users showed an eagerness to become actively involved in the story in addition to reporting it. This desire to participate manifested itself as posts which by their very presence helped the `Wikileaks` organisation (e.g. by helping people find their website at a new location), or which called on `reddit` users to engage in some form of collective action related to the story.

In the 'Wikileaks' case study `reddit`'s coverage was found to present an alternative narrative to the conventional news media, this coverage was in many cases explicitly pro-`Wikileaks` and incorporated attempts to further this cause. The SOPA case study highlighted a different aspect to `reddit`'s coverage, the capacity to raise awareness of a situation or 'break' a story, in addition to serving as a platform for the discussion and implementation of collective action.

Attention is a valuable resource (Huberman et al., 2004) - and there appears to be a growing awareness among `reddit` users that placement on `reddit`'s Front page is exactly such a valuable resource. As `reddit` has grown the nature of its interaction with the mainstream media has shifted; in its formative years this relationship was a one-way street but even before the `Wikileaks` and SOPA stories mainstream news outlets had begun to report on things which happened on or through `reddit` (e.g. (Townsend, 2010), (Rotham, 2010)).

One high-profile case where this occurred was the 'Rally to Restore Sanity'. On August 31st 2010 `reddit` user 'mrsammercer' submitted a post (redd.it/d7ntl) titled "I've had a vision and I can't shake it: Colbert needs to hold a satirical rally in DC." suggesting that Stephen Colbert (host of satirical news show 'The Colbert Report') should hold a rally as anathema to Glen Beck's 'Restoring Honour' rally which was at that point a major story on `reddit`. This post was up-voted to the Front page and (in a familiar sequence of events) a website and subreddit were quickly created to try and bring this idea to fruition. One of the proposals to encourage Stephen Colbert to host this rally was to donate money to a charity which he sat on the board of ('Donors Choose') - the resulting surge in donations when this proposal appeared on the Front page crashed the charity's donations system (Friedman, 2010b). Stephen Colbert acknowledged the donation drive on his show and began to hint that such a rally was in the planning stages. The 'Rally to Resore Sanity and/or Fear', hosted by Jon Stewart and Stephen Colbert, happened in late October 2010 and attracted an estimated 215,000 people (Montopoli, 2010). It should be noted however that the Rally's organisers claim it was in the planning stages before the `reddit` post and donation drive. Reddit's users certainly felt that they had been instrumental in bringing the rally about, and there may have been a resulting sense of confidence in users' collective 'power' through `reddit` which influenced coverage of the `Wikileaks` story that began shortly after. This kind of interplay between

'old' and 'new' media makes `reddit` a particularly interesting venue for the study of 'Convergence Culture' (Jenkins, 2006) - or the contemporary 'Hybrid Media System' (Chadwick, 2013).

Front page posts can act as a springboard for campaigns of collective action (for some further examples see redd.it/d7ntl and redd.it/hy24b) - or give a boost to some good-natured endeavour that has no marketing budget (e.g. redd.it/humy3, redd.it/elpko). The 'Humble Indie Bundle' (http://redd.it/elpko) is a good example of this. Several 'independently' developed computer games were packaged together and offered on a 'pay what you want' basis, with proceeds being divided between the developers and charities. When the bulk of this chapter was written there had been three humble indie bundles each earning over $1 million, and `reddit` had been the primary source of referrals for each by 'a decent margin' (source: private communication). By 2012 there have been many more humble bundles and this method of selling seems to be firmly established. In this period the 'Kickstarter' website has also become well established as a method of 'crowd-funding' projects and a number of similar endeavours have been launched (e.g. Steam's 'Greenlight', 'IndieGoGo'). It is not known whether `reddit` helped to generate publicity for Kickstarter in its early days, but several `reddit` posts linking to specific projects were observed in high-profile locations.

The attention received by Front page posts, coupled with the longer-term tracking of a user's contributions through karma scores, may also encourage some users to engage in pro-social behaviour where they might otherwise not have. Reddit has a history of supporting good causes and acknowledging acts of altruism (Voltier Creative, 2010, 2011). Some of `reddit`'s users also seem to actively police the website, there have been many popular posts in which a user has brought potential abuse of `reddit`'s voting system to light (e.g. redd.it/glgy8, redd.it/gjt26). While `reddit`'s voting system and Front page focus may give rise to a 'hive mind' it also seems to encourage pro-social behaviour among some users.

This kind of 'policing' is also relevant to the framing of `reddit` as a mechanism for the cultivation of a Common Pool Resource (CPR) (Ostrom, 1990) - users who seek to consume attention in an unscrupulous manner are identified and potentially punished by the down-voting of their submissions. The case studies in this chapter have shown that `reddit`'s users can also collectively deploy their CPR to a number of distinct ends - and some of the mechanisms through which the website's infrastructure mediates this have been discussed.

However, in more recent times such pro-social and altruistic activities occurring through `reddit` seem to be on the decline, and more attention has been paid to `reddit`'s darker side. The next Chapter details an examination of how `reddit` has changed over the course of this research.

The Front page feedback loop may also allow `reddit` to govern its own development in an interesting fashion. `Reddit` users often make suggestions for improving the website, either through some

modification of its software (e.g. redd.it/gnorc) or a change in the collective behaviour of its users (e.g. redd.it/fz3z1). When such suggestions appear in prominent locations they do so with the endorsement of users through the voting system and they are quickly accompanied by a comments page containing hundreds or thousands of comments - the scores of which offer a good insight into the community's feelings on the matter.

The 'Don't link to Fox News' example suggests that `reddit`'s users may respond to such suggestions. If this is the case it is quite remarkable. A member of this large group (the politics subreddit had at the time around 570,000 subscribers, although some proportion of these are likely inactive) with no special standing raised a suggestion and within 24 hours this had been discussed and appears to have been 'adopted' (to the extent that it has affected users' behaviour). All of this has been facilitated by the website's software infrastructure, no individuals with special status (moderators or administrators) and no special oversight was required. [4]

However, it must be noted again that this aspect of `reddit` has changed considerably since the bulk of this chapter was written. Many of the suggestions for how `reddit` should be changed (both socially and technically) were at one stage submitted to the /r/reddit.com subreddit, but in October 2011 this subreddit was disbanded. Such suggestions now have their own specialised subreddits (e.g. /r/ideasfortheadmins) where they will only be seen by users who have chosen to subscribe to these specialist subreddits. In addition, there seems to have been a shift towards ensuring that submissions to each subreddit are 'on topic', and a tendancy towards a greater number of more specialised subreddits. The net result of these trends in combination is that 'meta' posts of this nature most likely have a much smaller potential audience than they once did. Furthermore, these developments may have had the effect of fragmenting `reddit`'s user-base and fracturing what was once a relatively strong sense of community identity. If users are increasingly moving into a wider range of more specialised subreddits this could represent a problem from the public discourse - and `reddit` could lose some of its capacity to present users with a range of competing perspectives and 'information which they would not have sought out for themselves', in the process losing valuable 'shared experiences' (Sunstein, 2002).

Sections 8.4 and 8.5 consider these aspects of `reddit`'s development in greater detail.

The studies in this Chapter also allow us to situate `reddit` within the contemporary Hybrid Media System (Chadwick, 2013) - where it deploys an interesting mixture of older and newer media logics. Broadly speaking, `reddit` has performed in the roles of broadcaster and platform for collective action - and the capacity to broadcast is likely instrumental in `reddit`'s capacity to set

---

[4]By 2013 moderators have generally become more 'hands-on', a trend discussed in Chapter 8. In this environment a community-endorsed suggestion may lead to moderators adopting it as an official rule and listing it in the subreddit's 'side-bar'. In this scenario the moderators have greater power - they are autonomous with regard to the selection of which rules they list and enforce.

its onwn agenda for both news coverage and collective action - something which many bloggers and the *38 degrees* organisation do not even attempt (Chadwick, 2013). `Reddit`'s own 'organisational structure' is far removed from that of any other non-Social-News platform - and represents a step in the direction of "organisating without organisations" (Karpf, 2012).

Chapter 8

A longitudinal perspective on `reddit`

This Chapter considers Distributed Moderation from a longitudinal perspective. `Reddit` is particularly interesting in this regard because of its vaguely defined purpose ('The Front page of the Internet'), the breadth of content types which it caters to, and the fluidity with which it has developed. `Reddit`'s users vote to determine which posts are displayed on the website's Front page, and when a post is displayed on `reddit`'s Front page this is presumed to influence users' perception of the website's purpose or identity.

This Chapter addresses research questions relating to how `reddit` has changed during the time it has been under scrutiny. These are chiefly questions of scale - over the course of this research the number of people using `reddit` has increased by orders of magnitude. This affords an opportunity to consider `reddit`'s Distributed Moderation system as it naturally evolves, to see how the website and its user-base dealt with rapidly-increasing levels of activity. In addition to the increases in total number of users and level of activity - there appears to have been a change in the type of new users `reddit` attracted. `Reddit` was formerly seen as a niche technology-oriented website populated largely by technologically adept users - by 2013 it has become a much more 'mainstream' website. These two factors (increase in number of users and change in 'type' of users) must be taken into account in combination when we consider how `reddit` has developed over the course of the research.

In all of the previous Chapters we have probed `reddit` through data collected over specific time intervals (of usually no more than a few months). This is useful in understanding how `reddit` 'works' but will inevitably portray the website as static or stable. Informal observation suggests that `reddit` is in fact a highly dynamic website - the aim of this Chapter is to give an account of how the website has changed over the course of this research. This will begin with some aspects

which can be addressed through investigation of empirical data collected throughout the research. However, the nature of change on `reddit` has been such that 1) an increasing proportion of activity seems to be taking place away from the areas of the website which were part of the data collection regime, and 2) many of the changes concern type and quality of content.

For these reasons the latter part of this Chapter adopts a much less formal and data-driven approach. I felt that 1) much of what I've learned about `reddit` through informal observation was not captured by analyses conducted on empirical data, and 2) `Reddit` has changed most rapidly during the final stages of this research (mid-2012 onwards). Many of these changes do not fit neatly within the bounds of specific research questions, and in some cases there is no empirical evidence to present. However, I feel that excluding mention of these changes completely would be detrimental to the thesis, because the website I describe may not be recognisable to people who have become aware of `reddit` for the first time recently. This problem is compounded by the fact that `reddit` has become much more widely known while it has been changing most rapidly - there is almost certainly an association between these trends.

8.1   Analysing the Front page

Throughout this research `reddit`'s Front page has been observed and recorded at 30-minute intervals. As such, reliable information about the posts which have appeared on `reddit`'s Front page is available from June 2010 until March 2013 when data collection ceased; albeit with several gaps where my data-collection server was down. The first sections of this Chapter aim to provide an overview of how the content of `reddit`'s Front page has changed during this period.

In this time period, `reddit` has seen remarkable growth in its usage statistics. In June 2010 `reddit` had 8 million unique visitors and served 400 million page views (Reddit-Blog, 2010b). For January 2011 these figures had increased to 13.7 million unqiue users and 999 million page views (Reddit-Blog, 2011a), for May 2011 a further increase to 18.8 million unique users and 1.228 billion page views was reported (Reddit-Blog, 2011b), and in December 2011 the figures increased to 2 billion pageviews and 34.9 million unique visitors (Reddit-Blog, 2012). How has this rapid increase in user activity affected `reddit` and its Front page?

There are also a number of events which occured during this period which may have influenced the development of `reddit`'s Front page. As noted in section 3.3.1 August 2010 saw a major update to `Digg.com`, `reddit`'s main Social News competitor. This update to `Digg` was followed by a revolt of `Digg` users which resulted in `Digg`'s Front page being filled with links to `reddit` posts for around one week, during this time `Digg` saw a decrease in its usage statistics and `reddit` saw a corresponding increase, suggesting that a significant number of former `Digg` users became `reddit` users at around this time. `Reddit`'s Front page also saw many posts by or about the new users

from `Digg` at this time (e.g. redd.it/d5qne). Has this influx of `Digg` users affected `reddit`'s Front page in any discernable way?

At around this time `reddit` also began to be referenced more by conventional media outlets and in the comments threads of `Youtube` videos which were featured on `reddit`'s Front page, among other sources. This suggests a possible shift in how `reddit`'s new users have discovered `reddit`. `Reddit` was once seen as the domain of users who were well-educated and technologically savvy, it is likely that as `reddit` has grown over this period the proportion of its users who match this description has dwindled (this notion is supported by the removal during this period of subreddits related to programming and technology from the list of default subreddits which feed into the Front page).

Reddit's administrators have also changed the makeup of the default Front page directly by adding or removing subreddits from the list of default subreddits on several occasions. The most significant change made by `reddit`'s administrators took place on 18th October 2011, when the 'reddit.com' subreddit (initially the only 'subreddit' and subsequently a catch-all for posts which did not fit neatly into other subreddits) was disbanded and the list of default subreddits expanded from 10 to 20 (Reddit-Blog, 2011c). The criteria used to determine which subreddits would be included in the expanded set of default subreddits were primarily based on unique visitors. This expansion of the number of default subreddits has likely resulted in major changes to the dynamics of `reddit`'s Front page, and will be treated as something of a watershed moment in the analyses which follow.

There are six subreddits which have been constantly present in the list of default subreddits throughout the study period, and the analyses which follow will focus on these six. The 'reddit.com' subreddit was also a permanent fixture in the default subreddits list until it was closed; and will therefore be included in the analyses of the time before the default subreddits list was expanded to 20 subreddits. We repeat the analyses in two study periods seperated by the watershed moment when the default list of subreddits was expanded from 10 to 20 - the first running from 10th June 2010 until 18th October 2011, and the second from 18th October 2011 until 20th August 2012.

8.2 Reddit's Front page from 10th June 2010 until 18th October 2011

The known gaps in recording of `reddit`'s Front page are as follows: 20th July 2010, 28th December 2010 until 19th January 2011, 1st May 2011 until 16th May 2011, 2nd June 2011 until 12th June 2011, 8th August 2011 until 1st September 2011. There may be other short gaps where `reddit` was down or under heavy load. In creating the longitudinal data-sets care has been taken to exclude data which relates to the first observation after a period of no data collection. Many of

the measures considered here relate to a post's first observation on the Front page - if this first observation occurred after a gap in data collection the posts may have been on the Front page for some time before they were observed there and this would distort the analyses. Therefore for observations to be considered valid they must have been recorded within a certain temporal window following the previous observation (20 minutes - 2 hours). Data recorded after a gap which is outside this range has been marked as problematic and excluded from analysis. Similarly, where a post was first observed at one of these problematic observation points all data from that post has been excluded from analysis.

As many of the measures of interest here are counts the decision was taken to use Poisson regressions to assess longitudinal trends on `reddit`. These have been augmented with splines and smoothers to provide estimates of the uncertainty associated with coefficient estimates and to allow the slopes to fluctuate in time (as opposed to straight regression which would be limited to describing a monotonic increase or decrease in time).

These sections look at broad trends relating to `reddit`'s Front page and therefore a graphical modelling approach has been adopted - this allows a large number of trends and variables to be considered in a manner which is quick to execute and easily interpreted. Natural (cubic) spline models were fitted to the data-sets - where the response variable is a count the model is of the Poisson family, in other cases normal linear regression models have been fitted. To determine the degrees of freedom afforded to the splines (in the figures selected for presentation here) we have considered whether the data shows a monotonic trend or fluctuations. In relationships characterised by a monotonic trend (e.g. voting rates steadily increasing over time) the splines have been afforded 3 degrees of freedom. For relationships which show fluctuation (e.g. prevalance of subreddits on the Front page) the splines have been afforded either 5 or 10 degrees of freedom, allowing these fluctuations to be displayed without the regression lines becoming overly convoluted.

Formally, the Poisson models can be written as:

$$
\begin{aligned}
Y_i &\sim \text{Poisson}(\mu_i) \\
P(Y_i = y) &= \frac{\mu^y}{y!}e^{-\mu} \\
log(\mu_i) &= \beta_0 + f(X_i)
\end{aligned}
$$

Where $f(\cdot)$ is a natural spline function (Venables et al., 1994; Green and Yandell, 1985) and X is observation time.

The Normal models can be written as:

$$Y_i \quad \sim \quad \text{Normal}(\mu_i, \sigma_\epsilon^2)$$
$$Y_i \quad = \quad \mu_i = \beta_0 + f(X_i) + \epsilon_i$$

Where $f(\cdot)$ is a natural spline function (Venables et al., 1994; Green and Yandell, 1985), X is observation time, and $\epsilon_i \sim N(0, \sigma_\epsilon^2)$

8.2.1 Which subreddits do posts appearing on the Front page come from?

The first step in these analyses was to look at `reddit`'s Front page in terms of the subreddits of posts which appeared on the Front page at each observation point. 19,400 observation points between 10th June 2010 and 18th October 2011 are included in figure 8.1; by numbering these sequentially the gaps in records are effectively ignored. This is appropriate here because the aim is to look at general trends over this period. In this case the response variable (number of posts from subreddit on Front page) is a count and therefore Poisson regressions have been selected for modelling the data.

It is clear from this figure that the number of slots on the Front page occupied by posts from the /r/funny subreddit has shown a considerable increase over the observation period; steadily increasing from a mean of 2 to just under 4. The pics subreddit has also increased its share of the Front page, though not by as great a margin. The number of slots occupied by posts from the worldnews and AskReddit subreddits has declined, although with some fluctuation. The politics subreddit initially saw an increase in its share of the Front page (co-inciding with the time when the `Wikileaks`' release of US diplomatic cables was a big story) but has subsequently declined, concluding with a lesser prominence than it began the observation period with. The prevalence of the gaming and `reddit.com` subreddits has remained steady througout the observation period.

What sets the pics and funny subreddits apart from the others which are considered here? Why have posts from these subreddits become more prevalent on the Front page? The answer most likely lies with the nature of content submitted to these subreddits. Posts to the pics and funny subreddits tend to be both easy to comprehend/assess and entertaining in nature. This could affect the prevalence of these subreddits on the Front page in two ways. Firstly, reddit's voting users are volunteers and it seems likely that `reddit`'s new users have found browsing and rating content in these subreddits more enjoyable or appealing than the content which is submitted to subreddits which are more serious or information-oriented. Secondly, posts to these subreddits tend to be short and easily digested; this likely results in a higher number of votes per man-hour spent browsing these subreddits than on subreddits which play host to longer or more difficult to
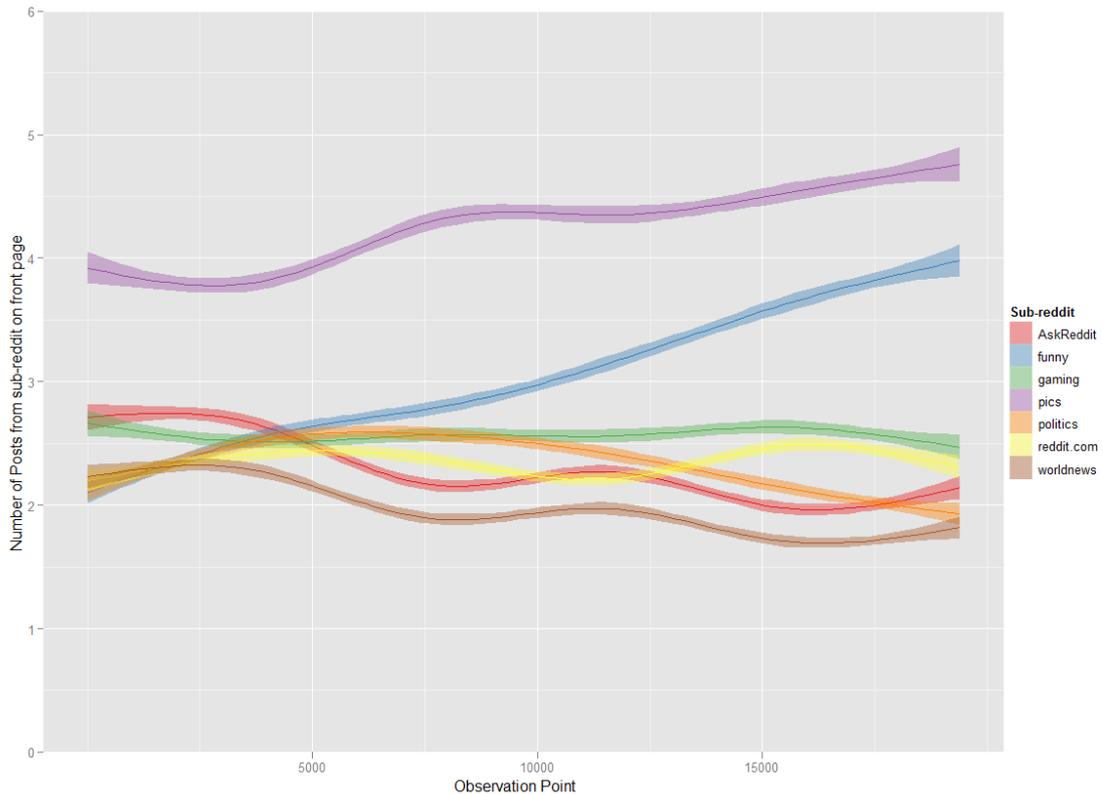
Figure 8.1: Mean number of Posts from each considered subreddit on the Front page per observation point, Poisson regression with natural cubic splines having 5 degrees of freedom, shaded area denotes 95% confidence interval - 10/06/10 - 18/10/11

assess posts (e.g. politics, worldnews).

Whatever the individual-level causes of this trend might be, the macro-level effect is clear. Considered as a news resource, the default `reddit.com` Front page has moved towards entertainment content and away from more conventional news subjects over the time period considered here. To relate this back to the newspaper analogy, the site has 'gone down-market'.

It is also possible to look at subreddit prevalence on an hourly basis; are there fluctuations in the makeup of `reddit`'s Front page throughout the 24-hour cycle? Poisson regressions with natural cubic splines having 10 degrees of freedom were applied to a Front page data-set covering the whole observation period. Figure 8.2 shows that for the majority of subreddits the number of Front page slots they occupy remains relatively stable throughout the day. However, there is a pronounced dip in the number of slots occupied by /r/pics (and to a lesser extent /r/funny) posts in the afternoon (GMT), corresponding to early morning in California.
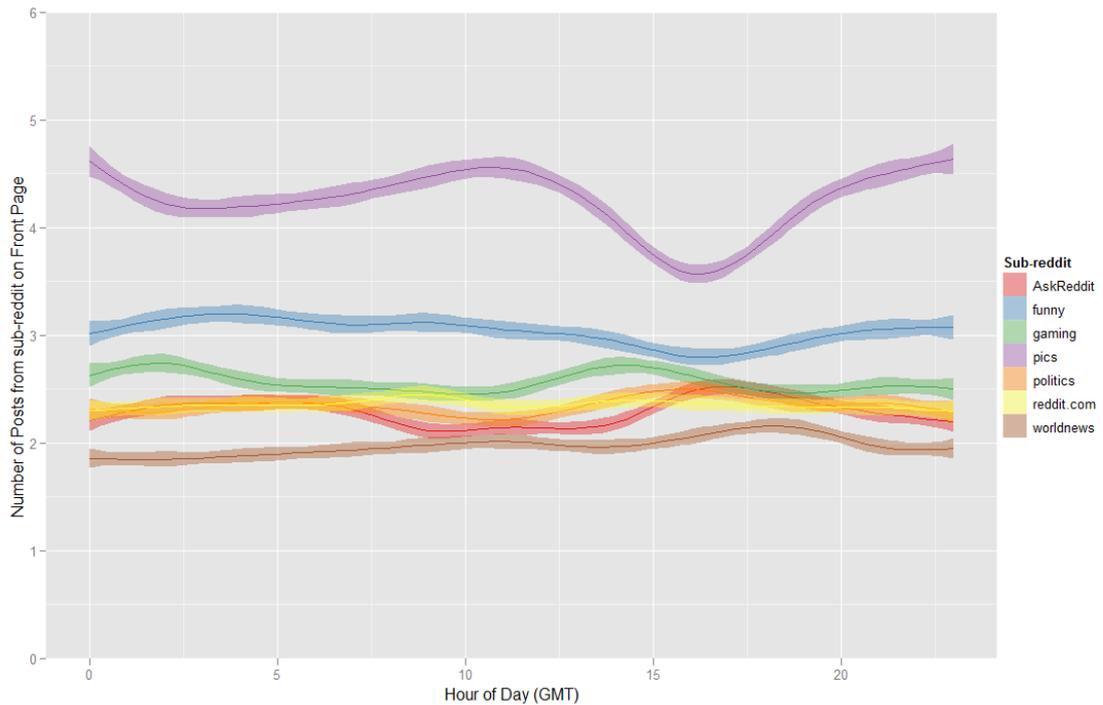
Figure 8.2: Number of Posts from each considered subreddit on the Front page per hour of the 24-hour cycle, Poisson regression with natural cubic splines having 10 degrees of freedom, shaded area denotes 95% confidence interval - 10/06/10 - 18/10/11

8.2.2 What does it take to reach the default Front page?

The rise in prevalence of the pics and funny subreddits on the Front page suggests that these subreddits have seen a greater increase in activity levels relative to the other subreddits which have decreased in prevalence or maintained a stable number of Front page slots. In this section we will consider the activity levels (i.e. score and number of comments) associated with posts the *first time they are observed on the Front page*; these measures reflect activity which has occurred before a post reached the default Front page. However, given that observations occur at 30-minute intervals a post will have actually been on the Front page for a short time before its first observation there, and as a result all measures will be slightly inflated. It is possible to apply this approach to all subreddits of interest, as it relies only on observations of the Front page.

The first set of results to be considered was generated by fitting normal regression models with splines having 3 degrees of freedom to the score at first observation of 38,136 posts which appeared on the Front page during the observation period. The exploratory variable in this model was the week of observation (ranging from 0 to 69, the week before the list of default subreddits was expanded to 20). The score at first observation gives a rough indication of the score a post required to compete with and win out over posts submitted at the same time to both the same subreddit and other subreddits. It can therefore be taken as a reflection of the voting activity level within

its subreddit.

Figure 8.3 suggests a substantial increase in the score required to reach the Front page for posts to all of the subreddits considered during the observation period. The score of a post at its first observation on the Front page has quadrupled between the start and end of the observation period for posts to all of the considered subreddits. The posts submitted to the pics and funny subreddits typically require the highest scores to reach the Front page; indicating that these subreddits may be host to a higher rate of activity on pages which 'feed into' the `reddit` Front page (i.e. the new and 'main' pages for the subreddit).
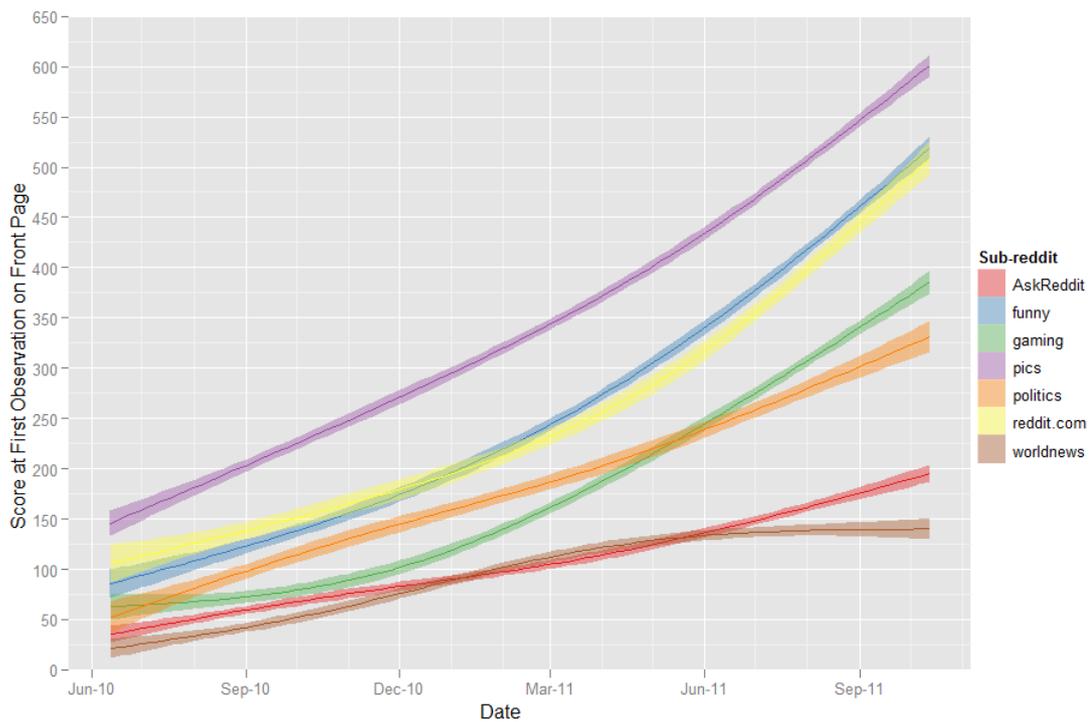
Figure 8.3: Normal regression splines with 3 degrees of freedom for the score of a post the first time it is observed on the Front page as a function of week, shaded area denotes 95% confidence interval - 10/06/10 - 18/10/11

One variable not accounted for in figure 8.3 is time to Front page appearance. There may be systematic differences in how long it takes a post to the various subreddits to make its way unto the Front page; and such differences would alter the meaning of first observed scores. It is possible to derive such measures from the data. An approximation of how long it took for a post to reach the Front page can be produced by comparing its creation time with the first time it was observed on the Front page. An approximation of a post's duration on the Front page can similarly be produced by comparing the timestamp of its first observation on the Front page with that of its last.

Figures 8.4 and 8.5 show normal regression splines on these measures with week number as the

explanatory variable. Confidence intervals have been excluded from these plots because they are wide and overlapping, making the regression lines difficult to see.

Figure 8.4 suggests that for the first 20 weeks of the observation period the mean time taken for posts to reach the Front page was declining for all subreddits. Beyond this point the subreddits diverge; for /r/politics, /r/worldnews and /r/AskReddit the time taken for posts to reach the Front page increases while for /r/pics, /r/funny and /r/reddit.com it continues to decrease. Figure 8.5 suggests that the mean time individual posts spend on the Front page has also been decreasing, with the pics and funny subreddits showing the most marked decrease in time spent on the Front page. Taken together, these figures suggest that the turnover rate of Front page content on `reddit` has been increasing; and that the subreddits which have increased their share of the Front page are the same subreddits with the highest rate of turnover.
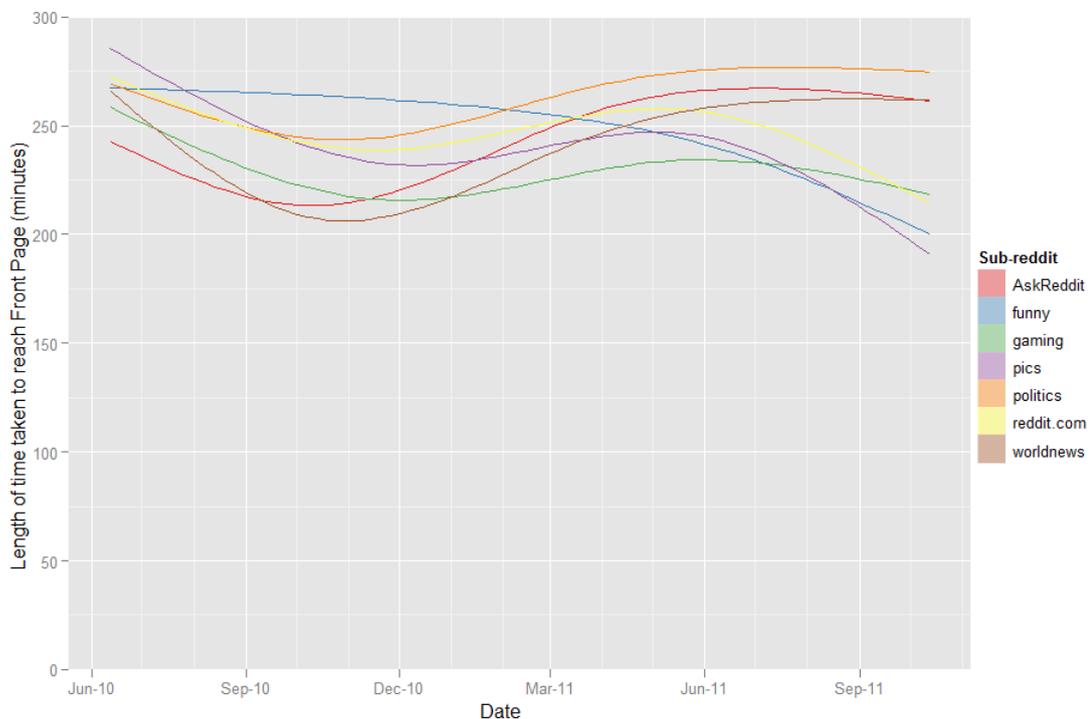
Figure 8.4: Normal regression splines with 3 degrees of freedom for the number of minutes taken for a post to reach the Front page - 10/06/10 - 18/10/11

What could be causing this increase in turnover and presence on the Front page for the pics and funny subreddits? To address this question we combine measures of activity with those of duration. For each post the score at first observation was divided by the number of hours taken to reach the Front page - producing a measure of score change per hour before reaching the Front page. A similar measure was created for score change per hour while on the Front page.

Figure 8.6 suggests that the hourly rate at which posts gain in score has increased most substantially on the pics and funny subreddits; the same two that have increased their prevalence on the Front
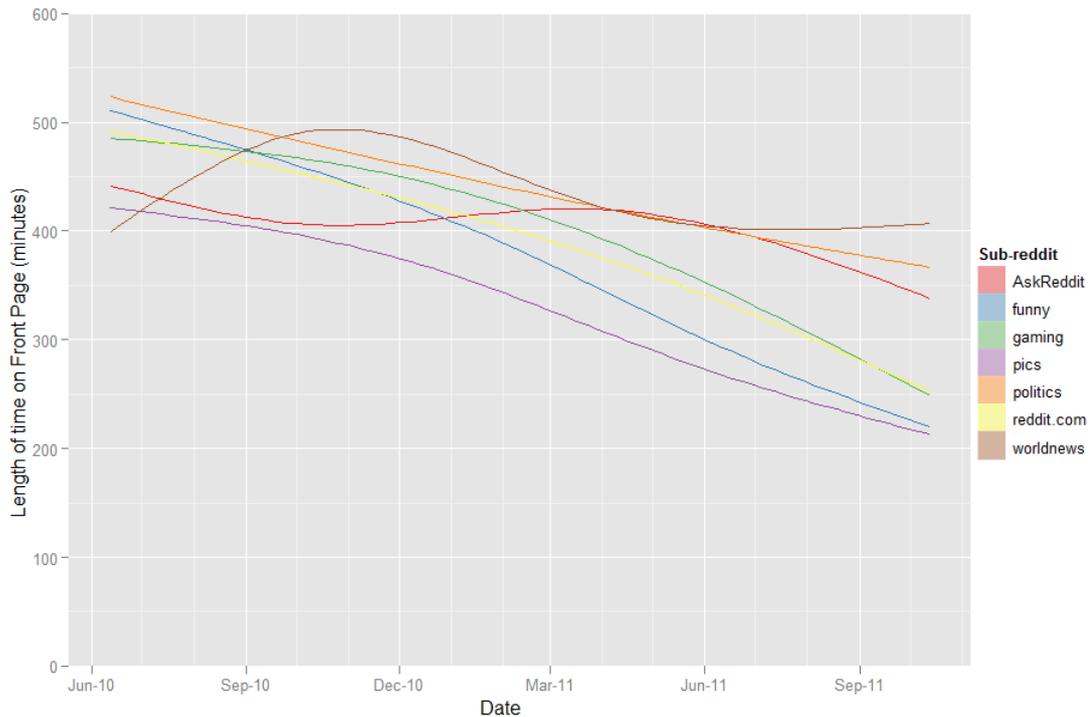
Figure 8.5: Normal regression splines with 3 degrees of freedom for the number of minutes spent on the Front page - 10/06/10 - 18/10/11

page and for which rates of turnover have increased the most. This is in contrast to the hourly change in score when posts are actually positioned on the Front page (figure 8.7); on this measure there appear to be much smaller differences between subreddits and these differences have remained relatively stable through time.

Taken together these two figures are quite revealing. The rate at which posts appearing on the Front page accumulate points has increased throughout the observation period; likely caused by a larger number of people browsing and voting on the Front page. This increase in Front page voting rate is comparable for all of the subreddits considered, suggesting a general increase in the level of Front page voting activity. The rate at which posts accumulate points before they reach the Front page reveals a much more striking disparity between the subreddits considered. It seems likely that the change in the Front page's makeup (in terms of the prevalence of posts from different subreddits) has been caused by a much larger increase in voting on the New and Main pages of certain subreddits than on others.

There is a peculiarity in these figures which must be addressed; towards the end of the observation period the voting rate for posts to certain subreddits before they reach the Front page appears to have eclipsed their voting rate while on the Front page. This has occurred for the pics, funny, reddit.com and gaming subreddits. This appears to be at odds with results reported in sections 5.1.3 and 5.1.3 which suggest a much higher voting rate on the Front page than for other pages.
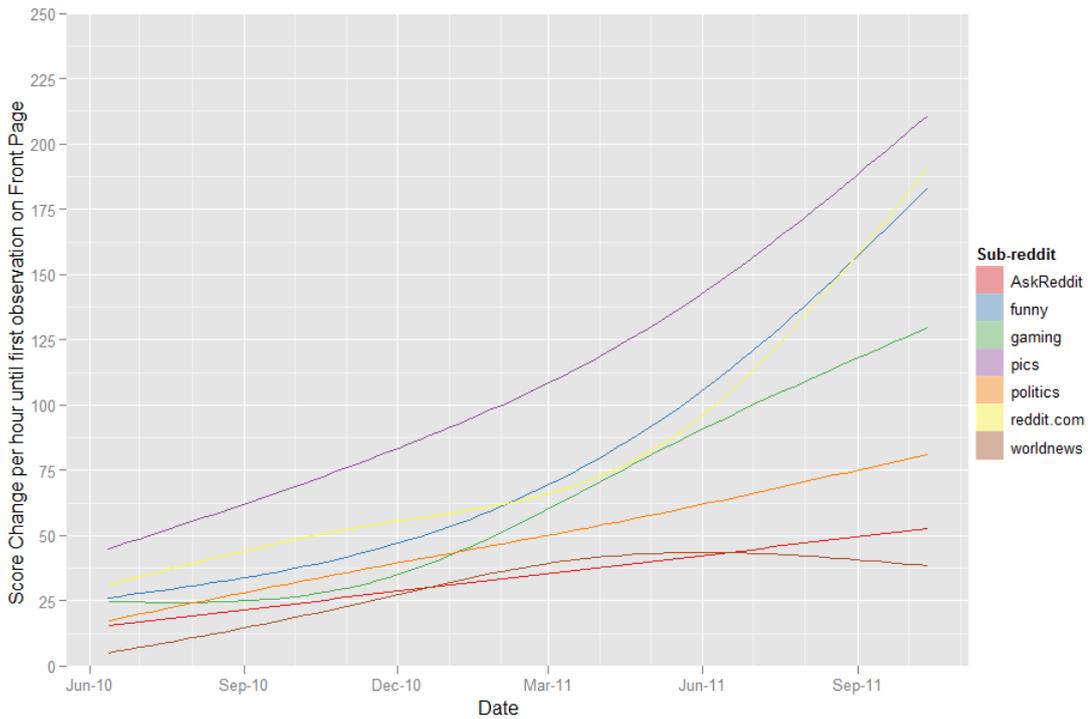
Figure 8.6: Normal regression splines with 3 degrees of freedom for the hourly change in score of posts before reaching the Front page - 10/06/10 - 18/10/11

Part of this discrepancy arises from a difference in the set of posts which the data modelled here relates to - here we are only considering the voting rates for posts *which ultimately were displayed on the Front page*, and one would expect these to have much higher voting rates than the posts which did not go on to appear on the Front page.

It is also possible that `reddit`'s users have become more familiar with how `reddit`'s voting system works, with votes cast on the Front page having less impact than those cast on pages which feed into the Front page. This awareness may be a function of increasing scores on the Front page; perhaps when one is voting on posts with a score of 2000 the futility of one's vote is more apparant than when the posts have a score of 500.

During the observation period trends have emerged which may also account for the shift in focus towards voting on pages which are not the Front page. Firstly, section 6.1.4 described messages which appeared at the top of the 'pics' and 'videos' subreddits in May 2011 urging users to vote on the 'new' page and presented some evidence that these messages had an effect.

Secondly, mid-way through the observation period posts began to appear on `reddit`'s Front page which thanked or in some other manner directly addressed or referenced users who vote on the New pages; often referring to these users as the 'knights of new'. A search of Front page records (post titles only) detects eleven such posts which have appeared on the Front page (although there are likely many more which were not captured by the regular expression search used). The first of
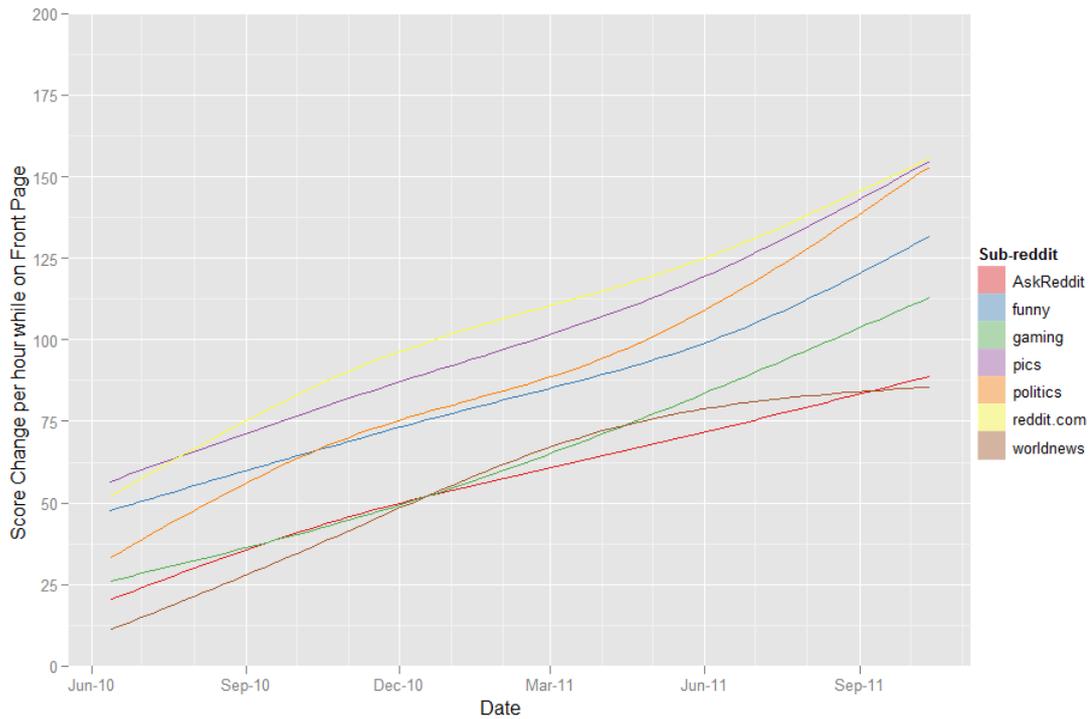
Figure 8.7: Normal regression splines with 3 degrees of freedom for the hourly change in score of posts while appearing on the Front page - 10/06/10 - 18/10/11

these posts appeared on the Front page on 3rd July 2011 (week 55 of the observation period) and was submitted to the pics subreddit (redd.it/ifm66). Of the eleven posts of this nature detected five were submitted to the pics subreddit, five to the `reddit.com` subreddit and one to the funny subreddit.

There are two ways of looking at these 'knights of new' meta posts; the frist is that they recognise and in some sense reward behaviour which is necessary to `reddit`'s successful operation (namely people browsing and voting on the New pages). In this sense they serve a purpose, encouraging behaviour which is pro-social; and the fact that these posts tended to appear on subreddits which also saw the greatest increase in voting activity off the Front page would tend to support this.

An alternative perspective on these posts is that they aim to reach the Front page (generating attention and karma points for the submitting user) by addressing (or pandering to) the people who will have the most influence over whether this occurs. For as long as `reddit` has been observed some users have addressed their posts to the `reddit` community (e.g. with titles beginning 'Hey `reddit`'), perhaps the 'knights of new' meme is a refinement of this approach.

As the number of active `reddit` users has increased so too has the voting rate on the Front page and pages which feed into the Front page. The pics and funny subreddits have seen a particularly large increase in the voting rate on pages which feed into the Front page, and this in turn appears to have increased the prevalence of these subreddits on the Front page.

The politics subreddit is interesting in this regard. The voting rate for politics posts on the Front page has increased more than for any other subreddit, suggesting that `reddit`'s users are interested in politics posts when they appear on the Front page. However, the voting rate for politics posts before they reach the Front page has increased much more slowly than for posts to the pics, funny, reddit.com and gaming subreddits. One interpretation of this trend is that users find good posts submitted to politics interesting, but are not so willing to wade through politics posts which have not yet been judged as interesting by other users (the length and density of posts submitted to /r/politics may be off-putting as compared to the ease with which one can consume and rate posts submitted to the pics and funny subreddits).

There are however other factors which must be taken into consideration. It is important to bear in mind that the analyses presented here consider only those posts which reached the Front page, a very small sub-set of the posts which are submitted to `reddit` and voted on. It is possible that voting rates on the politics subreddit have increased substantially but that the votes for individual posts are more balanced (i.e. more negative votes than on the pics subreddit leading to lower score increases). It is also possible that voting on the pics and funny subreddits converges on the ultimate 'winners' more quickly, perhaps because users on these subreddits have a more shared understanding of what makes a good or interesting post than users on the politics subreddit.

8.3 Reddit's Front page from 18th October 2011 until 20th August 2012

This section repeats some of the above analyses for a period of around ten months after the number of default subreddits was increased from 10 to 20 on 18th October 2011. Figures 8.8 and 8.9 show the presence of subreddits on the Front page for this period. These subreddits have been split across two figures because for this period there are 20 default subreddits - too many for a single figure. Also, confidence intervals around the lines have been removed because the lines for these subreddits are tightly grouped.

These figures show much more stability in the prevalence of subreddits on the Front page than was in evidence for the earlier period. All of the values are smaller now, as one would expect because there are now 20 subreddits competing for slots on the Front page as opposed to 10 previously. In the early stages of this period there is a drop in the presence of subreddits like /r/funny and /r/pics - suggesting that the full impact of a switch to 20 default subreddits was not felt instantly but after a short delay wherein more people began to be aware of and use the newly default subreddits.

Figures 8.10 and 8.11 show the mean hourly rate at which posts accumulated points *until their first observation on the Front page*. The dominant pattern is again a general increase in this rate,
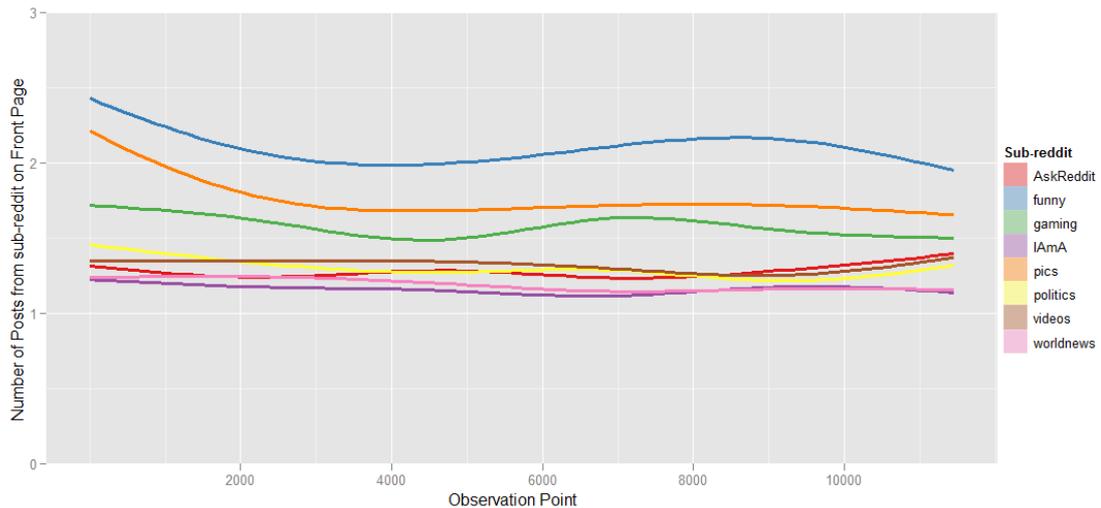
Figure 8.8: Poisson regression splines with 5 degrees of freedom for number of posts from each subreddit on the Front page per observation point - 18/10/11 - 20/08/12 - part 1

although the rate has been increasing more quickly for some subreddits than others. The greatest relative increases are for some of the newly default subreddits (e.g. rates for AdviceAnimals and Aww almost tripled), with established subreddits like /r/funny and /r/pics showing strong increases in their absolute rates.

In general these rates increased more slowly during this period than during the earlier period considered above. This could be due to a slowing of `reddit`'s growth rate or a decrease in the proportion of users who engage in voting on the 'early stage' pages of these subreddits.

Figures 8.12 and 8.13 show the rate at which posts accumulated points *while they appeared on the Front page*. Again the dominant trend is a general increase in this rate, although it is not increasing as quickly as it was during the earlier study period (before the switch to 20 default subreddits).

There is no evidence of a slowing in the growth of `reddit`'s usage statistics (the site continues to attract new users and increase the number of pageviews it serves), leaving two explanations for the slowing of the growth in voting rates. Either these new users are less likely to vote (perhaps viewing `reddit` without creating an account) or users are voting more in non-default subreddits.
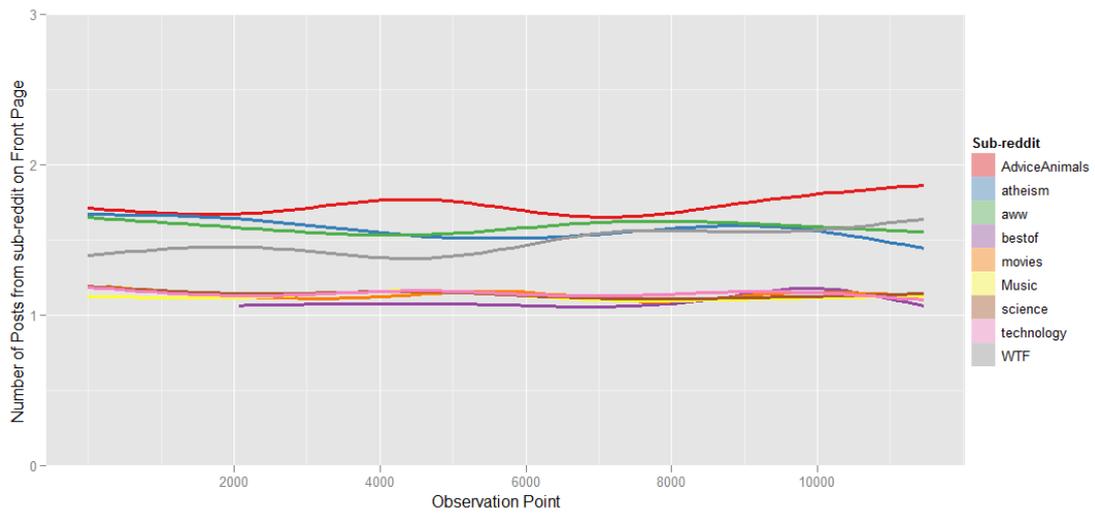
Figure 8.9: Poisson regression splines with 5 degrees of freedom for number of posts from each subreddit on the Front page per observation point - 18/10/11 - 20/08/12 - part 2
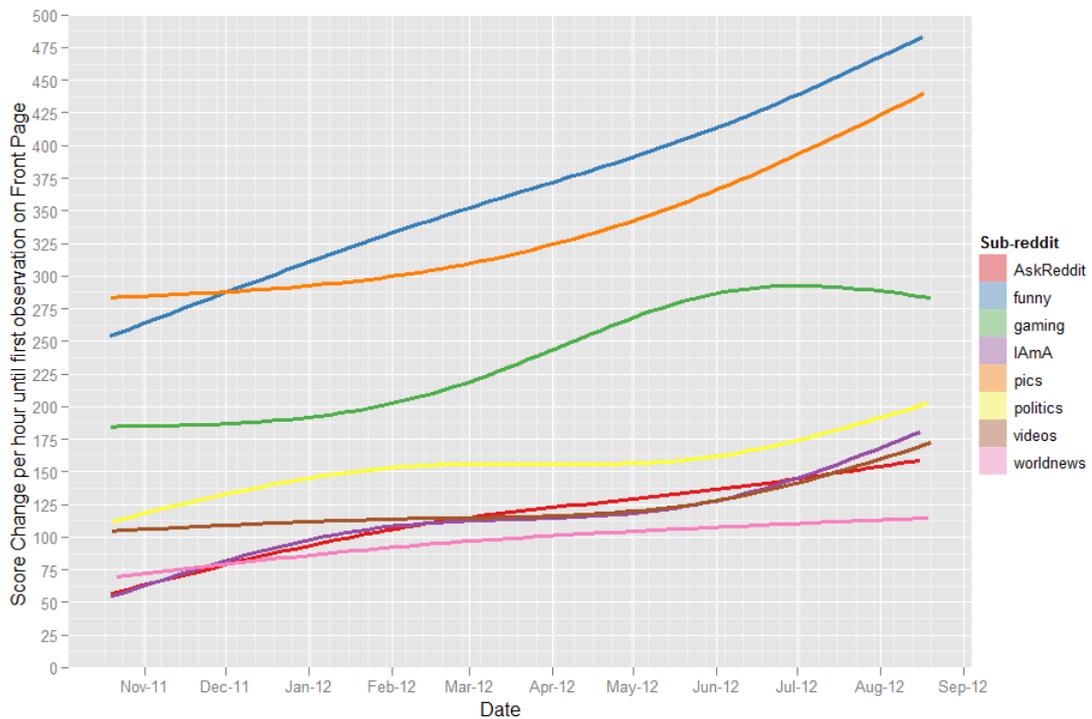

Figure 8.10: Normal regression splines with 3 degrees of freedom for the hourly change in score of posts before reaching the Front page - 18/10/11 - 20/08/12 - part 1
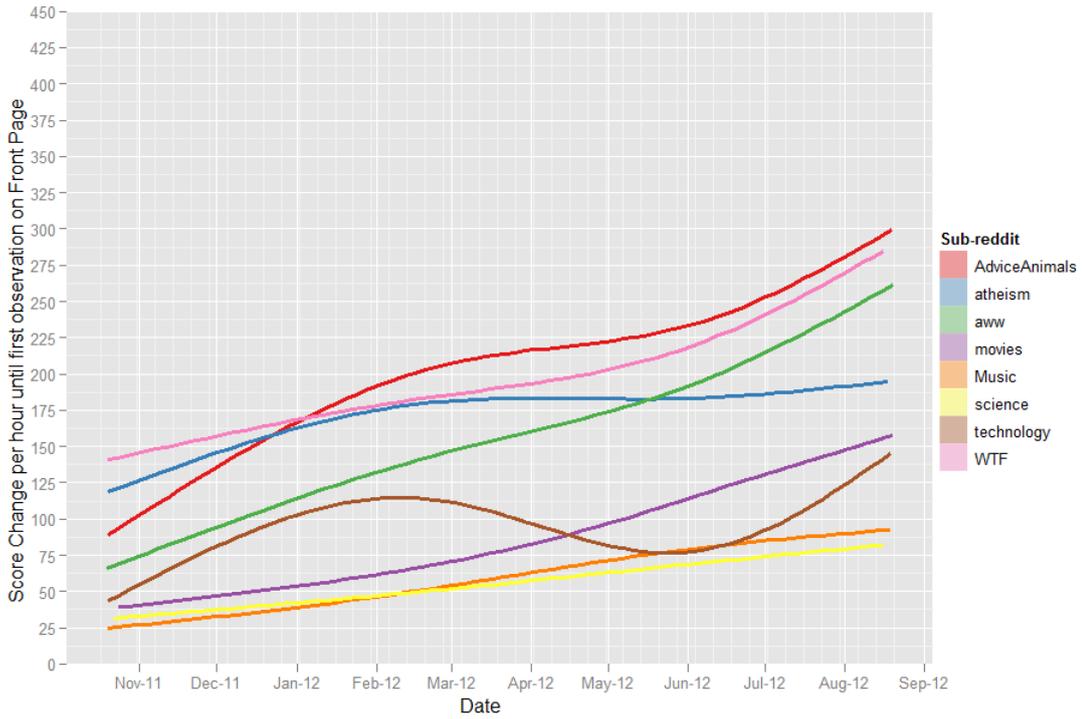
Figure 8.11: Normal regression splines with 3 degrees of freedom for the hourly change in score of posts before reaching the Front page - 18/10/11 - 20/08/12 - part 2
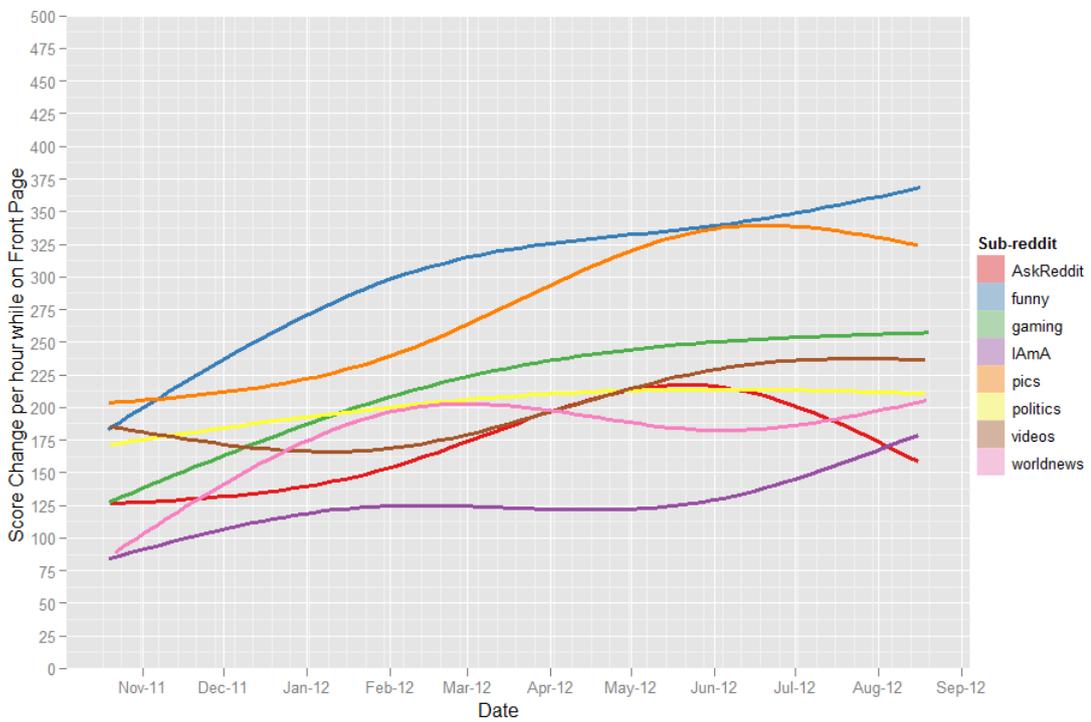
Figure 8.12: Normal regression splines with 3 degrees of freedom for the hourly change in score of posts while appearing on the Front page - 18/10/11 - 20/08/12 - part 1

341
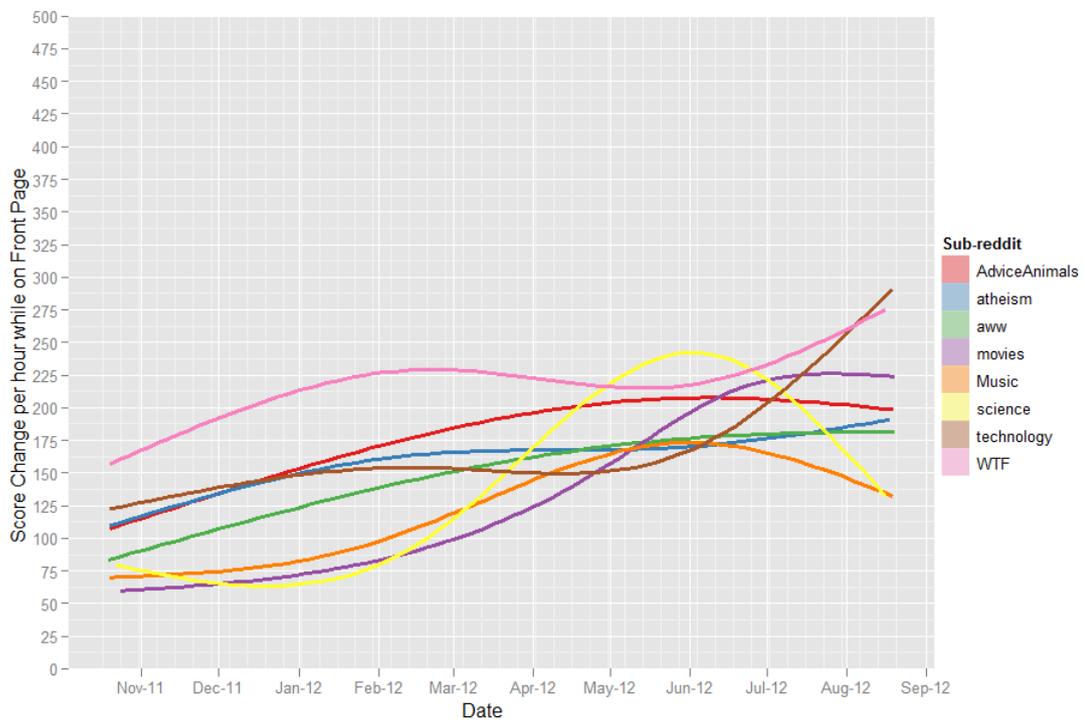
Figure 8.13: Normal regression splines with 3 degrees of freedom for the hourly change in score of posts while appearing on the Front page - 18/10/11 - 20/08/12 - part 2

8.4 A shift away from a shared Front page?

When casual observations of `reddit` began in 2008 comments often indicated that users thought of the 'reddit Front page' as a fixed entity - if something had appeared on 'the Front page' one `reddit` user could more or less expect that another frequent `reddit` user was also familiar with it. As the number of people using the website has increased rapidly over the course of this research one might expect that this increased user-base can support a much larger set of active subreddits and that users would come to diversify the makeup of their Front page through subreddit subscription choices. Casual observations have suggested an increase in comments where users talk about subscription choices. Popular comments now often refer to un-subscribing from /r/atheism and /r/politics, many even go so far as to suggest new users should unsubscribe from most of the default subreddits because the smaller subreddits tend to be better. The reasons for this are cited as a decline in the quality of content being up-voted to prominence through some of the default subreddits. For example, post redd.it/18gf7d asks "Is the tide now against /r/politics on `reddit`?". This issue of a perceived decline in quality on the default subreddits will be re-visited throughout the Chapter and particularly in section 8.10.

The question of whether redditors are increasingly tailoring their `reddit` experience through subreddit subscription choices is important from the perspective of the fragmentation of public discourse (Sunstein, 2002). As the number of active subreddits grows so too does the range of choices available to each user, and `reddit` can be customised to provide material along ever more specific criteria - raising the spectre of the Daily Me (see section 2.5.4). There are now a whole panoply of subreddits devoted to individual countries, states, cities, universities, sports, teams, games, scientific disciplines, etc. This raises a distinct possibility that users can largely filter out 'information which they would not have sought out in advance' from the feed they receive through `reddit`.

The degree of overlap between users' subscription portfolios is also important in that it will determine `reddit`'s capacity to provide 'shared experiences' for users - a valuable asset for a society and one which is diminished by excessive fragmentation of public discourse (Sunstein, 2002). Shared experiences are also likely to be critical to the development and maintenance of an over-arching `reddit` community and identity. The large default subreddits with the most users will naturally have the greatest capacity to produce shared experiences - and therefore an increase in users 'opting out' from these subreddits would be of particular significance. Widespread attention to large default subreddits was also seen as key to `reddit`'s ability to mobilise collective action in response to SOPA (section 7.2).

The present section aims to determine whether such a change in the subscription choices of users has occured. The obvious measure to consider here would be the number of subscribers to each subreddit. However, this is problematic because new users who create an account are automatically

subscribed to the default subreddits of the time and if these users do not continue to use their accounts this inflates the number of subscribers to default subreddits - such that the number of subscribers is not an accurate reflection of the number of people who actually use the subreddit. It remains very easy to create a `reddit` account and one suspects that many accounts are created and rarely used - or are not deleted once they cease to be used. In section 6.1.7 a survey was conducted in which around half of the respondents had created multiple `reddit` accounts.

This issue was illuminated in 2012 when `reddit` began showing the number of active users for subreddits in addition to their number of subscribers. For example, consulting the `stattit.com` website on 30th October 2012 - the 20 subreddits which are currently 'defaults' are the 20 subreddits with the most subscribers. In another table showing the most active subreddits in the preceding 24 hours by average number of active users - there are 6 non-default subreddits in the top 20 and the lowest-ranked default subreddit (/r/bestof) is ranked 51st by activity level. In section 5.1.14 subreddit subscription growth was considered (in particular see Figure 5.24), using the `redditmetrics.com` website, and figures suggested that in the 30 days up to 14th December 2012 around 38% of new users had un-subscribed from /r/atheism and 27% had un-subscribed from /r/politics. It is not clear how many of the 157,589 new users who remained subscribed to /r/funny and the other default subreddits are legitimate new users - it is likely that some proportion of these are 'throwaway' accounts or accounts that were not used for long.

An alternative measure of subreddit activity is required. Voting activity is another obvious choice but to know about voting activity levels for a specific subreddit is is necessary to collect data from the pages of that subreddit specifically. The decision was taken to first consider the /r/all page. The 'All' page can be thought of as an alternative to the Front page - it uses the same algorithm to determine post rank but it considers posts from every subreddit as opposed to only those subreddits a user has subscribed to (or the default set for a user not signed into an account). If the attention of `reddit`'s users is focused on the default subreddits one would expect the /r/all page to be similar in composition to the Front page (with most of the posts coming from default subreddits). If the attention (and voting activity) of `reddit`'s users has diversified away from the default subreddits one would expect to see an increase in the number of posts from non-default subreddits appearing on /r/all. This hypothesis is based on the premise that a popular post to a subreddit with higher voting activity levels will have a greater chance of appearing on the /r/all page than a popular post to a subreddit with lower voting activity levels.

This section considers the number of posts appearing on /r/all which come from default and non-default subreddits over the time period from 4th November 2010 until 18th October 2011. During this time period only 25 posts were being collected from the Front and All pages so analyses will be limited to Leaf 1 of these pages. These analyses also concern observations rather than posts - if a post is observed on the All page 10 times it will contribute a value of 10 because it occupied a slot

on that page at 10 observation points. During the time period `reddit`'s administrators sometimes tinkered with the set of default subreddits and it was not uncommon for a subreddit to be added to the default list but removed shortly afterwards. For the purposes of these analyses subreddits which had posts observed on the Front page fewer than 500 times (e.g. /r/ffffffuuuuuuuuuuuuu and /r/trees) are counted as non-default subreddits. Other subreddits which were default subreddits for some but not all of the observation period (e.g. /r/science, /r/WTF) are counted as default subreddits. These choices will lead to a slightly conservative figure for the number of posts appearing on /r/all from non-default subreddits.

The subreddits counted as defaults for the purposes of this analysis are therefore: announcements, AskReddit, blog, funny, gaming, IAmA, pics, politics, programming, reddit.com, science, today-ilearned, videos, worldnews and WTF. During the period there were 12,704 discrete observations, usually at 30-minute intervals. At each observation the top 25 posts on /r/all were recorded - resulting in 317,600 observations of posts appearing on the page. Posts to default subreddits account for 74% of these observations with posts to non-default subreddits accounting for the remaining 26%.

Figure 8.14: Showing the percentage of posts on /r/all (leaf 1 only) which come from non-default subreddits per week

Figure 8.14 shows the percentage of posts on the first leaf of the All page which came from non-default subreddits. There is no sign of an increase in the prevalence of posts to non-default subreddits during this time period - with the percentage of /r/all slots being occupied by posts to non-default subreddits fluctuating around 26% and actually decreasing somewhat towards the end of this period.

Section 8.5 will look for evidence of such a shift after `reddit` expanded the set of default subreddits to 20 and continued to grow rapidly.

However, an omission in both data-sets must be noted. There are no posts from 'Adult' subreddits

in the data from /r/all - whereas consulting websites like `stattit.com` and `redditmetrics.com` suggests that some highly active subreddits cater to 'Adult' material (i.e. pornography or material of a shocking nature). Posts from 'Adult' subreddits are likely excluded from pages such as /r/all unless a user is signed into an account and has elected to see posts from this type of subreddit. In any case, data on such posts has not been collected - and this may lead to a lower percentage of posts from non-default subreddits on /r/all than would otherwise be the case (because all of the 'Adult' subreddits are not defaults and it is possible that posts from these subreddits would have appeared on an un-filtered version of /r/all).

8.5 A shift away from a shared Front page - part 2

This section considers the prevalence of posts from non-default subreddits on the /r/all page after the `reddit.com` subreddit was decomissioned and the number of default subreddits was expanded on 18th October 2011 (Reddit-Blog, 2011c). The subreddits which are classed as defaults for this period are: AdviceAnimals, announcements, AskReddit, askscience, funny, atheism, aww, bestof, blog, gaming, IAmA, movies, Music, pics, politics, science, technology, todayilearned, videos, worldnews, WTF. Askscience warrants a mention at this stage because it was added as a default subreddit initially but later removed because the influx of new users to the subreddit was seen as problematic - this incident will be considered further in section 8.6.

Looking initially at just the first leaf of the All page the data-set contains 360,725 post observations recorded at 14,429 observation points. Only 15% of these post observations represent non-default subreddits - suggesting that the prevalence of non-default subreddits on /r/all was in general lower than before the number of default subreddits was expanded. This is to be expected because some highly active subreddits have been re-classified from non-default to default. However, figure 8.15 reveals an unexpected trend whereby the prevalence of non-default subreddits on the first leaf of /r/all has actually decreased within this observation period.

From 1st December 2011 data was being collected from the first 4 leafs of the /r/all page - allowing for analysis of the prevalence of non-default subreddits on all four leafs as opposed to just the first leaf. This data-set consists of 100 posts observed at each of 12,353 discrete observation points - and in this data-set 24% of posts come from non-default subreddits. Figure 8.16 shows the percentage of posts on leafs 1-4 of /r/all which originate in non-default subreddits. There is very little evidence of a trend in this data - and if a trend is present it is for a slight reduction in the number of posts to non-default subreddits appearing on /r/all.

Therefore it must be concluded that these analyses *do not* indicate a shift in activity away from the default subreddits. This does not rule out the possibility that such a shift has occurred, but it constrains the possibilities for how this might have occurred. There is no evidence here that posts
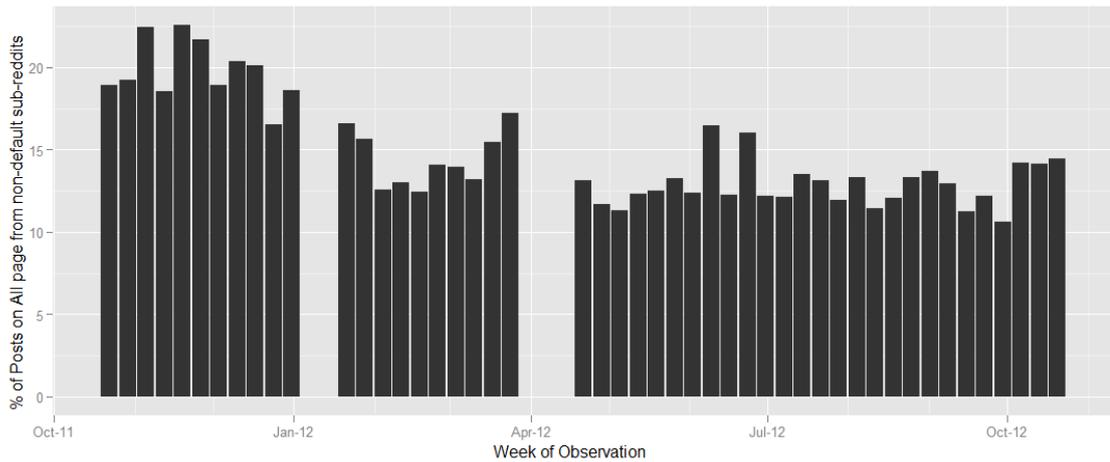
Figure 8.15: Showing the percentage of posts on /r/all which come from non-default subreddits per week.
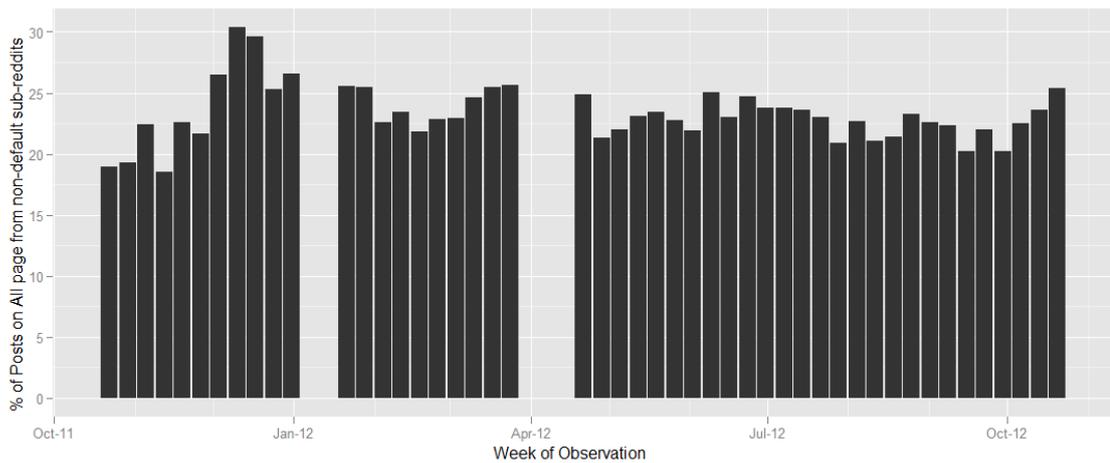
Figure 8.16: Showing the percentage of posts on /r/all (first 4 leafs) which come from non-default subreddits per week.

to non-default subreddits are out-scoring their rivals on default subreddits more frequently over time. Sections 8.2 and 8.3 also suggested that voting rates on default subreddits had in most cases been growing consistently over the duration of data collection for this research.

It is however still possible that a significant *proportion* of voting activity has shifted away from default subreddits - and that this voting activity is progressively more thinly spread across a greater number of non-default subreddits. In this scenario one would not expect to see an increase in the number of posts to non-default subreddits appearing on /r/all because only subreddits with a substantial level of activity stand a chance of having their posts appear on /r/all.

It is a source of some frustration that the available data do not allow for a thorough investigation of the hypothesis that `reddit`'s users are increasingly active on a wider range of non-default

subreddits. This is especially so given that with access to the full history of `reddit`'s back-end procedural data - analysis of the distribution of votes and active users between subreddits over time would be trivial to conduct.

While the data does not show a shift in user activity away from default subreddits such a shift is apparant if one spends any length of time browsing `reddit` in depth. In addition to changing the nature of the website this shift has caused problems for the present research's approach to data collection. This approach was devised in 2009 when the default subreddits were central to the website and other subreddits tended to have very low levels of user activity. In 2013 there are a greater number of active subreddits than ever before and many would argue that the prototypical `reddit` experience now incorporates picking and choosing from the many subreddits which to subscribe to rather than looking exclusively at the default subreddits. Furthermore, there are now many active 'meta' subreddits (e.g. Bestof, TheoryOfReddit, SubRedditDrama) whose sole purpose is to provide commentary on or discussion of events which have taken place elsewhere *on `reddit`*. The manner in which these subreddits interact with other subreddits adds layers of complexity to the functioning of `reddit` as a whole. Developing an understanding of exactly how these new cross-subreddit dynamics affect the website would require a new programme of data collection and analysis - this is beyond the scope of the current research, and as such for some questions we will discuss specific posts or examples withnessed on `reddit` rather than subject them to detailed empirical scrutiny. These are topics which absolutely must be addressed by any comprehensive piece of research about `reddit` as a whole which is completed in 2013 - but they have largely come to light in the later stages of the present research, and therefore cannot be addressed in any depth.

8.6 The effect of being on the list of default Front page subreddits

Reddit's Front page is a focal point for user attention on the website, and for the many users who are not registered or do not sign into an account the makeup of this Front page is determined by the list of default subreddits. Aside from this capacity to reach the audience of unregistered users, how does being on the list of default subreddits affect the activity which takes place within a subreddit?

The /r/askscience subreddit is a good venue for investigating this effect because within the data-collection span of this research it was initially not a default subreddit, it was then added to the list of default subreddits when the list was expanded to 20 (18th October 2011), removed from the default list on 18th December 2011, added again by 18th January 2012 before finally being

removed around 22nd March 2012. It is unusual for a subreddit to be added and removed from the list of default subreddits so frequently over a short period of time - and this has happened because (vocal) users perceived a decline in the subreddit's quality as a result of its being added to the default set of subreddits.

AskScience follows a similar model to AskReddit in that posts are intended to be questions and top-level comments are intended to be answers. There are two major differences between AskReddit and AskScience. Firstly, AskScience deals with questions of a scientific nature and answers are intended to be 'scientific' ("based on repeatable analysis published in a peer reviewed journal") and free of anecdotes or 'layman speculation'. Secondly, the rules governing the discourse on AskScience have historically been much more rigidly enforced than on AskReddit - posts or comments which contravene the subreddit's guidelines are routinely deleted by its moderators.

When AskScience became a default subreddit this was seen as causing problems because new users were being automatically exposed to AskScience and did not have to find the subreddit and subscribe themselves. As a result, these new users of AskScience were perceived to not be following the subreddit's guidelines - there was a perceived increase in behaviour which would be perfectly acceptable elsewhere in the default subreddits but was not acceptable on AskScience (e.g. making jokes in comments).

Unfortunately, comment data was not being collected from AskScience during the critical periods when the subreddit was being added to and removed from the list of default subreddits. Data on these comments has been collected retrospectively and this puts limits on the types of analysis which can be conducted on comments. I will first consider post submission and voting, as suitable 'live' data was collected for these activities.

Figure 8.17: Showing the number of new posts observed on /r/askscience per week. For weeks with only partial data collection the weekly rate has been estimated from their mean. Shading indicates the time periods when AskScience was a default subreddit.

Figure 8.17 shows the number of new posts on AskScience observed per week of observation. Zero counts relate to weeks where the data collection server was not functioning - and where data

collection was only operational for part of a week the weekly rate has been estimated from the mean for the observed interval. Shaded areas on these graphs indicate time periods when AskScience was a default subreddit. On the first occasion when AskScience became a default subreddit there was an immediate increase in the number of posts being submitted, increasing from around 800 per week to around 1500 per week. When Askscience ceased being a default subreddit the rate of new post submission seems to have been declining to previous levels (before the data collection server broke down). When data collection resumed in January AskScience was back on the default list but rates of post submission are lower for this interval than the first period when AskScience was a default subreddit. The data collection server broke down again in late March 2012 and when data collection resumed AskScience had again been removed from the default list of subreddits (but was experiencing the highest rate of post submission observed at any point).

In terms of post submission rates, the first time AskScience became a default subreddit this seems to have had a much greater impact than the second time this occurred. This is perhaps to be expected - when AskScience was first added to the default subreddits list many `reddit` users will likely have 'discovered' it for the first time, when it was subsequently removed from the default set many of these 'new' users will likely have remained as subscribers. On the second occasion when AskScience became a default subreddit it will only have been 'discovered' by users who did not discover it during the previous interval (i.e. users who created their account in the preceding month or during this second interval while AskScience was a default subreddit). It is regrettable that data collection was not operational when AskScience was removed from the default list for the second time because something peculiar seems to have occurred here (a higher rate of post submission when data collection resumed, despite AskScience no longer being a default subreddit).
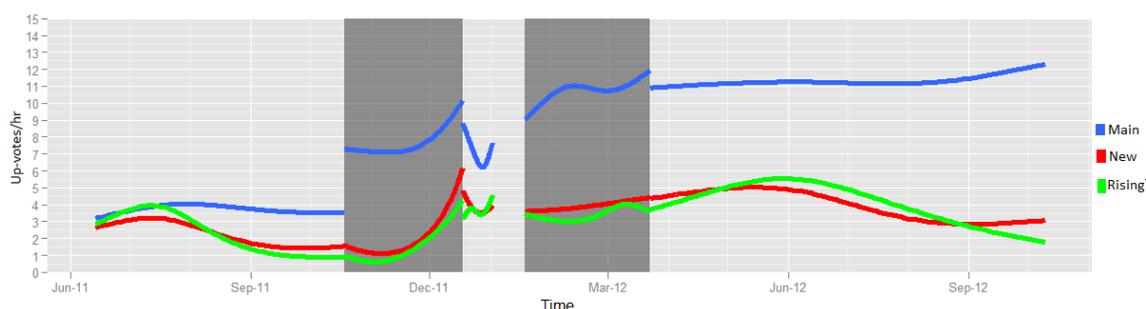
Figure 8.18: Showing the average up-votes per hour rate for the Main, New and Rising pages of /r/askscience. Shading indicates the time periods when AskScience was a default subreddit.

Figure 8.18 shows smoothed poisson regression lines for the hourly rate of up-voting when a post appeared on the New, Rising and Main pages of AskScience. Seperate lines have been drawn for each of the five 'stages' of this observation period (before becoming a default subreddit, first interval as default subreddit, interval when it was removed from default list, second interval as

default subreddit, time after being a default subreddit). As with the rate of post submission, when AskScience becomes a default subreddit for the first time there is an immediate effect on up-vote rate for posts appearing on the Main page (doubling from four up-votes/hour to eight up-votes/hour). Some proportion of this increase will be attributable to posts which appeared on `reddit`'s default Front page at the same time as they appeared on AskScience's Main page. There is no immediate effect on up-vote rate for posts appearing on the New or Rising pages but these increase sharply (along with up-voting on the Main page) towards the end of the first interval when AskScience was a default subreddit. When AskScience was then removed from the default list these voting rates show a sharp decline followed by a sharp increase. On the second occasion when AskScience became a default subreddit the effect on up-vote rates is less pronounced and more gradual. When AskScience was finally removed from the default subreddits list this had little impact on its voting rates. Figures depicting rates of down-voting and commenting show similar patterns over time and have not been reproduced here.

Figure 8.18 suggests that there are both immediate and delayed effects of AskScience being added to or removed from the list of default subreddits. However, the regression lines plotted in this figure concern models of each page type *in isolation* - therefore the voting rates for the New and Rising subreddits have likely been artificially inflated by posts which appeared simultaneously on the Main page and the New/Rising pages. To circumvent this problem negative binomial regression models (of the type utilised heavily in Chapter 5) have been fitted which include all three page types as indicator variables. Including each page type as a coefficient in the same model allows the model to control for the fact that posts appeared on more than one page type at the same time - and therefore produce a better estimate of the voting rates which relate to specific page types. These models were fitted to up-vote rates seperately for each of the five 'stages' of the observation period and details of the models are displayed in table 8.1.

These models suggest that being on the list of default subreddits has primarily influenced the voting rate on the Main page - this is to be expected because posts at the top of the Main page will also appear on the default Front page during these intervals. For instance, between stages 1 and 2 the up-vote rate on the Main page more than doubled but rates on the New and Rising pages did not increase at all. In Chapter 5 I argued that voting rates on a subreddit's 'early stage' pages were a good indicator of its 'health', and as these rates remained stable when AskScience was added to the default subreddits list this would suggest that the 'health' of the subreddit (in terms of the scores of posts which appear in high-visibility locations) may not have changed.

Figure 8.19 shows the number of posts which reached leaf 1 of the AskScience Main page and whether they had a final observed score which was positive (> 0) or negative ($<= 0$). When AskScience was added to the list of default subreddits there has actually been a decrease in the number of posts with a negative score appearing on leaf 1 of the Main page. However, one must
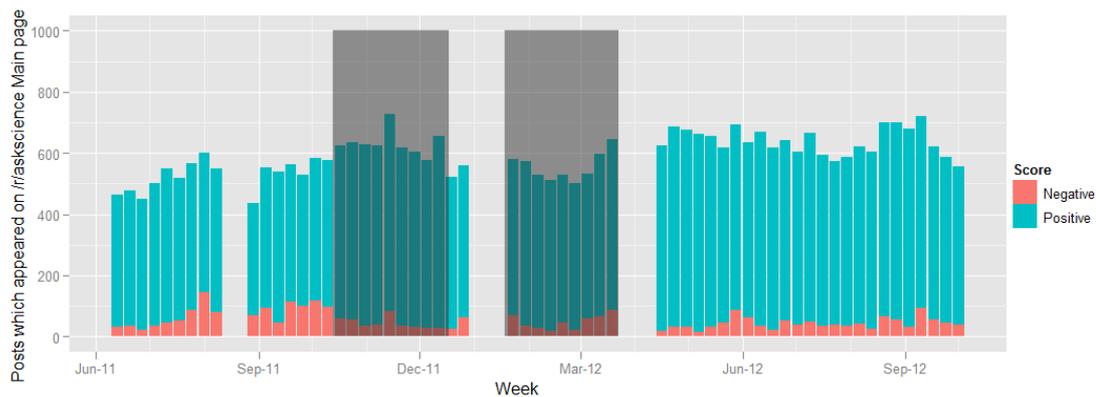
Figure 8.19: Showing whether the last observed scores for posts which appeared on the /r/askscience Main page were positive or negative. For weeks with only partial data collection the weekly rate has been estimated from their mean. Shading indicates the time periods when AskScience was a default subreddit.

also consider the rate of post submission - before being added to the default list most of the posts submitted to AskScience appeared on the Main page at some point, after the change the rate of post submission almost doubled whereas it appears that the subreddit had already 'saturated' the Main page - and consequently a much lower *proportion* of posts appeared on leaf 1 of the Main page. This is most likely the source of the trend in Figure 8.19.

In considering the voting rates and scores of posts to AskScience over the full observation period there is no sign that the subreddit was suffering a decrease in post quality as a consequence of being a default subreddit. However, one must also consider that when AskScience was added to the default subreddit list it saw an influx of new users - and that these new users had the same level of access to the voting system as pre-existing users. If we conceive of AskScience voters at this stage as being composed of two groups; pre-existing users and new users, both sharing the same voting system, this can account for the sense among some users that the subreddit was decreasing in quality. This decrease in quality is not apparant when looking at post scores because the combined set of new and pre-existing users, in aggregate, actually like the content which is appearing in high-visibility locations. The 'problem' is that this content does not match the ideals of pre-existing users, in particular the moderators, and these users are likely more vocal (and certainly more influential) than the new users.

AskScience has always been a relatively heavily moderated subreddit with clear rules about the discourse which takes place there. One of the complaints voiced during the first period when AskScience was a default subreddit was that the moderators could no longer keep up with the volume of incoming submissions. In February 2013 data on the (visible) comments of all the AskScience posts considered above was collected (a total of 652,851 comments). As this data was collected post-hoc only the 'final' ranks, scores and text of the comments is available - this means

that it is not possible to analyse comment voting on the basis of which page a post appeared on, or to analyse the change in comments' ranks over time.

It is however possible to consider comment deletion. Where a comment has been deleted its text field displays '[deleted]'. This identifier does not distinguish between comments which were deleted by their author and those which were deleted by the subreddit's moderators. Also, it is likely that some deleted comments were down-voted to the point where they were no longer displayed and these will not be present in the data on comments collected sevel months later.
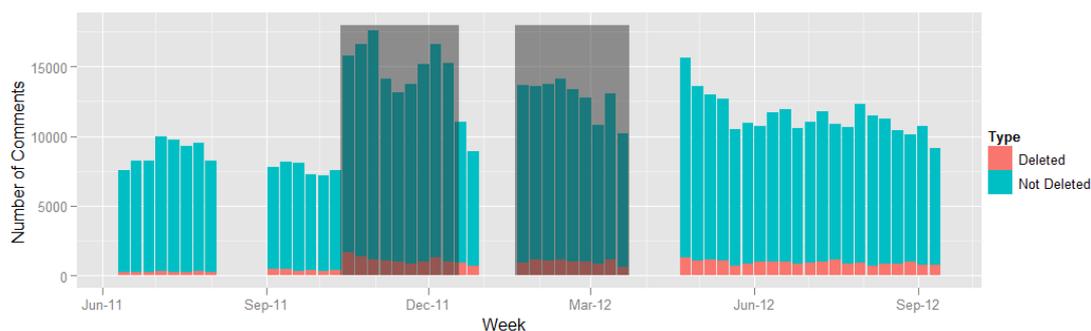
Figure 8.20: Showing the number of comments per week which were deleted and not deleted. Weeks with only partial data collection have been removed from this Figure. Shading indicates the time periods when AskScience was a default subreddit.

Figure 8.20 shows a clear increase in the number of comments which were being deleted after AskScience was added to the default list. The rate of comment submission follows that of post submission in Figure 8.17, doubling when AskScience was first added to the list of default subreddits. The number of comments being deleted also shows a sharp increase at this point, and the proportion of comments which were deleted does not subside to previous levels for the remainder of the observation period (even when AskScience was removed from the default subreddits list). Table 8.2 shows details of a logistic regression model on whether a comment would be deleted with 'stage' (binary, before or after AskScience became a default subreddit) as the explanatory variable. This model suggests that the percentage of comments which were deleted increased from 3.9% to 8.5% after AskScience became a default subreddit for the first time. This doubling of the proportion of comments which 'required' deletion, coupled with a doubling in the rate of comment submission, must have resulted in a substantially greater workload for the subreddit's moderators.

A further model was fitted with coefficients for each of the five 'stages' in this time period, and this model revealed very little difference between stages 2-5, with the estimated percentage of comments deleted ranging between 8.1% and 8.7%.

The insistence on a certain style of discourse for this subreddit, which is distinct from all of the other major subreddits, is where the problems of AskScience being a default subreddit arise. On

other default subreddits more responsibility for sorting 'good' content from 'bad' is delegated to the voting system. On these subreddits an influx of new users can change the nature of how posts are voted on, but because there are no objective standards against which content is judged a shift in the type of post which scores well is likely to be more subtle. On AskScience, with its stricter rules, the comments pages for posts were often littered with 'Deleted Comment' placeholders - a clear sign to readers that there was a developing problem.

This high level of moderation allowed AskScience to retain its specific style in the short-term. Eventually, the moderators of AskScience persuaded `reddit`'s administrators to remove it from the set of default subreddits. AskScience moderators used the tools at their disposal (deletion and the capacity to display messages on the subreddit) to fight the perceived slide in quality, until they were able to shut this off at its source. Thus AskScience was able to retain its own distinctive style of discourse.

This case reinforces some points made earlier about the suitability of `reddit`'s voting system for certain types of content. Content which is quickly and easily digested has an advantage on `reddit` because it will convert views to votes at a faster rate than content which is more effortful to appraise. When AskScience was placed on the default list it was exposed to many new users who did not necessarily know or care about the rules particular to content on this subreddit. In this environment the same type of comments that tend to score well on other subreddits (short and humerous) had an advantage over the longer comments which provided scientific answers with reference to the literature. The AskScience subreddit had been constructed as a space *for comprehensive scientific answers with references to literature* and is therefore not compatible with `reddit`'s voting system and mass of 'Front page' users in combination.

A good analogy in this case is of a garden where the nature of the soil (voting system) encourages a certain type of plant (content) to grow. Gardens (subreddits) where this type of plant is desired can be left mostly to grow wild and free, with the best plants naturally out-competing their rivals (e.g. /r/funny). Gardens where a different type of plant is desired (i.e. /r/AskScience) must be maintained, the plants which naturally thrive must be weeded out to allow the desired plants to grow. It has always been the case on AskScience that the gardening staff (moderators) were willing to weed out the undesirable plants, but when AskScience was added to the default list this immediately changed the nature of the 'soil' (voting users as a whole were behaving differently) and doubled the size of the garden, while the number of gardeners remained the same. The moderators were able to stall the change in the style of content on their subreddit by frantically 'pulling out weeds' - but the only lasting solution in their eyes was to remove AskScience from the list of default subreddits.

The problems encountered when AskScience became a default subreddit would likely also apply to any other 'moderated/curated' subreddit which strives for a style of discourse which is not typical

of the default subreddits.

	Coefficient	Std Error	p value	Est. Up-votes per minute
Stage 1 - 13/06/11 to 17/10/11 - Not Default Subreddit - N = 163,007 observations				
Intercept	-5.1351	0.0154	< 0.001	0.005
New - Leaf 1	-0.1653	0.0122	< 0.001	0.005
Rising - Leaf 1	0.4219	0.0177	< 0.001	0.009
Main - Leaf 1	2.3897	0.0110	< 0.001	0.064
Stage 2 - 18/10/11 to 18/12/11 - Default Subreddit - N = 172,286 observations				
Intercept	-5.3065	0.0101	< 0.001	0.005
New - Leaf 1	-0.0976	0.0122	< 0.001	0.004
Rising - Leaf 1	-0.0097	0.0177	0.584	0.005
Main - Leaf 1	3.2913	0.0110	< 0.001	0.133
Stage 3 - 19/08/11 to 17/01/12 - Not Default Subreddit - N = 77,345 observations				
Intercept	-5.6061	0.0141	< 0.001	0.003
New - Leaf 1	0.0866	0.0241	< 0.001	0.004
Rising - Leaf 1	1.5248	0.0373	< 0.001	0.017
Main - Leaf 1	3.3858	0.0202	< 0.001	0.108
Stage 4 - 18/01/12 to 22/03/12 - Default Subreddit - N = 331,833 observations				
Intercept	-5.5511	0.0068	< 0.001	0.004
New - Leaf 1	0.1526	0.0114	< 0.001	0.004
Rising - Leaf 1	1.4811	0.0273	< 0.001	0.017
Main - Leaf 1	3.7822	0.0101	< 0.001	0.171
Stage 5 - 23/03/12 to 09/08/12 - Not Default Subreddit - N = 878,144 observations				
Intercept	-5.6338	0.0042	< 0.001	0.003
New - Leaf 1	0.2893	0.0074	< 0.001	0.005
Rising - Leaf 1	2.0452	0.0126	< 0.001	0.028
Main - Leaf 1	3.9118	0.0061	< 0.001	0.178

Table 8.1: Showing model parameters for negative binomial regression models of number of new up-votes by page type (leaf 1 only) with an offset for the time since previous observation. The dispersion parameters for these models (θ) ranged between 0.31 and 0.44.

	Coefficient	Std Error	p value	% of comments deleted
Intercept (Before AskScience became default)	-3.2379	0.0149	< 0.001	3.9%
After Askscience became default	0.7691	0.0158	< 0.001	8.5%

Table 8.2: Showing model parameters for a logistic regression model of whether comments were deleted. The explanatory variable is whether the comment was created before or after AskScience was added to the list of default subreddits for the first time on 18th October 2011.

8.7 Meta Subreddits

Meta subreddits are subreddits whose subject is `reddit` itself or some portion therof, their increasing prevalence is an indicator that `reddit` as a whole is becoming larger and more complex. Some of these meta subreddits are intended to mitigate this increasing scale and complexity (e.g. BestOf) and while they may be useful to their subscribers in doing so - their very presence serves to increase the overall complexity of `reddit` still further.

Some examples of Meta subreddits are as follows:

- BestOf - Posts submitted to /r/BestOf must link to a pre-existing post or comment on `reddit`. The expressed purpose of this subreddit is to highlight high-quality items of content which might otherwise have gone unseen by `reddit`'s users en masse (e.g. good comments submitted too late to be up-voted to prominence on their post's page, posts to small subreddits). In August 2012 the decision was taken to exclude default subreddits from /r/BestOf because it was perceived that there was too much overlap between /r/BestOf and the default Front page. /r/BestOf is considered further in section 8.7.2.

- TheoryOfReddit - This is a particularly interesting subreddit, wherein users discuss the workings of `reddit` and propose theories/experiments or report on the findings generated by their own studies. TheoryOfReddit is considered in more detail in section 8.7.1. TheoryOfReddit users have also compiled a list of subreddits which they consider to be 'Meta' subreddits [1]

- SubRedditDrama - Posts submitted to /r/SubRedditDrama link to comment threads where fights or other 'drama' has taken place.

- ShitRedditSays - Users submit links to 'horrible' comments that have a score of +10 or more - the intention is to criticise `reddit`'s users for up-voting these comments.

- CircleJerk - Users satirise popular trends on `reddit`.

- CircleBroke - Describes itself as "A place for discussing or complaining about trends on `reddit`. It's like /r/circlejerk without the satire, or /r/TheoryOfReddit for people who hate `reddit`."

The last three of these meta subreddits are explicitly 'anti-`reddit`' - and their presence on `reddit` with active user-bases tells us something about how `reddit` is changing. In addition to many of its users abandoning the default subreddits, some of these users are congregating in subreddits whose purpose is to criticise or satirise the activity on (mostly) large subreddits. This kind of sentiment is also routinely expressed with regard to specific subreddits (/r/atheism and /r/politics in particular) in comments on default subreddits.

---

[1] `www.reddit.com`/r/TheoryOfReddit/wiki/meta_subreddits

8.7.1 TheoryOfReddit

TheoryOfReddit describes itself as "a mildly navel-gazing space for inquiring into what makes the `reddit` community work and what we in the community can do to help make it better". The subreddit has existed for several years and appears to have been created to facilitate discussion among users who found `reddit`'s voting-mediated interaction and social dynamics interesting. As `reddit` has increased in complexity TheoryOfReddit has become a useful subreddit for users who want to know how it all 'works' and fits together. On this subreddit users submit observations, theories and even reports on their own studies of `reddit`.
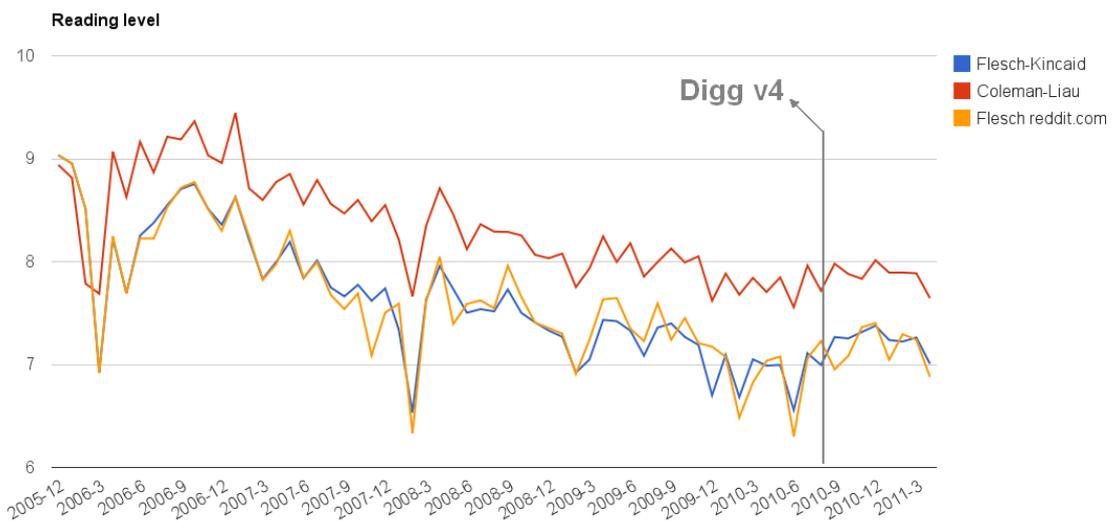
Figure 8.21: Showing the reading level of comments on `reddit` over time. Taken from redd.it/l8id4 post by user LinuxFreeOrDie.

The top-scoring post of all time (although /r/bestof has influenced this) to TheoryOfReddit (redd.it/l8id4) details a study of `reddit` comments conducted by user LinuxFreeOrDie. This study analysed 'millions of comments' (although it does not state how comments were sampled) and looked for differences in metrics (reading level, comment length, presence of swearing and Internet slang) both over time and between subreddits. The study concludes that the reading level of comments (calculated with both 'Flesch-Kincaid' and 'Coleman-Liau' methods) had dropped by about one grade between `reddit`'s launch and March 2011 (Figure 8.21). The length of comments has also declined during this time (Figure 8.22), shrinking to one-third of their original length on average. One interesting aspect of these graphs is that the changes have mostly occurred during `reddit`'s early history and from a point in 2008 onwards all of these measures are quite stable.

One of the questions which this post sought to answer was whether the influx of `Digg` users in August 2010 made `reddit` 'dumb' - and LinuxFreeOrDie concludes that "Digg had little, if any, effect. Reddit's rise in popularity took place well before `Digg`, and the jump from 50k people
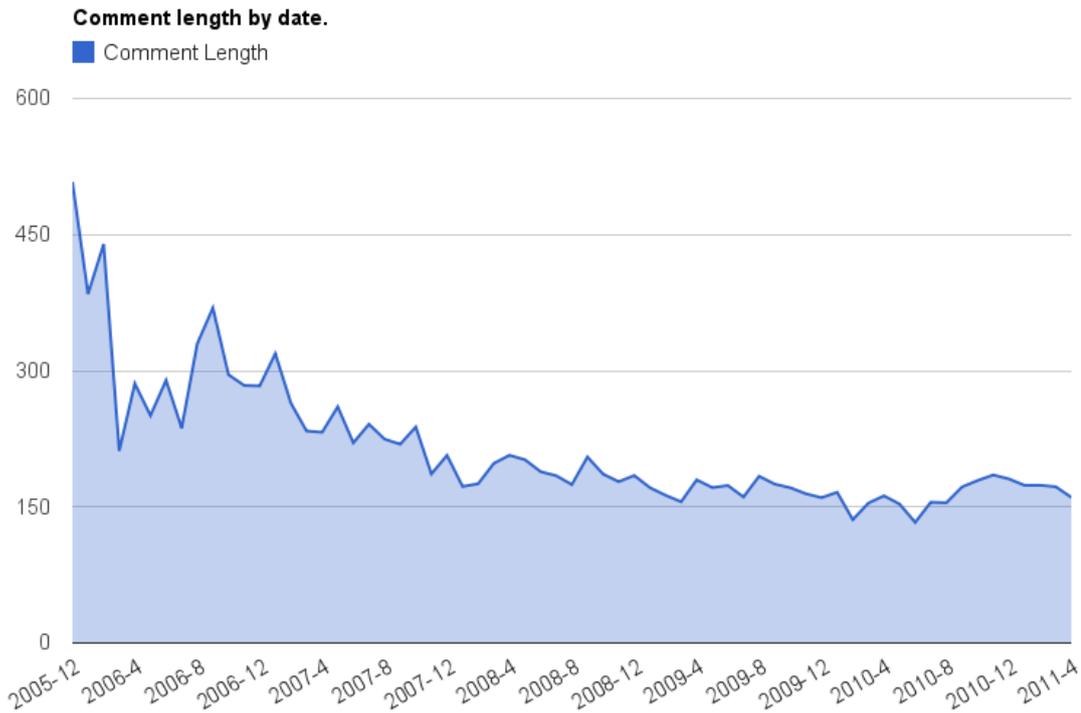
Figure 8.22: Showing the length of comments on `reddit` over time. Taken from redd.it/l8id4 post by user LinuxFreeOrDie.

to 300k has a much greater effect than 300k to a million it seems." It is not clear where these user figures are taken from and it seems unlikely that `reddit` had 50,000 users in December 2005 (shortly after it launched). However it is clear that in absolute terms `reddit` saw much greater growth in the latter half of this observation period than the first. It is fair to say that the changes documented in this post occurred as `reddit` transitioned from being a small website to being a moderately active one - beyond which point (as `reddit` became a large and recognisable website) all of the measures remained stable.

There are limits to the weight which can be placed on these analyses however because the data collection methods are not clearly described and for the early period of observation analyses are based on relatively few comments. Also, during its early stages `reddit` was changing structurally - the website began with only one 'subreddit', with the system of subreddits being added some time after its launch and the capacity for users to create their own subreddits not being added until 2008. This has been taken into account for the graph of reading level by analysing the `reddit.com` subreddit seperately and showing that its reading rate was similar to that derived from comments to other subreddits. However, the analysis of comment length is likely to be confounded by the introduction of new subreddits - one would not expect /r/pics posts to have long comments and a decline in global comment length could be attributed to an increase in the number of comments from subreddits like /r/pics in the sample.
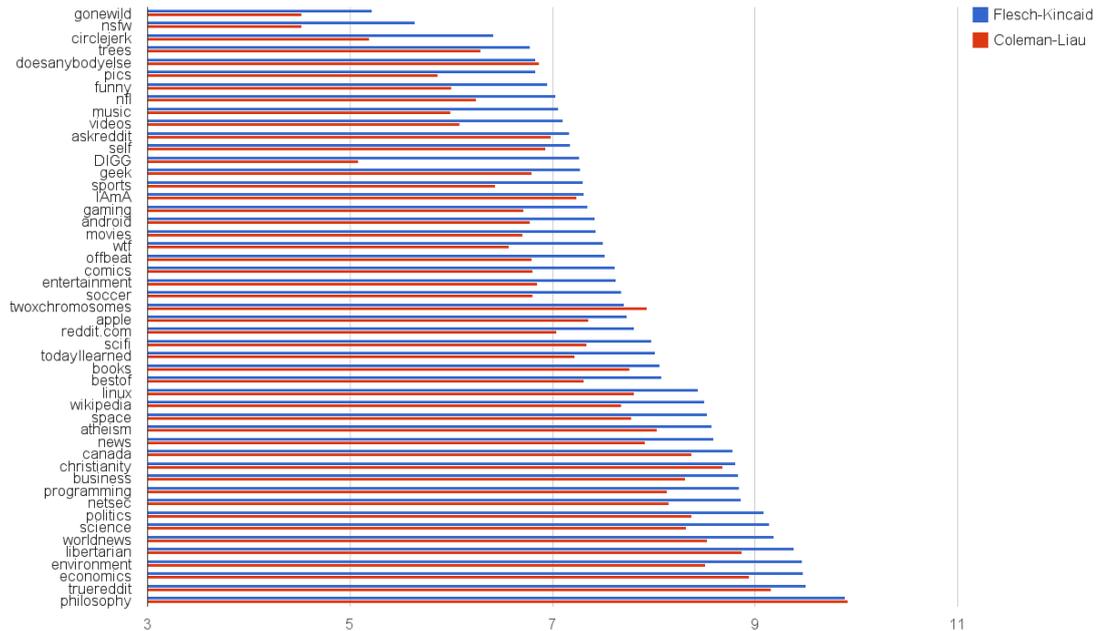
Figure 8.23: Showing the reading level of comments on `reddit` for individual subreddits. Taken from redd.it/l8id4 post by user LinuxFreeOrDie.

This post also reports an analysis of reading rate which compares comments from a selection of subreddits (see Figure 8.23). While many of the same caveats about sampling apply here this graph shows something of `reddit`'s diversity - with subreddits for 'adult content' and satire unsurprisingly being found to have low reading rates while more 'high brow' subreddits like economics and philosophy have comments of greater literary prowess.

This study is in some ways typical of the type of study users submit to TheoryOfReddit. While these studies lack the rigour, detail and formality of peer-reviewed publications one must bear in mind that they have not been produced by or for professional researchers/academics. That `reddit` users who (one assumes) are not professional researchers will work to produce research on `reddit` (involving data collection, analysis and visualisation), for the edification of other `reddit` users, speaks to the motivational power of `reddit`'s infrastructure and a sense of community which is palpable in certain subreddits.

As `reddit` has increased in scale and complexity, and a greater proportion of activity takes place away from the default Front page, TheoryOfReddit has become increasingly valuable as a way of merely finding out what's happening on `reddit`. Here TheoryOfReddit's strength is in its number of users - there are perhaps thousands of TheoryOfReddit users who are active on `reddit`, and the eyes of all these users, coupled with their awareness of which events are significant from the perspective of wanting to understand `reddit`, make TheoryOfReddit a useful place to keep abreast of certain types of event on `reddit`. In fact several examples described in this thesis were first

encountered through TheoryOfReddit posts (e.g. the comment where 'fuzzing' of vote counts was first described).

8.7.2 BestOf

The BestOf subreddit is also a useful resource for finding out what's happening on `reddit`. Posts submitted to /r/BestOf must link to a pre-existing post or comment on `reddit`. The expressed purpose of this subreddit is to highlight high-quality items of content which might otherwise have gone unseen by `reddit`'s users en masse (e.g. good comments submitted too late to be up-voted to prominence on their post's page, posts to small subreddits). In August 2012 the decision was taken to exclude default subreddits from /r/BestOf because of a commonly expressed opinion that there was too much overlap between /r/BestOf and the default Front page.

The process whereby this decision was taken and implemented is interesting in itself as an example of moderators leading an effort to change the purpose of a subreddit while gauging feedback from users and attempting to build consensus. This idea seems to have first been prominently discussed in a post (redd.it/rhkm7) to /r/BestOf by a moderator on 28th March 2012 - titled "The BestOf mods are considering a one-week experiment: NO posts from default subreddits. Thoughts?". The highest-scoring comments on this post are not supportive of the idea and most of these make two points. Firstly, they argue that it would change the nature of the subreddit such that it no longer highlighted the *best* content on all of `reddit`. Secondly, in many of the top-scoring comments users state that they no longer subscribe to default subreddits and that they like being shown only the best items from these subreddits through /r/BestOf.

Several months later, on August 5th 2012, another moderator makes a post announcing a one-week experiment during which no comments from default subreddits would be allowed. This post lists the contemporary default subreddits which would be banned and, strangely, references the earlier post as an indicator of support from /r/BestOf users - it is possible that the voting scores for comments changed in the intervening period, but in its current form (observed early 2013) this post did not seem to show support for the idea. The comments on this later post are more mixed but again there are prominent comments which criticise the idea.

One week later a moderator makes another post about the "results of the 'no defaults' experiment" (redd.it/y3ufx) in which it is announced that "As of this post, and until further notice, /r/bestof will no longer allow comments from default subreddits to be submitted here." The reasons for this decision are cited as 1) a perceived increase in the quality of submissions to /r/bestof over the preceding week, 2) the support of users in an unspecified poll, and 3) links to posts where users have discussed the change (acknowledging that not everyone agrees with it).

This post also describes the change as a 'golden opportunity' to provide a tool for the discovery

of new subreddits - and notes that this is a feature often requested in /r/IdeasForTheAdmins. In support of this case the post references an increase in subscribers for subreddits which were featured on /r/BestOf in the previous week, quoting figures of a few thousand new subscribers to specific small subreddits. This post clearly expresses recognition that the decision being taken is a controversial one, and closes by suggesting some alternative subreddits which might fill the gap left by the exclusion of default subreddits from /r/BestOf - these include /r/DefaultGems, which is essentially a counterpart of /r/BestOf that *only* deals with comments from default subreddits. The top-scoring comments on this post again express discontent with the decision - and one must assume that beyond this point the /r/BestOf community has fractured to some degree, with users who wanted to see good comments from default subreddits going elsewhere.

The split among /r/BestOf users seems to have occurred because there were two distinct sets of users. One set had already abandoned the default subreddits in favour of more specialised and lower activity subreddits - they wanted /r/BestOf to keep them appraised of high-quality contributions to the default subreddits because they no longer subscribed there. The other set were likely still subscribed to the default subreddits and they wanted to be shown content from all of the non-default subreddits, perhaps hoping to discover interesting subreddits in the process.

Despite their claims to the contrary, this appears to have been a unilateral decision taken by the BestOf moderators. The top-scoring comments for all of the posts discussed here oppose the move, it was certainly not a democratic decision and there does not appear to have been consensus among /r/BestOf users that it was the right decision. However, this does not necessarily mean that it was a bad decision. Active users of BestOf at the time opposed the decision, but this is a population that almost by definition liked things the way they were. Users who wanted a subreddit which showed them content from smaller subreddits would likely not have been active on /r/BestOf before this decision because it did not serve that purpose. For the global `reddit` community, this may have been the right decision because `reddit` did appear to need a way to direct people's attention beyond the default subreddits.

As /r/BestOf is itself a default subreddit this may also have had an effect whereby the default Front page regained some of its capacity to show users what was happening on `reddit` as a whole. In the early stages of this research (2009) much of `reddit`'s activity was occurring in default subreddits and the highest-scoring posts were being broadcast to all on the Front page. Over the years, as users moved away from the default subreddits and a much greater number of subreddits had high activity levels, the default Front page lost this aspect because posts to non-default subreddits could not appear there. Now, a selection of (the 'Best of') these posts are appearing on the Front page, and these serve as examples of the different types of content which `reddit` provides beyond the default Front page. In a sense BestOf is now bridging the gap between the 'old' `reddit` (typified by the large default subreddits) and the 'new' `reddit` (typified by a massive diversity of smaller

subreddits).

BestOf can also be thought of as working against the fragmentation of discourse (Sunstein, 2002) on `reddit`. Each subreddit provides its own individual sub-community with 'shared experiences', and in the case of /r/BestOf these are shared experiences drawn from `reddit` in its entirety - no subreddit is too small to be precluded from having a post featured on /r/BestOf. While this subreddit is not going to reverse the broader trend towards fragmentation, users who view content on /r/BestOf will at least become acquainted with a sample from the plethora of discourses taking place in subreddits they may never have heard of. Its counterpart (/r/DefaultGems) may be able to provide a similar service for users who have unsubscribed from the default subreddits - and in fact many of the users who opposed the exclusion of default subreddits from /r/BestOf cited this as their reason. It is perhaps a reason for optimism that users who had found no value in the default subreddits and unsubscribed from them *still wished to maintain a level of familiarity with the activity taking place there.* It is possible that users could sense that they were walling themselves off from the 'well of shared experience' or mass of 'ordinary users', and saw browsing /r/Bestof in its old form as a way to maintain contact with this 'default' side of `reddit`. The massive scope of /r/BestOf also implies that users who browse there are going to be presented with 'information that would not have sought out for themselves' (Sunstein, 2002).

Subreddits like /r/BestOf and /r/DefaultGems can also be thought of as 'aggregators of aggregators' - and comments which appear there prominently will have been 'aggregated' several times. Consider a hypothetical comment submitted to a post on /r/TheoryOfReddit - the number of users who see this comment will be determined firstly by the score/rank of its parent post and secondly by the comment's own score, the more users who see this comment the more likely it is that one of them will submit it to /r/BestOf. Once a link to the comment is submitted there the comment is subjected to a third level of aggregation, now as a post in its own right (and therefore beginning life on a sparsely-upvoted New page, but with the potential to make it all the way to the default Front page). The posts which score well on /r/BestOf and /r/DefaultGems could be described as 'triple-distilled' - and it remains to be seen what impact this nesting of aggregation processes might have, or how many layers might eventually be added.

The decision to only allow non-default subreddits on /r/BestOf also had some unforseen consequences related to 'subreddit invasions' and 'voting brigades', these are discussed in the following section. Incidentally, this topic was re-visited on /r/TheoryOfReddit six months later (because by then /r/BestOf had also banned text posts) for a post-mortem of sorts in which views are mixed but concerns are expressed about the issue of subreddit invasions (redd.it/18nvk6).

8.8 Subreddit invasions and voting brigades

In the above section the post about /r/BestOf banning submissions from default subreddits (redd.it/y3ufx) described the growth of small subreddits as a positive outcome of this decision. For example: "/r/WritingPrompts jumped from 237 users to 2771 thanks to being featured here. Thats a pretty significant spike in traffic, and means the difference between a struggling new community vs. a prospering, self-sustainable one."

However, since then it has become clear that this kind of growth is not always a positive development. Consider for a moment the perspective of pre-existing users of a small subreddit - a small group (237 users in the case of /r/WritingPrompts) may have been using this subreddit for an extended period of time and found it useful, perhaps developing a strong sense of community along the way. Then one day someone posts a link to a particular comment on /r/BestOf, its a good comment and it scores well, appearing prominently on /r/BestOf and perhaps on the default Front page as well. Within one day 2,534 new users subscribe to the /r/WritingPrompts subreddit and the pre-existing community of users now becomes a minority on 'their own' subreddit, representing less than 10% of its total users. The term 'subreddit invasion' has been coined to describe this phenomenon.

The user Jess_than_three, moderator of /r/ainbow (a subreddit for LGBT (Lesbian, Bi-sexual, Gay and Transgender) issues) made a post about this to /r/TheoryOfReddit on December 15th 2012 (redd.it/14wpow). In this post Jess_than_three cites a number of harms to the subreddit resulting from links (to /r/ainbow content) submitted to /r/SubRedditDrama in particular. Chief among these is the 'flipping' or distortion of comment scores. When a link to content appears prominently on /r/SubRedditDrama this brings an influx of users to that particular comment/post and the voting behaviour of these 'external' users is different to that of regular /r/ainbow users (this is what's referred to by the term 'voting brigade'). This influx of 'external' users can result in these users outnumbering the regular users on a specific post's comments page. The result is that comments which had been scoring negatively can be flipped so that they score positively and vice versa - in short the comments pages for these posts do not accurately reflect the opinions of long-term users of the subreddit. This can lead to the alienation of pre-existing users, in the case of /r/ainbow comments which were sometimes anti-LGBT scored positively, making "the subreddit feel hostile to members of its own community".

This post also serves as an example of `reddit`'s users attempting to find ways to redress problems on `reddit`. In the post Jess_than_three calls for the utilisation of 'No Participation' URLs and CSS, linking to a previous post by KortoloB (redd.it/10mqi3) which introduces this concept. CSS (Cascading Style Sheets) controls the way a web page looks, individual subreddits can modify part of their own CSS files to give their subreddit a distinctive look by, for example, changing

the background or the icons for voting buttons. KortoloB had created a modified CSS file which removed the voting and commenting buttons when a link to the subreddit was pre-fixed by 'np.'. KortoloB referred to this as a 'No Participation' stylesheet. The rationale behind this was that meta subreddits could prefix their links with 'np.' and as a result when users of these subreddits followed these links (to a subreddit which had installed the 'No Participation' stylesheet) they would not have buttons whereby they could interact with the linked post. In early 2013 /r/SubRedditDrama now mandates that all links should utilise the 'np.' prefix to mitigate the effect of 'voting brigades', and while the precise series of events which lead to this change are not known it is likely that Jess_than_three's post to /r/TheoryOfReddit has played a part in this.

The capacity for individual subreddits to modify their visual style is one of the ways in which `reddit` is an 'open' platform, and this case is one example of Redditors using this control in an attempt to fix a newly apparant problem. There have also been cases where subreddits wished to remove the down-vote button entirely and have achieved this through modifying their CSS (e.g. redd.it/18ddyy).

However, there are constraints on the effectiveness of this method because it operates on the subreddits' visual display only. `Reddit`'s users have the option to ignore customised styles for subreddits - if they have checked this option then subreddits always appear in the same way, including the presence of voting buttons. Similarly, many `reddit` users browse with customised extensions (the 'Reddit Enhancement Suite' or RES being the most popular) and these will ignore a subreddit's CSS and show the voting buttons. While it is possible to change the *look* of a subreddit (even so far as not displaying voting buttons) it is not possible to modify any aspect its 'back-end' code - users who want to can work around the non-display of vote buttons and if they do their votes will be counted in the normal way by `reddit`'s algorithms.

Concepts like 'subreddit invasions' and 'voting brigades' are themselves ominous signs from the fragmentation of discourse perspective (Sunstein, 2002) - if there are already isolated communities which are hostile to one another this implies a level of balkanization which does not bode well for the future.

The example of voting brigades and the 'No Participation' stylesheet is yet another one where `reddit`'s administrators or paid employees are conspicuous by their absence. It is 'ordinary' `reddit` users (although mostly moderators of at least one subreddit) who have been involved in first identifying and documenting the problem, then producing a solution and pushing for the adoption of this solution.

Reddit (the company) increasingly describes itself as a facilitator of many communities (subreddits) and opening up some control over how a subreddit functions (e.g. allowing only certain types of vote or only votes from certain users, choosing which algorithm to use to determine the ranks of

content) would be a logical next step, and there certainly seems to be an appetite for this among the moderators of subreddits. The problems with doing so are likely of a technical nature.

8.9   `Reddit`'s murky and dubious depths

`Reddit`'s administrators have always had a hands-off approach to content and embraced 'free speech' as one of `reddit`'s guiding principles. Content which is demonstrably illegal has always been an exception to this, but `reddit`'s administrators have tended to tolerate content which is legally questionable, even when it is morally objectionable to most users. In both of these regards (freedom of speech and illegal content) `Reddit`'s is a US-centric interpretation of these concepts.

This unwillingness to 'impose censorship' on the part of `reddit`'s administrators has allowed certain subreddits which are host to deliberately shocking or outrageous content to persist on the website *until they are publicised by an external entity*, at which point `reddit`'s adminisrtators have tended to shut down the offending subreddit(s) and/or amend their user agreement to forbid the offending content. There are two well known occasions where this has occurred.

The first of these happened on 30th September 2011 when CNN's Anderson Cooper ran a piece about `reddit` which focused on /r/Jailbait - a subreddit which was purportedly catering to sexually suggestive images of minors. A post about this piece appeared on `reddit`'s Front page through /r/AskReddit (redd.it/kvzx4) and in the ensuing days there was much discussion of this and other subreddits catering to dubious content. Around 11 days later the /r/jailbait subreddit and a number of other dubious subreddits were shut down (redd.it/l7q74) and `reddit`'s terms of service were amended to disallow sexually suggestive images of minors.

There were a number of strong themes to the discussion which took place around this issue. Firstly, some `reddit` users saw Anderson Cooper's initial piece as an attack on `reddit` because it did not clearly state that /r/Jailbait was a small subreddit which was not characteristic of the website as a whole. Secondly, users expressed concern that if /r/Jailbait was shut down this would mark the end of `reddit` as a haven for free speech and that other subreddits dealing with dubious topics of a different nature (e.g. drugs, downloading of copyright material) might also be subject to closure in future. Thirdly, some comments suggest that /r/Jailbait being featured on CNN had the effect of making the subreddit worse - they hypothesised that many people who were not previously `reddit` users will have seen the piece on CNN and then deliberately sought out the /r/Jailbait subreddit. One post to /r/WTF on 10th October 2011 showed evidence that a user who had posted an image of a young child to /r/Jailbait had subsequently been 'hounded' for nude images of the child. It was on the following day that /r/Jailbait and a number of other subreddits were shut down, with the administrators stating that they 'threatened the structural integrity of `reddit`'.

In posts about the shutting down of these subreddits there are some high-scoring comments which endorse the decision - users who did not want the website they used to be associated with this type of content or behaviour. There are other high-scoring comments which express disapproval *if* the decision was taken because of the piece on CNN, but who would not object if the decision was taken because the subreddit had been facilitating the exchange of child pornography (as suggested by the previous day's /r/WTF post).

The second such incident rose to the attention of news outlets in October 2012 and concerned a subreddit called /r/creepshots. On the Creepshots subreddit users would submit sexualised photos of young women, usually taken without their consent in public. The first time when Creepshots was mentioned on the `reddit` Front page was on 27th September 2012 - the first post about this was from a user stating that "I found my picture on r/creepshots and I know exactly who posted it. Can this be used to shut down the entire subreddit? I'm a minor.". This post was only observed on the Front page for one hour, and as it disappeared from the Front page another post about the issue took its place, titled " Why is it if people take pictures of attractive girls and post them to a website it's creepy, but when people post photos of fat people who are in walmart to a website, it's hilarious? ". Both of these posts appeared on the Front page through /r/AskReddit, the second post was observed on the Front page for ten hours. The first of these posts is accusatory, the second seems to be making a point in defence of /r/creepshots - and it is the second post which appeared on the Front page for much longer. This could be read as evidence that AskReddit's voting users on the whole were supportive of /r/creepshots - but all of the posts about this topic have since been deleted and therefore it is not clear whether 1) the first post was deleted/hidden by a moderator (it may have named the offending party and therefore broken a central rule on `reddit` regarding the posting of users' personal information) or was down-voted off the Front page, and 2) whether comments on the second post were in agreement with the tone of the post's title.

A third post about the topic appeared on the Front page 2 days later on 29th September - this was a post to the IAmA subreddit where a user identified themself as a 'student who was taught by Christopher Bailey, AKA, the substitute teacher who posted on r/creepshots' and presumably answered questions about him (this post has also been deleted). In the interim it seems that the culprit in the first post about this topic had been identified and searched then arrested by the authorities (apparantly for having sent lewd photographs to students and not for their activity on /r/creepshots). This part of the story is recounted in an article published 10th October (Baker, 2012) on `jezebel.com`, a website describing itself as 'CELEBRITY, SEX, FASHION FOR WOMEN'. The same article advocates the naming of individuals involved as a strategy for shutting down /r/creepshots and describes an endeavour ('Predditors') being undertaken by 'anonymous redditors' which is attempting to do this.

This article seems to have been at the forefront of a flurry of media interest in the story over the

ensuing days. Here we must distinguish between two distinct but inter-twined stories. There are further articles about the fact that /r/creepshots exists and about users' attempts to identify and expose individuals who are involved (Guardian, 2012a), but a second story also appears at around this time. This second story concerns the identification of an infamous `reddit` user (Violentacrez) who had been the creator of /r/jailbait and was also involved in /r/creepshots and a selection of other objectionable subreddits. An article by Adrien Chen on the Gawker website (Chen, 2012) revealed the identity of Violentacrez and suggested a direct line of communication between him and `reddit`'s administrators whereby Violentacrez served as a monitor and informant of sorts on outright illegal content submitted to 'his' subreddits.

At this point it is important to introduce some background information in order to understand what happened next. Firstly, there already existed some level of animosity between `reddit`'s users and Gawker (Adrien Chen in particular). This anomosity stems from articles which were critical of `reddit` and a particular post in which a teenager dying from cancer (lucidending) spent their last hours (before ending their life) on `reddit` talking to other users in comments. Many of `reddit`'s users were moved by this post and Adrien Chen later stated on `Twitter` that lucidending had been him playing a prank on `reddit` (he subsequently dismissed this tweet as a joke).

Secondly, one of `reddit`'s most rigidly enforced site-wide policies relates to the posting of users' personal information (known in hacker communities as 'doxxing'). `Reddit` is intended to facilitate pseudonymous interaction, users have usernames which identify them but `reddit` does not hold information on their off-line identity. This rule exists partially in response to several cases where users who were perceived to have committed some wrong were identified, had their contact information posted, and then suffered harassment at the hands of some `reddit` users (e.g. redd.it/ky6u7; redd.it/mcy7l). Posting a user's personal information on `reddit` is taken very seriously by the administrators and is likely to result in a ban from the website.

In this context '`reddit`'s' response to the Gawker article which identified Violentacrez is more un-derstandable - an entity which had already been established as an 'enemy' of `reddit` was breaking the site's cardinal rule in a very public fashion and in relation to one of its most recognisable users. This response took the form of several high-profile subreddits (including /r/politics) banning links to the Gawker website (Guardian, 2012b). Perhaps tellingly, these bans were implemented by the moderators of the subreddits - it was perceived that Violentacrez's status as a moderator on the objectionable subreddits had made him a target for Chen.

One of the outcomes of this story was the stimulation of debate on free speech and privacy - debates which periodically rise to prominence on `reddit`. On one side of the free speech debate there are people who believe that objectionable subreddits should be allowed (provided that they are legal) because when one wishes to allow 'free speech' one must accept that there are people who will use this right in ways which the majority find distasteful. People on this side of the debate would not

wish to see any subreddit banned unless it was a venue where users actually broke the law.

On the other side there are people who believe that `reddit` is under no compulsion to allow totally 'free speech' and that attempting to do so is untenable in the long run. These people would be happy to see objectionable subreddits banned because their presence tarnishes `reddit`'s image and because they do not wish to share a web space with the offending parties. From this perspective the question becomes one of where to draw the line - if the legal/illegal dichotomy is abandoned as the yardstick for which subreddits will be allowed to continue then how 'objectionable' does a subreddit need to be before it is banned and who makes this judgment?

The perspective voiced by external sources on these issues with regard to `reddit` and /r/creepshots are somewhat different and characterised by a perception of hypocrisy. The first form this hypocrisy takes is in the willingness to allow users to post surreptitiously taken sexualised photographs of young women without their consent or knowledge - while objecting to the identification of the users who have submitted these images to `reddit`. This is an absolutely legitimate complaint and the situation is not 'right' in any moral sense. It is however understandable, in that the offending parties were 'citizens' of `reddit` and were availing of the 'free speech' protections which `reddit` affords to all users - the victims in this case were generally not 'citizens' of `reddit` and therefore received no protection from `reddit`. The content on the subreddit was not generally perceived to be illegal, and therefore people who objected to /r/creepshots had no legal recourse and the subreddit was allowed to exist under `reddit`'s rules at the time. Posting information which identified a `reddit` user was however not allowed by the rules. Thus by logic internal to `reddit` a subreddit like /r/creepshots was allowed but identifying the individuals involved was not.

However, this logic comes unstuck when we consider that `reddit` is not self-contained and isolated but rather an openly accessible, inherently public, website. While a user may have a right to 'free speech' (which varies depending on the country they are posting from) they do not have a *right* to participate anonymously on-line. The capacity to participate anonymously on `reddit` is granted *by `reddit`* (in that the administrators will not reveal a user's IP address or any information which could identify them without a subpoena/warrant, and they will not allow other users to post this information on `reddit` irrespective of how it has been obtained). If one believes that a subreddit like /r/creepshots should be allowed to exist purely because it doesn't break US laws and doesn't contravene `reddit`'s rules - then one must also 'allow' other users to exercise their rights to 'free speech' by attempting to identify and publicly name the individuals who captured and/or published the images. Provided that this activity does not occur *on `reddit`* (where it would constitute a breach of the rules) there doesn't seem to be any logical reason why it should be decried unless one believes that the rules of `reddit` should be enforced beyond the website's domain. This is likely also the reason why the 'Predditors' group who sought to name and shame the submitters of creepshots did not co-ordinate on `reddit` but rather on `Tumblr`.

It is hypocritical to allow /r/creepshots users to post sexualised images of women taken without their consent or knowledge and to disallow the identification of these users. In applying its pre-existing rules on these matters this is exactly how `reddit` (the company/management/administrators) 'acted'. However, it cannot be said that `reddit`'s administrators or *collective* user-base endorsed this as a morally legitimate stance - some users may have argued that this was the case but they were likely a small minority, and `reddit`'s administrators, to my knowledge, never spoke publicly about this story.

The second form of hypocrisy which is evident in this case relates to '`reddit`'s' response - the banning of links to Gawker (Guardian, 2012b). This is a form of censorship and it is obviously hypocritical for a website which allows a community like /r/creepshots to exist because of 'free speech' to then issue a punitive blanket ban for links to a journalistic website. Here we must however acknowledge that this ban was not implemented by `reddit`'s owners/administrators (although these appear to have banned links to the specific article for a short period because it contained a user's personal information) or `reddit`'s collective user-base but by small groups of moderators on their own subreddits. The reasons for this ban most likely include Adrian Chen's previous history on `reddit` and the fact that Violentacrez was seen to have been targeted (subsequently losing his job and being widely vilified) because he was a moderator - thus other groups of moderators saw this course of action as justifiable. This ban was however often criticised by 'ordinary' `reddit` users as hypocritical and counter-productive.

The mass of 'ordinary `reddit` users' do not feature directly in this story and their role is confined to discussing the rights and wrongs of the events. These discussions take place on `reddit` and therefore tend to focus on what the developments mean *for `reddit`*. Reddit's users value the relative lack of rules governing their interaction on the website. Two of the most important rules that `reddit` has are that users should not post illegal content and that users should not post the personal information of other users - and these rules are generally accepted as being necessary. Where `reddit`'s 'ordinary users' speak out in favour of allowing a subreddit like /r/creepshots to exist it is from the perspective of not wanting `reddit` to begin adding rules which govern the type of content that is acceptable or unacceptable on the website - for these users the issue is one with a strong political dimension. Furthermore, many `reddit` users do not support this position and would be happy for subreddits dealing with morally dubious content to be removed - for these users the issue is more of a practical one, they do not want the website they use to be associated with the content on /r/jailbait or /r/creepshots.

The actions which have defined '`reddit`'s' role in this story instead originate with a few small groups; the small number of /r/creepshots subscribers, the administrators who decide which subreddits are allowed to exist, and the moderators who decided to ban links to Gawker on their own subreddits.

`Reddit`'s administrators adopt a hands-off approach to subreddits in general and their position with regard to dubious subreddits appears to be that if these are not explicitly breaking the law (as it stands in the US) they will be allowed to persist. However, in both of the cases discussed here `reddit`'s administrators appear to have changed their position on whether a subreddit should be allowed *when that subreddit began to draw a significant volume of negative attention to `reddit`*. As such, their status as 'defenders of free speech' is compromised and their policy with regard to dubious subreddits begins to appear as characterised by cowardice and/or laziness. This also sends mixed messages to `reddit`'s users, with the implication being that subreddits catering to any type of content, no matter how morally dubious it might be, will be allowed *unless* the subreddit begins to show `reddit` in a negative light to an external audience. If `reddit` continues to grow in the future it will need a more robust set of rules about the types of content which are allowed or incidents like these will continue to occur as 'troll' users find new ways to push the boundaries of provocation.

The /r/creepshots saga is also interesting from a perspective of `reddit`'s long-term development. This case is distinct from all previous cases where `reddit` was 'making the news' in that the story was not prominently 'covered' on `reddit`'s Front page. Only a handful of posts in relation to this story appeared on the Front page until the Gawker article was published and some subreddits banned links to Gawker. The posts appearing prominently at this stage tended to focus on the Gawker article and the banning of links to Gawker. I myself found out that this controversy was unfolding only when someone showed me an article in *The Guardian*. Previously when `reddit` had been featured in the conventional media this tended to be the subject of many Front page posts/discussions and this coverage would begin almost immediately upon the story first breaking.

This could be due to either the way in which `reddit` has been changing or the nature of this particular story. It is possible that as `reddit` has become more fragmented and established users have moved away from the Front page, redditors no longer tend to see news about '`reddit`' (but a part of `reddit` that they don't use) as relevant to their interests. There is also the issue of which default subreddit coverage of this story would belong in - the /r/reddit.com subreddit would have been the most likely subreddit to carry this kind of news in the past, but since it has been closed there is no obvious default subreddit where this story would belong. Following the trend towards increasingly specialised subreddits, there is now a small subreddit devoted to external articles about `reddit` (r/redditinthenews).

It is also possible that this story did not feature prominently on the Front page because users' feelings on the subject were mixed - if opinion was divided then posts about the topic would tend to receive many down votes and therefore would be unlikely to be displayed prominently. Users may also not have had a clear opinion on this story, disapproving of /r/creepshots but not wanting `reddit`'s administrators to increase its legislation on which types of content were allowed.

Finally, gender has likely played a subtle role in how this case unfolded. `Reddit`'s users are mostly male (see section 6.1) and there is a clear gender dimension to this case whereby the subjects were female and the actors male. Gender was never explicitly discussed in any of the posts or comments which were observed, but it is possible that a user population with a more even mix of males and females would have identified /r/creepshots as a problem more quickly and more vigorously. This is not to say that `reddit` is sexist (although there are parts of it which certainly are) but perhaps `reddit`'s male users did not realise the gravity of the situation because being featured on /r/creepshots was not something that could happen to them.

8.10   Eternal September

For as long as this research has been underway some of `reddit`'s users have been complaining about a degredation in the quality of its content and user-base. The root cause of this problem, as with any formerly niche Internet location which grows into the awareness of the general population, is perceived to be an influx of new users. One term for this is the *eternal September*, originally coined on `Usenet` in reference to September 1993 when it was perceived that the influx of new users from `AOL` had degraded standards of discourse and behavior on `Usenet` and the Internet in general. This is a term which one often sees on subreddits where long-time users congregate - users who may have seen the same process unfold in a number of discussion spaces they formerly frequented. In these subreddits it is almost taken as a given that `reddit`'s eternal September started a long time ago, although the precise date when this began would likely differ in the estimation of each user dependant on when they themselves 'discovered' `reddit`.

The crux of the problem is a feedback loop of sorts. When new users flock to any discussion-based website it is inevitable that their participation will change the nature of that space - particularly if they have learned of the website en masse by its being featured in some other media (thus ensuring that these new users share something in common that is not the website they are entering). Pre-existing users may like the direction the the website goes in but there tend to be a significant (and vocal) number who do not appreciate the nature of this change. These users will each have some degree of change they can tolerate before the website's nature has changed to the point where they no longer want to participate - and go elsewhere. Once pre-existing users begin dropping out this marks the beginning of an irreversible trend because their lack of participation shifts the balance still further towards the new, pushing more pre-existing users out and establishing an endless cycle which can persist through whole 'generations' of users. This is the essence of eternal September.

With regard to `reddit`'s growth, the website served 1 billion pageviews to 13.7 million 'unique visitors' in January 2011 (Reddit-Blog, 2011a), increasing to 2 billion pageviews and 34.9 unique visitors in December 2011 (Reddit-Blog, 2012) and as of February 2013 `reddit`'s 'about' page

was quoting figures of 4.2 billion pageviews and 56 million unique visitors for the previous month. During this time there have been many events which exposed `reddit` to a new audience and as a result new users have probably arrived in large numbers from the same source.

`Digg`'s upgrade to version 4 of its software triggered one such event when `Digg`'s users revolted and filled its Front page with links directly to `reddit` (Reddit-Blog, 2010a). Before this `reddit` and `Digg` had been of comparable size and `reddit`'s users frequently made posts which were critical of the content on its better-known rival (the most frequent criticism being that content from `reddit`'s Front page would appear on `Digg`'s Front page hours or days later). The migration of `Digg` users to `reddit` was marked by a large number of prominent posts from 'ex-Diggers' about their perception of `reddit` in comparison to `Digg`, and by posts about what this migration meant for `reddit`. For some `reddit` users this event likely marks the start of `reddit`'s eternal September.

More recently, the /r/IAmA subreddit has likely been drawing in new users in large numbers on the occasions where well-known individuals have participated in AMA interviews. This is because these individuals often publicise the AMA post elsewhere (e.g. `Twitter`) and so 'fans' of the individual will come to `reddit` to read or participate in the AMA interview because these are usually interesting and ask questions which are distinct from those asked in typical interviews. A good example of this (from the perspective of drawing in new users, not answering interesting questions) is Barack Obama's AMA post (redd.it/z1c9z) which he also tweeted about - the influx of traffic for this post brought `reddit`'s servers to a grinding halt despite their capacity having been expanded by 20% in preparation for the post (Reddit-Blog., 2012). The blog post about this also reveals some interesting information about attention on `reddit`, namely that at any given time the Front page accounts for 15-20% of all traffic on `reddit`. The Obama AMA post was the first in `reddit`'s history to draw more traffic than the Front page itself - and the blog post reports that very popular posts rarely exceed 5% of `reddit`'s traffic in total.

A common effect of this kind of growth coupled with a perceived decrease in quality of content is for an out-migration of users to occur. People who liked the website the way it was in its earlier stages will often leave the website in favour of a smaller website that serves a similar purpose - this can result in the fracturing of the community with large groups breaking off and moving to different websites. For `reddit`, 'Hacker News' (`www.news.ycombinator.com`) is one such similar website - it is much simpler than `reddit` (having no equivalent to subreddits), has a smaller user-base, and deals only with ITC-related news.

In `reddit`'s case there is no evidence of a mass exodus of users thus far. This is in one sense surprising, because `reddit` does not have any obvious 'lock-in' factor. For example, `Facebook` has a kind of 'lock-in' effect because people use it to communicate with other people they know - if one's friends all use `Facebook` then it is not feasible for an individual to decide that they would rather use a different Social Networking site because the people they wish to communicate with

do not use this other site. In `reddit`'s case users could easily decide to use a different Social News website and all they would lose is their history of participation and karma scores - if users could find better content on another Social News website one imagines that they would have little hesitation in using that website instead.

The key to why this has not occurred is `reddit`'s open and federated system of subreddits. If a user doesn't like the content which is appearing on the default Front page any more they can unsubscribe from the default subreddits and this content will no longer be part of their `reddit` experience. Similarly, if a user thinks that the quality of discourse on a subreddit they value is decreasing they can easily start their own rival or complementary subreddit. Many of the default Front page subreddits which deal with 'news' now have exactly such a lower-activity counterpart. For /r/worldnews there is /r/worldevents or /r/news, for /science there is /r/hardscience (and many discipline-specific subreddits), for /r/gaming there is /r/games or /r/truegaming. There are also whole subreddits dedicated to showcasing these subreddit's with more 'Depth' (/r/DepthHub). One of these is /r/TrueReddit - there are a whole slew of 'true' subreddits (e.g. TrueMusic, TrueFilm, TrueAskReddit) and the intimation from this naming convention is clear - that the default subreddits no longer show high-quality content and these smaller subreddits are where the 'true' `reddit` experience (i.e. more like it used to be) is to be found.

On `reddit` a continuous migration of users away from the default subreddits is undoubtedly occuring - and at the same time new users are coming to the website in droves, so that despite this out-migration the default subreddits maintain a high level of activity. The displacement of established users by new users on the default subreddits started a long time ago and the 'quality' of these subreddits has been in decline ever since as users who prefer 'high quality' content are continuously 'pushed out' by sliding standards. If these users were migrating to a different website one might assume that this other website would come to be seen as the new 'hip location' on the Internet and `reddit`'s popularity would eventually wane.

However, this migration seems to be occurring largely *within* `reddit`, users are abandoning some subreddits in favour of others rather than abandoning `reddit` entirely. Discussions have been observed in which redditors talk about an eternal September *for specific subreddits* and the possibility that a subreddit invasion could trigger such a phenomenon. If `reddit` continues to develop in this way there will be further 'layers' to this phenomenon. The formerly small subreddits to which users fled from the hordes on the default subreddits are themselves growing in size and there is nothing to prevent the same processes from unfolding again once some users start to perceive a decline in the quality of these formerly small subreddits. There are already subreddits like /r/TrueTrueReddit and /r/TruerReddit, indicating that /r/TrueReddit is no longer 'true' enough for some users.

If the disparity between the default Front page and the smaller subject-specific subreddits continues to grow this will cause some problems for `reddit`. Users who might find some of these smaller

subreddits very interesting or useful may not realise they exist. It is easy to invisage a user who types '`www.reddit.com`' into their address bar for the first time now (2013) being discouraged by the content they are presented with on the default Front page and leaving with the impression that `reddit` is a website about (increasingly poor) jokes, image macro memes, cute animal pictures, etc. At present there are sometimes one or two 'interesting' posts on the default Front page (in my personal opinion) but the frequency with which these appear seems to be dwindling. `Reddit`, as a whole, needs the default Front page to display some indications of its 'deeper' regions if new users are to persist long enough to discover the thriving 'subreddit scene' - or failing this new users are only going to reach these 'deeper' regions if they are directed there by pre-existing users. However, as was shown with the example of /r/AskScience not all subreddits are suited to being added to the default list - so mixing up the default subreddits is not likely to be a suitable approach to addressing this problem. In this context having /r/BestOf as a default subreddit which accepts only links to comments on non-default subreddits is likely beneficial to `reddit` as a whole (if one assumes that continued knowledge and growth of non-default subreddits is desirable).

What `reddit` needs for sustained and productive growth now is a better facility for subreddit discovery - something which its users have been requesting for years. `Reddit`'s administrators have done little in this regard aside from talking about non-default subreddits in blog posts, nominating a 'subreddit of the day' and having some of the annual awards (where users vote to determine the winner) relate to subreddits. These approaches are themselves problematic because they are likely to lead to a rapid influx of new users (to the subreddit of the day or which was mentioned in a blog post) over a short period of time - it would be better for the health of these subreddits if new users joined them at a steady rate. `Reddit`'s users have been attempting to create resources to aid subreddit discovery (redd.it/8lon3 is one post where users have compiled a list of these resources) with some success - the problem here being that these resources are discussed in areas of the website which new users are unlikely to see (e.g. /r/TrueReddit).

To address these problems `reddit`'s administrators should 1) prominently display some information about the system of subreddits on the Front page so that new users understand this aspect of the website more quickly, and 2) link to some tools to aid new users with the discovery of subreddits which interest them.

8.11 Chapter conclusions

This Chapter has considered `reddit`'s evolution over the course of the research and described a number of longitudinal trends and important developments. In some cases it has been possible to analyse empirical data which provides evidence in support of the descriptions put forth - in pursuing other themes this has not been possible and instead casual observation and anecdotal

evidence have had to suffice.

The Chapter began by looking at metrics related to individual subreddits over the course of the research. These analyses showed that `reddit` has been moving towards content which is better described as entertainment than news. Posts from subreddits which deal with this type of content have become more prevalent on the Front page, and there is evidence which suggests these subreddits have seen a greater increase in their activity levels.

This Chapter also considered the effect of being a 'default' subreddit (taking /r/askscience as an example). There is undoubtedly a sharp rise in the level of attention which content on a subreddit receives at the moment when it becomes a default subreddit - in the case of /r/askscience this appears to have had a much greater effect the first time it occurred than on the second occasion.

There also appears to be a secondary effect whereby the popular content on that subreddit becomes more similar to the type of content which is popular on other default subreddits. This most likely occurs because of a commonality between users who are active on default subreddits in terms of their opinion on what constitutes 'good' content, this leading to voting behaviour which follows particular patterns. A non-default subreddit has the potential to cultivate a particular style of discourse because its users must find and subscribe to the subreddit for themselves. When a subreddit becomes a default choice this 'barrier to entry' disappears and it is instantly presented to the mass of users who consume content through the default Front page. In contrast to the earlier users who first sought out the subreddit and decided for themselves that they wanted to subscribe - these 'Front page' users may be less likely to know or care about the specific style of discourse which the subreddit is intended to foster, and the specific set of rules/guidelines which have been put in place to achieve this.

This case also highlighted the increasing centrality of moderators on `reddit`. The moderators of /r/askscience were in fact the major actors in this story - they protested the changing nature of popular content on the subreddit vociferously, they deleted content which did not match their conception of what the subreddit was for, and ultimately they had it removed from the set of default subreddits. This, and a number of other examples which have been recounted, belies a change in how `reddit`'s 'mass of voting users' are perceived by a (growing) contingent of users. Where once this broad collective was celebrated as the driving force behind `reddit` and the reason it was good - it is increasingly seen as problematic or even 'dumb'. Direct moderator/administrator intervention used to be viewed with suspicion, sometimes being seen as tyrannical - in many cases it is now seen as necessary to preserve those areas of `reddit` which are 'deep' and prevent their slide towards content which is shallow or inane.

Another strong theme to `reddit`'s longitudinal development has been the fragmentation of its userbase. `Reddit` is becoming an increasingly 'broad' website - in that it now has active subreddits

which cater to an immensely broad selection of subjects. Over this time the number of users who actively manage their subreddit subscriptions has increased, as has the tendancy for some users to un-subscribe from the default subreddits. In aggregate these individual choices have changed the very nature of the website. As users' subscription choices have diversified so has the content of their Front page, this no longer has the capacity to broadcast content to `reddit`'s user-base 'as a whole'. The concept of a single `reddit` community and identity has as a result been steadily eroded.

This is a significant development from a public discourse perspective (Sunstein, 2002). At the outset of this research `reddit` appeared to have the potential to mitigate information overload by allowing a very large group of users to collectively pick and sort their way through the vast assortment of content on the web - and broadcast these 'picks' to the community of users as a whole. If users are tending towards a highly specialised set of subreddit subscriptions this instead casts `reddit` as a site which could become a provider of a resource akin to the 'Daily Me'. There are two concerns of particular relevance here, 1) that users will filter out all information which does not match their pre-existing interests and attitudes and 2) that reddit's capacity to provide users with 'shared experiences' (and thus shared awareness/understanding) will be diminished.

While the general trend may be towards increasing customisation of subreddit subscripitons (implying fragmentation) there are some causes for optimism in this regard. Raw activity levels (votes/comments) for most default subreddits rose steadily over the full span of data collection (sections 8.2 & 8.2) - an increasing *proportion* of users may be abandoning these subreddits but they are by no means deserted 'ghost towns' and are in many cases thriving.

There are also a number of means by which users can mitigate fragmentation - even if their primary use of `reddit` is to obtain a Daily Me. The /r/All page is one such mechanism, showing the highest scoring posts from all subreddits irrespective of a user's subscriptions. Comments have been observed in which users discuss using /r/all to keep track of what's happening in the 'default' areas of `reddit` which they no longer frequent - if this behaviour is widespread then /r/all could serve as an 'antidote' to fragmentation on `reddit`. The fact that this behaviour and motivation is discussed at all is grounds for optimism and indicates that not all users are keen to embrace the Daily Me to the exclusion of all else - but perhaps switch between different 'modes of consumption' on `reddit`, using it to provide both 'general interest intermediary' and Daily Me content at different times. 'Meta' subreddits like /r/BestOf and /r/DefaultGems complicate this scenario further but may themselves offer alternative 'solutions' to the problem of fragmentation by allowing submissions which link to a wide array of subreddits and employing an additional layer of aggregation to this broad pool.

Research by Munson and Resnick (2010) suggests individual differences in whether users wish to encounter or avoid diversity in their information sources - with some users being 'diversity-seeking'

and others 'challenge-averse'. This, along with the casual observation of `reddit` users talking about /r/all and /r/BestOf in this way - suggests that there is likely a contingent of users on `reddit` who wish to receive information from the website in a very broad manner. It is regrettable that lack of access to relevant (procedural) data prevents the running of some simple analyses which would allow us to discuss fragmentation on `reddit` in much more concrete terms. With anonymous access to data on users' subreddit subscription choices it would be possible to quantify the degree to which fragmentation is taking place - and knowing how many users browse on the /r/all page, and how regularly, would put a number on users who seek an antidote to fragmenation.

`Reddit` is also becoming increasingly complex, and the presence of an increasing number of active 'meta' subreddits is a central factor in this trend. These 'meta' subreddits can be thought of as new and independent ways in which users' attention is being channeled. Without 'meta' subreddits it is possible to estimate the level of attention an item will receive based on the subreddit it was submitted to and the page(s) it appears on - the attention which a comment can receive is a function of its parent post's visibility and the comment's rank. Meta subreddits add new methods of directing user attention - a low-scoring comment to a low-scoring post on an obscure subreddit can now reach a sizeable audience if it is submitted to a meta subreddit and scores well there (e.g. such a comment could appear on the Front page through /r/bestof). Each of these meta subreddits has its own criteria for voting on content - and in some cases (e.g. /r/ShitRedditSays) it is the *worst* submitted items which are upvoted to prominence.

Meta subreddits which are themselves 'anti-`reddit`' suggest a user-base which is in conflict with itself. These, along with 'subreddit invasions' and 'voting brigades' are ominous signs because if there are already isolated communities which are hostile to one another this implies a level of balkanization which does not bode well for the future (Sunstein, 2002).

Some long-time users are disappointed with the direction in which `reddit` has evolved. They hark back to an era when `reddit` had an identity, broadcast through the Front page, which they identified with and could maybe even take pride in. As the Front page ebbed towards increasingly shallow content these users have deserted the default subreddits, they have been supplanted by new users who like this content. The Front page still broadcasts a `reddit` identity of sorts but for many long-time users one suspects that this is not an identity they take pride in. Personally, I barely recognise the default `reddit` Front page of 2013 as the same place where the case studies reported in Chapter 7 (conducted 2010/2011) unfolded.

To re-visit the newspaper analogy, `reddit`'s default Front page has become much more 'tabloid' in style while a 'broadsheet' edition can still be obtained by subscribing to non-default subreddits. The dischord between these styles is problematic for some users (imagine your favourite broadsheet sitting in the newsagent with the front and back pages of a tabloid - to a casual observer who doesn't also read that newspaper it is these pages which identify the newspaper you are buying).

It is possible that this kind of 'internal conflict' on `reddit` is largely symptomatic of a transitional phase, and that this will subside once the process of fragmentation is 'complete'. Indeed it would not be entirely surprising if in future `reddit` abandoned the concept of a default Front page in favour of an alternative entry point that emphasised its diversity and guided new users towards specifying what they wanted from the website. This would of course have major implications for `reddit`'s role in the Hybrid Media System (Chadwick, 2013) and would change its impact on the fragmentation of public discourse entirely (Sunstein, 2002).

Chapter 9

Discussion and Conclusions

This research has considered `reddit` and `Stackoverflow` as two examples of 'Distributed Moderation' principles being applied to meet different ends. These websites have been considered along the following criteria:

- How do the websites 'work'? What purpose does the voting system serve and how are the actions of individuals aggregated to produce an end resource? Chapter 5

- Why and how do individuals participate? Is there a standard form that this participation takes or do individuals specialise in different types of activity? Chapter 6

- What are the broader social implications stemming from websites which utilise DM heavily? In the context of Social News how does this diverge from conventionally produced news? Chapter 7

- How have the websites under scrutiny changed over time? As these websites have been experiencing rapid growth this is also a study of how these systems scale up and how increasing levels of activity affect their functionality. Chapter 8

This research has been exploratory in nature - when it began there was very little published research on Distributed Moderation and so the research has attempted to build an understanding of these systems without the benefit of a body of existing research. As a result, there is a substantial methodological component to the research - a variety of methods have been devised and those which provided useful outcomes have been detailed in this thesis. One might say that the most fundamental research question has gone largely unstated - what can we learn about these systems through analysis of the data which they naturally collect and store in their day-to-day operation? Here a distinction must be made between data which is publicly available and can be extracted through an API, and data which exists on the websites' servers but is not publicly accessible and can only be accessed with the cooperation of the website's administrators. This research has

utilised data primarily of the former type (with one month of post voting activity from `reddit` being the only exception), but later in this Chapter I will also discuss some of the types of question which could be addressed if one had unfettered access to procedural data.

One of the themes which has emerged from the research concerns the difficulty in making direct comparisons between different websites which use DM - in this case `reddit` and `Stackoverflow`. The problems with making such comparisons are two-fold. Firstly, these websites are structured in very different ways and their voting systems serve different purposes - for `reddit` up/down voting is its defining characteristic and is absolutely central, whereas `Stackoverflow` utilises a much broader set of 'Social Media tools' including tagging and Wiki features, with up/down voting playing more of a supporting role. Secondly, although both of these websites are fundamentally public and many types of data are publicly accessible - there are significant differences in exactly what this data represents and one often finds that differences in the availability of data limit the extent to which comparisons can be conducted.

These factors also limit the degree to which the results of this research are generalisable. Most of the findings of this research are specific to the websites which have been studied - for instance, the investigation of `reddit`'s functionality cannot be assumed to tell us anything about other Social News websites because differences in the algorithms used to convert votes into ranks, differences in the pages used to display content, or differences in how individuals participate could lead to very different outcomes. This is however not as great a problem now as it was at the start of the research when `reddit` was one of a few Social News websites with comparable levels of activity - with the decline of `Digg`, the stagnation of `Slashdot` and the failure of other sites (like 'Hacker News') to achieve widespread recognition, `reddit` is now arguably *the* Social News website.

`Reddit`'s rapid growth does however cause problems of its own for this research, stemming from the fact that social aspects of the website appear to have changed considerably - although its software and structure have thankfuly remained largely unaltered. There are two questions one must ask of the findings related to `reddit`.

Firstly, are they are still applicable or are they outdated? This is particularly relevant for analyses based on the procedural data dump of March 2009 (Chapter 6 used this data almost exclusively) because it has not been possible to obtain similar data from a later time period. Since March 2009 the number of people using `reddit` has increased substantially and the type of content which the website displays on its Front page has changed considerably - it is quite possible that users' 'participation profiles' have also changed during this time. The analyses in Chapter 5 were initially conducted on data collected in 2010 (not reported) but have been repeated and expanded with data collected in 2012 (with the results being quite similar) - therefore these analyses are much less likely to be outdated.

Secondly, the research on `reddit` has primarily considered the activity on large default subreddits and the default Front page. At the beginning of this research `reddit` seemed to be largely centred on the default subreddits - there were relatively few non-default subreddits with even moderate activity levels and many of the discussions belied the sense that `reddit`'s users were all looking at more or less the same Front page. Over time this situation has changed and in comments one is now more likely to see users talk about un-subscribing from default subreddits than something they saw on the Front page last week. The expansion of the default list to 20 subreddits in October 2011 has likely been a pivotal point in this process. The issue which this fragmentation raises is whether results are generalisable between subreddits - Chapter 5 reported notable differences between the 'behaviour' of highly active and moderately active default subreddits. It is likely that smaller subreddits behave differently in still further ways. For instance a lower volume of submissions means less 'time pressure', one will often see relatively old posts near the top of smaller subreddits - it also means that proportionally more posts will appear on the 'high visibility' parts of the subreddit (i.e. its Main page).

There are also an increasing number of 'curated' subreddits - subreddits which are intended to host a very specific style of discourse and which are (sometimes heavily) moderated to achieve this - with content deemed to not meet the subreddit's ideal/standard being deleted or censored by the moderators. This suggests that either 1) something has 'gone wrong' with `reddit` as its number of users has grown, or 2) its voting system and structure is best suited to a particular type of content and style of discourse - if one wants their subreddit to cater to an alternative type of content one must either hope that a great majority of the users who subscribe are of like mind, and/or start moderating the subreddit to exclude undesirable content.

Indeed the increasingly central role played by moderators suggests that `reddit` is moving away from the ideal of democratically mediated discourse. Moderators are generally not 'elected' (although there are several instances where a subreddit's community have held a vote to pick new moderators) and `reddit`'s software makes no provision for doing so. Instead it is a subreddit's creator who becomes its first moderator and has the power to appoint additional moderators. If one thinks about this as a form of governance the 'direct democracy' which was formerly central to `reddit` is being relegated to a secondary role behind the (hopefully benevolent) dictatorship of moderators.

Let us now consider what this research can offer by way of answers to the research questions which have guided it.

9.1   Discussion of Research Questions

The broader questions which have guided the research (i.e. How does Distributed Moderation work? How and why do users participate? How does the presence of DM affect the nature of

the resource which is produced or the interaction which takes place?) will not be discussed here because there are whole chapters which constitute the answers to those questions (Chapters 5, 6 and 7 respectively).

Instead the more specific research questions raised in section 2.12 will be re-visited here. Many of these questions are only relevant to `reddit` and the research has tended to focus largely on `reddit` because it is conceptually more interesting than `Stackoverflow`.

A) How well do Distributed Moderation systems fulfill their task of picking the 'best' items of content to place in the most highly visible locations?

This question has been addressed with respect to `reddit` because picking the 'best' items of content to place in the most highly visible locations is not one of `Stackoverflow`'s main objectives. This research has rarely attempted to assess the 'quality' of material on `reddit` by any objective standards - given the scale at which the website operates and the breadth of content types it deals with any meaningful assessment would have required a great investment of time and this time was spent pursuing other lines of enquiry. The qualities of individual items have been discussed when and as they relate to these lines of enquiry (Chapter 7). However, for most of this research it is 'quality' through the eyes of `reddit`'s users which is of interest - and the specific characteristics of this 'quality' are bound to vary significantly between different subreddits. Thus for sections like the present one a concise but useful definition of quality might be "is it worthy of an up-vote"?

The central problem in addressing this question is that the only measure of an item's 'quality' on `reddit` is its score, and this is intimately tied to the locations in which the post has been displayed. A content-rating experiment (section 7.3) was designed which would generate an 'external' measure of a post's 'quality' (external to `reddit`'s voting system, but deliberately recruiting participants who were `reddit` users). The results suggested that there was a relationship between a post's performance on `reddit` and the ratings it received in the the experiment, but that this relationship was not particularly strong. Posts which appeared on the default Front page received on average slightly higher ratings than those which were observed on the Main page - the fact that these posts achieved much higher scores on `reddit` is *due to their appearance on the Front page* and not to any massive disparity in 'quality'. There were also several items for which experimental ratings did not match their voting performance on `reddit` at all - 'poor' posts (low ratings in the experiment) which appeared in high-visibility locations with high scores, and 'good' posts (high ratings in the experiment) which received low scores on `reddit` and were only seen by a few users.

Given the previously conducted explorations (see Chapter 5) this 'random' element was anticipated. `Reddit`'s 'early stage' pages have relatively low levels of voting activity and on highly active subreddits (like /r/pics, the subreddit which experimental items were drawn from) they see a

high rate of post submission. The result is that posts have a very short window before they are pushed off the New page in which they must begin to accumulate points and earn placement on subsequent pages - and given that the voting rates for these pages are low this decision as to which posts proceed to the next 'stage' will be made by a relatively small number of users.

Section 5.1.5 outlined a 'hurdling' or 'filtering' process which seems to describe the passage to the Front page well for highly active subreddits. Posts which ultimately reach the Front page must first earn a position on the Rising page, then the Main page, and then the Front page. Each of the pages in this sequence brings with it a significant increase in voting rate, the result is that a post which has progressed to a given page (let's say the Main page) is effectively only competing with the handful of other posts that are concurrently making their way up the Main page ranks. This account of the 'passage to the Front page' allows for the possibility that 'poor' posts will appear in high-visibility locations because 1) they are effectively only competing with the relatively small number of posts at a similar stage and 2) the decisions about which posts progress past the first stages are being made by very few users and very quickly.

Section 5.1.5 also describes an investigation into whether 'poor' posts (in this case defined as those posts which had a final observed score on `reddit` which was negative) could appear on the Main page for a number of subreddits. This was found to happen much more frequently on subreddits with relatively low voting rates on their early-stage pages - indicating that these voting rates can be taken as a measure of the subreddit's 'health'.

More recently, Gilbert (2013) has explored the phenomenon of posts which failed to attract a substantial voting response on `reddit` upon first submission but which went on to appear on the Front page when re-submitted at a later date. This paper reaches the conclusion that `reddit` suffers from an under-provision of votes on its New pages - a conclusion also borne out by the present research, although arrived at through a different approach.

Gilbert (2013) see instances where a post scored badly at first but later appeared on the Front page upon re-submission as evidence of underprovision of votes, but the analyses presented in Chapter 5 depict a system which is is inherently 'unstable'. The transition from New to Rising page is a key point in the lifespan of a post and it occurs (or does not) very quickly after the post's submission, based on the direction of the first few votes which are cast on the post. The analyses conducted as part of this research indicate that an impractically large number of users voting on the New page would be required for `reddit` to produce consistent judgments on posts. Greater consistency might instead be achieved by altering the structure of the site such that the New and Rising pages share less overlap - in particular, posts are pushed off the New page too quickly and can appear on the Rising page too easily (voting rates there are higher than on New [1]), so the New page is

---

[1] At the time when this research was conducted the New and Rising pages occupied the same 'New' tab, and the 'Rising' page was actually the default sub-tab when a user clicked 'New'. These two pages now have seperate tabs

relegated to a role of quickly filtering out 'bad' posts which won't appear on Rising.

This is understandable in `reddit`'s present configuration because the New page for a busy subreddit itself moves at such a fast pace - when new posts are being submitted every few seconds a post will have only minutes on the first leaf(s) of the New page. The frantic speed at which this page moves is one of the primary causes of the issue reported by Gilbert (2013), and therefore it is likely that this could not be easily remedied by simply having more people voting on the New page. When the decision about which posts to 'filter out' on the New page is being made in a matter of minutes there will inevitably be a random component to this decision.

If `reddit`'s administrators wished to make its voting system more consistent they might instead think about re-structuring the early stage pages entirely. When a user clicks through to the New or Rising pages they are implicitly 'volunteering' to pitch in their efforts where they are needed most. At present the New page shows submissions from the very recent past, and Rising shows a more stable set of items but these have already been filtered (rapidly so by necessity, and heavily so on busy subreddits). Users who 'volunteer' in this manner could instead be shown a random sample of items submitted within the last hour, and this would spread early votes more evenly between recent submissions and thus produce more reliable early-stage decisions. At present too much of a post's ultimate success on `reddit` rests on being seen within the very short interval before it is pushed off the New page (posts are often 'dead' within minutes) - and where a post is followed quickly by submissions from other users it is unfairly disadvantaged.

B) Social News websites focus a large amount of their users' attention on a very small sub-set of the items that are submitted. What are the consequences of this from a public discourse perspective?

> Do Social News sites offer a means of deriving shared experiences from the web for their users? Or are they likely to lead to further fragmentation of public discourse? (Sunstein, 2002) The presence of a default Front page where a small set of items judged to be 'the best' are displayed to all suggests a unifying effect. However, on `reddit` the proliferation of many specialised subreddits means that users who manage their subscription choices may see very few of the same items - indeed a highly specialised set of subreddit subscriptions would be a close conceptual match with the Daily Me. Can users configure a Social News account such that information they would not have sought out and which they have no interest in is entirely filtered out of their experience on the site? Do they choose to do so?

The presence of a 'Front page' is one of the defining characteristics of Social News, on par with

---

and as a result there may now be greater parity in the levels of voting which occurs on these pages.

the voting system in its level of importance. Where the voting system generates ranks for content it is the 'Front page' which effectively broadcasts the items which make it to the top of their respective categories. At the beginning of this research the default subreddits had much higher levels of activity than non-default subreddits and there were very few non-default subreddits with a reasonable amount of activity (e.g. in March 2009 there were only 35 subreddits which received 10,000 votes or more while the larger subreddits had upwards of 300,000 votes). On the basis of this and observations of redditors' discussions, it was assumed that the majority of users were subscribed to most if not all of the default subreddits. In this scenario the default Front page served the function of broadcasting the currently most up-voted and active posts to most of `reddit`'s users - in the process providing 'shared experiences' (Sunstein, 2002) which work against fragmentation.

Broadcasting the most popular items to 'all' `reddit` users in this way allowed for a reasonably strong sense of community to emerge among users - surprisingly strong for a website which already had upwards of 100,000 active users. It was hypothesised in Chapters 3 and 5 that by focusing the attention of users on a small sub-set of posts users could have a shared understanding of 'what's happening on `reddit`' and this in turn defined `reddit`'s 'identity'. The capacity for users to influence the decisions about what appeared on `reddit`'s Front page (and consequently `reddit`'s identity) through the voting system was cited as likely being an important motivator for users to participate (particularly by voting). The evidence for this sense of community is however largely anecdotal - observations of discussions among users at the time which referenced `reddit`'s 'identity', and the capacity for users to rally around a cause.

Over the course of this research the sense of a single `reddit` 'identity' and community have been steadily eroded as its number of users has rapidly increased. `Reddit` has transitioned from a website with 35 active subreddits to a website with a whole panoply of active subreddits devoted to individual countries, states, cities, universities, sports, teams, games, scientific disciplines, etc. This raises a distinct possibility that users can tailor their subreddit subscriptions extensively and largely filter out 'information which they would not have sought out in advance' from the feed they receive through `reddit`.

In addition to this 'natural' fragmentation as the range of choices have been expanded there is evidence of a schism (or schisms) in `reddit`'s user-base. Central to this schism is the perception that 'new' users are different to the users which had previously been using the website (age is widely seen as being key to this difference), leading to a change (some would say decline) in the nature of the content which appears on the default Front page. The result is an exodus of older users from the default subreddits, pushing these further towards entertaining or shallow content. The quintessential '`reddit` experience' for long-time users now involves selectively subscribing to only those subreddits which they deem worthy of their attention. There are still a large number of users active on the default subreddits, and probably a considerable number who still view the

default Front page - but these are, it seems, not generally long-time users, and they are in fact looked down on by many long-time users.

Writing now in 2013 it is more accurate to describe `reddit` as composed of many different communities (subreddits), and this is how `reddit` currently describes itself. The voting system and 'hot' ranking algorithm perform the same role for each of these communities, the items which are ranked most highly in subreddits will appear on the Front page for users who are subscribed to them, and users of a particular subreddit still have a shared understanding of what's happening on that subreddit. However, the degree to which all of these subreddits are aggregated to produce a shared Front page has been substantially diminished.

Relating this back to the literature - one can describe this as a move towards the realisation of the Daily Me (Sunstein, 2002). `Reddit` formerly defied this, acting as a means of dealing with Information Overload by centralising users' attention on a small sub-set of items 'picked' from the Web by user submission and voting activity. As `reddit`'s user-base becomes more fragmented `reddit` itself becomes more like the 'Daily Me'. It is a news resource but one where users select the 'streams' of information which they want to see, if this trend continues to its logical conclusion `reddit` will be a very close match with the 'Daily Me' concept.

As has often been observed, it is not the nature of a communications technology itself which will determine its impact on public discourse but the way in which people use it. Over the course of this research `reddit`'s role in public discourse has arguably transitioned from a unifying force to one which facilitates fragmentation - with this change being driven by changing patterns of user behaviour. The question of how fragmented `reddit`'s user-base is could be addressed through analysis of users' subreddit subscriptions - analysis of how this has changed over time would be particularly illuminating. Unfortunately these records are not publicly available, although they are likely in existence within databases on `reddit`'s servers. Without access to this data we can only speculate on how fragmented `reddit`'s user-base and discourse have become.

There are however some reasons for optimism on this front (sections 8.5, 8.7.2 8.11). The addition of an '`/r/all`' page can be thought of in this context as an 'antidote' to fragmentation on `reddit`. This page is similar to the Front page in that it is an aggregated 'Hot' page - but in this case posts from any subreddit (excluding 'NSFW' subreddits by default) feed into the page and its top ranks are occupied with the top-scoring posts from `reddit` as a whole. Comment discussions have been observed in which users talk about browsing on /r/all as a way to find out what's happening on `reddit` beyond the set of interests they have defined through their subreddit descriptions. This in itself is indicative of users who recognise the value of the 'shared experiences' (Sunstein, 2002) which the default subreddits/Front page can provide. It would be informative to know how often the /r/all page is accessed - and whether users tend to access both /r/all and the default Front page. If /r/all is being regularly accessed this suggests that perhaps `reddit`'s users sometimes

want to take a break from the Daily Me and be presented with 'information which they would not have sought out for themselves' (Sunstein, 2002). The presence of the /r/all page certainly implies that this avenue is open to `reddit`'s users, and that depending on how they use the site it can provide them with a Daily Me or more of a general interest intermediary.

'Meta' subreddits like /r/BestOf may offer another remedy for fragmentation through the addition of a further level of aggregation. On this subreddit users submit posts which link to comments elsewhere on `reddit` and the 'Best of' these are upvoted to prominence. This subreddit formerly allowed submissions linking to any subreddit, and was dominated by comments on default subreddits. Section 8.7.2 describes the decision to exclude default subreddits from /r/BestOf - and in discussions around many of the top-scoring comments were from users who said that they had unsubscribed from default subreddits and used /r/BestOf as a means of keeping track of what was happening on `reddit` outside their determined areas of interest. The presence of users who unsubscribed from default subreddits (presumably because they did not find them interesting) but still wish to retain contact with this 'default' side of `reddit` might be grounds for optimism - perhaps the Daily Me is not something people will always choose even if it is available to them (see also Munson and Resnick (2010)). A subreddit like /r/BestOf also has posts covering a very wide range of topics appearing on its high-visibility pages, and a subreddit like this is bound to present all users with 'information they would not have sought out in advance'.

C) Do the decisions reached by DM systems accurately reflect the opinions of their contributing users?

This question can be thought of as having two components. Firstly, with regard to 'neutral' content (like /r/pics and /r/funny) it is similar to the question of whether the 'best' items are being displayed most prominently (discussed above). Secondly, with regard to content which adopts a position on a particular topic (which users can agree or disagree with) it is a question of whether the judgments made through the voting system accurately reflect the opinions of the website's (or subreddit's) users. This is difficult to establish because there is no straightforward way to gauge the opinions of a group of users without conducting a survey with a representative sample of these users. The ability to down-vote posts raises the possibility that the users holding the majority opinion on an issue could suppress dissenting views, and if only the dominant attitudes can appear in high-visibility locations this is likely to lead to an 'echo chamber' effect and group polarisation (Sunstein, 2002; Myers and Lamm, 1976).

This research has attempted to address this question with data collected from `reddit` in the context of the story about `Wikileaks`' release of US diplomatic communications (section 7.1). In this case a sample of all submitted text posts was taken as representing /r/politics users' opinions on the topic - each post's 'tone' towards its subject being coded as positive or negative. The

breakdown of opinions in these posts was then compared to the breakdown of opinions expressed by posts which appeared on the Front page. This comparison suggested that `reddit`'s Front page accentuated the difference in opinions to some degree - a greater proportion of Front page posts were pro-`Wikileaks` than the same proportion for new submissions unfiltered by the voting system. However, `reddit`'s users (through the voting system) did not entirely exclude the minority opinion (i.e. that the release of the cables was bad) from prominent display. The treatment of minority opinions is considered further below.

Here it is important to identify a problem with `reddit`'s voting system with respect to content on which users are likely to have relevant pre-existing biases. The voting system is one-dimensional (and 'bi-polar'), and this dimension is intended to reflect only whether users think the content is 'attention-worthy' or not. For instance comment votes are intended to reflect whether a comment is "contributing to the community dialogue or discussion" and the reddiquette guidelines ask users not to "Downvote an otherwise acceptable post because you don't personally like it". Many of `reddit`'s users believe that this guideline is routinely ignored, and although none of the available data allows for any kind of analysis which is relevant here casual observation supports their position.

Firstly, this guideline is asking users to do something which is itself difficult to achieve - to disentangle their opinion on a post/comment and seperate whether they think it is of high 'quality' and should be prominently included in the subreddit's discourse - from their own subjective opinion on the post (whether they 'like' it). This has close parallels with the positivism/antipositivism debate which is at the core of social science - with many scholars of a Feminist or Marxist persuasion rejecting the idea that they can, or should try to, adopt a 'neutral' perspective informed only by the available empirical evidence. If there are professional social scientists who reject the idea that this kind of dissociation between subjective opinion and objective understanding is even possible - is it then reasonable to expect `reddit`'s users to routinely do this when they vote on content? I would argue that it is not.

Secondly, if we were to accept that this kind of objectivity is achievable for `reddit`'s users... is this something that they would *choose* to do? A user's votes are private, nobody will be able to review a user's votes and determine whether they have adhered to this guideline or not - there is therefore no external pressure to do so. Given that `reddit`'s voting system serves to determine which content will be broadcast, will users really vote to broadcast well thought out arguments which go against their own opinions? If one is interested in propagating one's own opinions in addition to or instead of merely fostering open discussion, this would appear to be counter-productive.

One of the biggest problems with this guideline is that it leaves no room for the expression of opinion through the voting system. It is not possible to up-vote a post because it makes a good contribution to the discussion *while also indicating that one does not agree with the argument it is making.* In this scenario some proportion of users will always choose to vote in a way which

reflects their opinion, because they want to express their opinion and because when something appears prominently on `reddit` it is seen, rightly or wrongly, as being endorsed by `reddit`'s users. The ambiguity about what votes mean is even reflected on `reddit`'s pages - despite the guidelines which state that users should not vote based on whether they 'like' the content, on each post's comments page the percentage of votes which are up-votes is expressed as a percentage who 'like it'.

The uni-dimensional voting system clearly confounds 'attention-worthyness' with 'agreement'. For subjects which have an 'agreement' dimension (e.g. politics, science, answers to questions) the utility of the voting system could be improved by adding a second dimension which users can use to reflect their opinion. If users could express their opinion by voting on a second dimension they may be more willing and able to up-vote posts they disagree with but which are 'attention-worthy' - and it would be possible for this kind of post to appear on the Front page with a clear expression that most of `reddit`'s users do not agree with the point it is making.

The idea that systems which utilise user ratings could operate on more than one dimension is not a new one. Lampe and Garrett (2007) conducted an experiment with the users of `NewsTrust` in which participants rated stories using one of four instruments with varying numbers and types of dimension. They found that users showed the greatest satisfaction with a multi-item reviewing instrument, but acknowledge that as the number of dimensions is increased so too is the burden on users and the complexity of the resulting resource.

The confounding of 'agreement' and 'attention-worthyness' undoubtedly causes problems for `reddit` in terms of accurately portraying its users' opinions about content. There are some subreddits which would benefit from the addition of a second dimension on which users could vote to reflect 'agreement' - and that the increased 'burden' of this second dimension would be out-weighed by the benefit to understanding what users' attitude towards a piece of content is. If `reddit` could more accurately capture and portray its users attitudes/opinions this would likely lead to an enhanced capacity to foster collective action endeavours among other benefits.

D) Are the opinions and judgments of all users weighted equally? Or is it possible for individual users or groups to 'game' the system to induce their desired outcome? To what extent are highly active minorities or 'superparticipants' (Graham, 2013) present on DM websites, and what are the characteristics of their behaviour?

This question is similar to the previous one in that it concerns `reddit`'s ability to accurately reflect the opinions of its users. The previous question was about `reddit`'s voting system and its capacity to represent the opinions of its users - in that case an assumption was made that all users' opinions were being equally weighted and the question was one of how `reddit` represented a

genuine difference in opinions among users. This question is about whether the opinions of all users are in fact equally weighted - and about whether 'expert' users can behave in ways which lead to the disproportionate representation of their own opinions. Keen (2007) is particularly concerned that Social News websites are far from democratic and are in fact dominated by their own coteries of elie users. "For all their claims to be more democratic and honest, these supposedly editor-free social news sites are actually creating a more oligarchic and corrupt media" (Keen, 2007, p. 94)

Again, there are two components to the answer. The first involves users who 'cheat' by breaking the rules of the website, the second involves ways in which users might maximise their influence without actually breaking any rules.

Firstly, it may be possible for users to coordinate their actions and thereby wield disproportionate influence as a group - an individual could achieve the same thing if they were able to operate a number of 'sock-puppet' accounts without being identified as doing so and excluded. There are numerous stories of groups which have attempted to 'game' the Distributed Moderation systems employed by Social News websites - either for the purpose of filtering out content on ideological grounds (the 'Digg Patriots', (Halliday, 2010; Oleoleolson, 2010)), or for marketing purposes. From this perspective the 'early stage' pages are Distributed Moderation's weak point. On `reddit` the first barrier to reaching the Front page is progressing from the New page to the Rising page, and voting rates on these pages are relatively low. It is therefore relatively easy for a group to prevent posts they don't like from progressing beyond this initial stage by casting early down-votes - if they can do so without being detected by the website's anti-cheating mechanisms. This was the primary tactic of the 'Digg Patriots' who communicated outside `Digg` via e-mail and IRC and would inform each other when they spotted a new post which contravened the group's political perspective. Using this tactic the `Digg` Patriots successfully buried many posts which they perecived as being 'too liberal' until they were infiltrated and exposed by another `Digg` user.

The opposite side of this is trying to ensure that certain posts are featured on the Front page, and this seems more difficult to achieve. Several companies have attempted to sell front-page placement of an item on `reddit`, which they would achieve through the control of many user accounts (redd.it/elqgs ; (Woyke, 2010)); this kind of behaviour had also been observed previously on `Digg` (Warren and Jurgensen, 2007). The early stage pages can be targeted to this end but here a small co-ordinated group will only be able to ensure that a post progresses past the first one of two hurdles - when the post begins to appear on pages with higher voting rates the influence of a small group will be diluted. The much larger number of co-ordinated user accounts which would be needed to ensure that a post appeared on the Front page is likely much easier for anti-cheating code to detect and nullify. Here some Social Media marketing companies have purportedly attempted to recruit prominent users to submit and vote on their content, but the attempt failed when some of the high-profile users they contacted immediately posted about the offer they had just received.

Without access to comprehensive back-end data it is not possible to investigate whether these attempts to game Social News voting systems are ongoing or whether some groups might be successfuly doing so. Here we must rely on the word of the website's administrators. `Reddit`'s administrators have always expressed confidence that they were successfully identifying and nullifying attempts to game `reddit`'s voting system through the co-ordination of many accounts. While there is no reason to disbelieve them, one must also accept that their main concern is with their website *being seen to have this kind of problem* - even if they were aware that their voting system was being rampantly gamed they may not report this to their users. This kind of manipulation could also conceivably be occurring completely undetected by administrators.

With regard to the second component of this answer it is possible to conduct investigations and present some empirical evidence. Are there users who behave in ways designed to maximise their influence without breaking the rules (i.e. without co-ordinating with other users)? Again the 'early stage' pages are the obvious place to vote if one wants to maximise one's influence. All posts must progress through these pages and their low voting rates mean that an individual's vote has proportionally more influence in determining whether a post will progress to the next page in the sequence.

There is also no limit to the number of votes `reddit`'s users can cast in any given time period - and users can have more influence simply by voting more often on a greater number of posts. A greater number of votes can be achieved by spending more time on `reddit` voting and/or by voting more rapidly in the time that one spends on `reddit`. If one's goal was to maximise their influence they might well cast a large number of votes in rapid succession, scanning the titles of posts on a page and quickly voting on those posts which they didn't like the look of without ever taking the time to examine the content of the post.

From the perspective of `reddit`'s general user-base this is a profoundly anti-social behaviour, adding a significant volume of 'noise' to the voting system which decreases the impact of votes cast by users who take the time to read a post's content and make a reasoned decision. The knowledge that this is possible may even discourage users from taking the time to read posts and make reasoned decisions. Suppose someone takes five minutes to read an article and decide whether it is attention-worthy - during this time a user who scans the titles and votes every 2 seconds will have made 150 votes. By taking the time to make reasoned decisions one is implicitly sacrificing influence to people who don't, and awareness of this could precipitate a 'race to the bottom' where users strive to vote more quickly.

Section 6.1.4 looked for evidence that users were engaging in behaviours intended to maximise their influence. Quick votes were found to be unexpectedly prevalent, with 19.5% of all votes being cast within 10 seconds of the user's previous vote and 7.6% of votes being cast within 2 seconds of the user's previous vote (these figures being derived from the March 2009 back-end data-set which

excluded all votes nullified by `reddit`'s anti-cheating code). There was also a strong association between quick votes, early votes and down-votes - votes which met one of these criteria were likely to meet the other two as well. Furthermore, all of these vote types were cast disproportionately by `reddit`'s most active users.

All of this evidence points to a small set of highly active users who were engaging in behaviours which seem likely to increase their influence over which posts would fare well on `reddit`. These users could be described as voting 'superparticipants' Graham (2013) - and the nature of their voting behaviour is such that it is consistent with an intent to 'curb' the capacity for other users to participate in the discourse. This does not however mean that 1) these users were deliberately aiming to maximise their influence or that 2) these behaviours successfully increased the influence of these users. It has not been possible to test either of these hypotheses with the available data and time. However, in one related analysis (section 6.1.5) it was found that these behaviours *are not* associated with better performance when users who engage in them *submit posts of their own.*

`Reddit` also has a set of users who are highly active post submitters - but post submission frequency and proportion of activity were also found to be associated with less successful posts. On a discussion board these users could be considered to be 'super-posters' (Graham, 2013) - although such a concept on `reddit` is more likely to relate to comments. Given that posts on reddit must acquire voting points to be seen, these users take on the aspect of failed superparticipants - whereas on a conventional discussion board the frequency with which 'super-posters' submit content practically ensures that their contributions will be widely seen.

Section 6.1.6 investigated longitudinal data on reddit's Front page to determine whether there are users who can repeatedly submit posts which reach the Front page. This section allowed us to consider whether `reddit` is dominated by an 'oligarchy of elite users'. In support of this claim Keen (2007) cites a *Wall Street Journal* article which reported small groups of 'influencers' who successfully manipulated Social News - in Digg's case one third of observed Front page postings were submitted by 30 highly infulential users. In `reddit`'s case, over a two year period the 27 most successful post submitters accounted for 4% of Front page posts. While `Digg` was well known as being dominated by 'Power Users', `reddit` seems much less prone to the domination of an elite oligarchy. 55% of all the posts pbserved on the Front page during this two year period were submitted by a user who only achieved this once - implying that more than half the slots on `reddit`'s Front page at a given moment will be occupied by posts submitted by a user who had never before seen one of their posts displayed there. It would appear that the large volume of attention for posts which appear on the Front page is within reach of any user who submits a post, in practice as well as in principle.

There are however a considerable number of users who submitted several Front page posts, and a small minority who have repeatedly done so (see table 6.11). Of Graham (2013)'s typology, these

users are best described as 'agenda-setters' because their highly visible posts act as the basis for discussion (with these users' posts seeing a combined total of 1.25 million comments).

However, we must bear in mind that `reddit` is not a discussion board, and there may be users who have no interest in comments pages but instead consume only the content of the posts themselves - in this context the term 'agenda-setters' does not fully capture the role these superparticipants play on `reddit`. It would be of considerable interest to investigate these users much more thoroughly, to try and determine the origin of their apparant knack for submitting Front page posts (do they submit certain types of post? at certain times of day? do they have a lot of friends on `reddit`?). However, the data which would allow one to address these questions was not collected as part of the present data collection regime - and much data of interest here is not publicly available at all.

`Stackoverflow`, unlike `reddit`, does limit the number of votes which a user can cast (limited to 40 per day). This immediately suggests a number of comparisons which could be made, but 1) voting serves a different purpose on `Stackoverflow` and there is less incentive for users to 'game' its voting system and 2) no data on the timing of users' votes is available for `Stackoverflow`. The only thing one can say with certainty about the vote limit on `Stackoverflow` is that it has only been reached by 6,114 users (because the 'Vox Populi' badge is awarded based on this criterion and on 11th March 2013 it had been awarded 6,114 times, by contrast the badge for users who have voted more than 300 times in total has been awarded 19,809 times). `Slashdot` also (heavily) limits the number of 'ratings' a user can make, when a user is granted this capacity they can make a total of five ratings.

The introduction of a limit on how frequently users can vote on `reddit` would be one way to curb the extent to which individuals can maximise their influence by voting rapidly over a prolonged duration. A limit of 40 votes per day would be much too low for `reddit` (40 votes per hour seems like a more appropriate limit) - but if administrators had an appetite to introduce this kind of limit they could, one suspects, quite easily experiment with the precise nature of the limit until they found something appropriate.

This would 'level the playing field' between active users who vote often enough to reach this limit. If highly active users knew that the number of votes they could cast was limited this would remove a lot of the incentive to vote rapidly - and users who take time to make reasoned decisions before voting would be less likely to be disadvantaged in terms of their level of influence. One of the problems with introducing a site-wide limit would be that the content in different subreddits requires a differing investment of time to make a reasonable assessment of its merits - one might imagine that 10 seconds is ample time to assess some of the posts in /r/pics, but it is not enough time to assess a scientific article. A limit on the number of votes a user could cast would need to take this into account - and subreddit specific limits might be a better approach.

The limitless nature of votes means that they have little perceived value, one user can spend their time scrolling down a page making 30 votes in a minute and if another user makes one vote in that minute the first user has exerted 30 times as much influence as the second - from the voting system's perspective there is no other difference. A limit on the number of votes one could cast would at least limit the duration for which one could apply the 'rapid-fire' approach to voting (without 'cheating' by using multiple accounts). Such a limit would likely also give users a sense that their votes had value (because they would become a finite resource) and this could influence their choices about how they 'spend' them for the better. This limit would not need to be stringent to be effective, and could be introduced at a level where it would only affect the most active 1% or 2% of users. Votes are the fuel of Social News and more votes should mean better collective decisions - but it is difficult to see any advantage for the website as a whole in allowing a single user to cast more than 23,000 votes in a month.

Finally, some caveats - the data which this recommendation is based on was collected in March 2009, the situation may have changed since then and quick-voting may have become less prevalent. `Reddit`'s administrators may even have quietly imposed such a limit without announcing it. Without access to voting data there is no way to tell.

E) Can we conceive of Distributed Moderation systems as mechanisms for the management of Common Pool Resources (Ostrom, 1990)? If so, how do they perform in this regard? Are they managing their CPRs to avoid a tragedy of the commons or are they susceptible to 'free riders'?

Section 2.8 introduced the concept of Common Pool Resources (CPRs) and in the case of `reddit` its CPR is attention and the voting system functions to regulate access to this resource. In this framing users who only submit posts are 'free riders', attempting to gain attention for their own posts without contributing much of their own attention to the common pool, and without participating in the administration of the CPR by voting.

Section 6.1.5 showed that this pattern of behaviour is associated with poorly scoring post submissions, suggesting that those who wish to 'free ride' on `reddit` do not prosper - and by inference that `reddit`'s DM system may be quite robust to attempts at free riding. Given that users cannot see each others' voting records, the concept of 'community involvement' was invoked in explaining these findings. This implies that users may be assessing the posts of others in terms of whether they were submitted by an 'outsider' or someone who is familiar with `reddit` (and hence probably a member of the community) - with those perceived to be outsiders usually denied access to the CPR.

Instances where `reddit` users have engaged in 'community policing' and 'detective work' to expose

perceived abuses of `reddit` (see section 2.6) also fit with this CPR framing. Indeed there are whole sub-communities dedicated to 'policing `reddit`'s CPR' (e.g. /r/reportthespammers), and there have been several instances observed where posts claiming to uncover and document a particularly egregious abuse have appeared on the Front page - indicating that users collectively may place importance on this kind of policing.

In cases where free riders or users who are otherwise abusing `reddit` are identified, users' collective down-voting of the offender's content may serve as a 'graduated system of sanctions' (Ostrom, 1990). If many users remember the offender's alias and continue to down-vote their submissions this serves to restrict the offender in future attempts to obtain attention on `reddit` - with the restriction lasting as long as voters remember their identity. There is also an indication that down-votes may be used in a purely punitive manner (i.e. reports of uses browsing through other user's submission history and down-voting their contributions with the intention of reducing their karma score).

Chapter 7 showed that reddit's users can collectively deploy their CPR to achieve a range of different objectives (by promoting a particular perspective on `Wikileaks` and serving as a 'counter-measure' for actions against it, then in raising awareness of SOPA and organising collective action against it). In the SOPA case `reddit` went beyond consuming its own CPR and actively sought to expand this resource by deliberately attempting to garner attention from outside its own domain - both the GoDaddy boycott and the blackout were successful in doing this and generated a considerable level of coverage from both older and newer media elites.

An interesting corollary of this CPR framing is that within it 'lurkers' are not perceived as 'free riders' but as contributors of their own attention who don't seek to gain attention for their own submissions in return. The presence of what is assumed to be a large number of 'lurkers' or 'readers' is one of the reasons why `reddit`'s approach can be construed as a non-zero-sum game. For the community of users there is a 'pay-off' for collectively maintaining a useful information resource, as this will attract and maintain new users, bolstering the `reddit`'s CPR in the process and increasing its value.

F) Are the users of Social News websites homogenous or heterogenous with respect to their outlook and opinions? Does there tend to be a dominant majority opinion on any given topic? Where minority opinions are voiced can they be heard?

> Social News voting systems seems to offer the majority a mechanism through which they could silence dissenting voices - are they used in this manner? Sunstein (2002) espouses the importance of exposure to a variety of opinions on a topic, can Social News provide this variety of opinion in highly visible locations? Are these sites echo

chambers, or could they reach the lofty status of 'general interest intermediary'?

One of the potential problems with applying Distributed Moderation to political discourse in particular is that it might create an environment which allows for a 'tyranny of the majority' - where a there is a clear majority of users who hold a certain opinion they may suppress dissenting points of view. It is important that a variety of perspectives can be heard - as groups where all discussants are of like mind tend to move collectively towards more extreme formulations of this position (Myers and Lamm, 1976), the 'echo chamber' effect. A more common term for this on `reddit` would be the 'hive mind' - incorporating both the exclusion of opinions which dissent from the established norms and the reinforcement of popular opinions/trends.

With this in mind a case study was conducted on `reddit`'s coverage of the release of US diplomatic communications by `Wikileaks` in late 2010. A sample of posts observed on the New page of /r/politics (and thus unfiltered by the voting system) were taken as representative of users' opinions - and the tone of these posts was then compared to that of posts which appeared on the Front page. In this case the majority opinion was pro-`Wikileaks` and Assange, critical of the way the US government and corporations reacted to the story and critical of how the conventional news media covered the story. Posts which espoused contrary opinions were less likely to appear on the Front page *but were not entirely excluded* from appearing there. For example, 19% of New page text posts had a negative tone towards the `Wikileaks` organisation whereas of the Front page posts 10% had this attitude.

In this case users' votes led to an over-representation of the most popular opinions on the Front page, but did not lead to the total exclusion of minority opinions. There does not appear to be any single characteristic of `reddit`'s voting system which leads to this outcome - but this is a trend which has been observed in a number of other contexts on `reddit`. This is perhaps related to the 'random' components of `reddit`'s voting system and the presence of users who are *willing to vote based on the 'objective' merits of a post even when it makes an argument they do not agree with.* When a 'high-quality' post argues against a widely accepted '`reddit` norm' it can reach prominent locations if it gets off to a good start and is effectively competing with posts which are of inferior 'quality' (even if these support the majority opinion).

The comments for posts also tend to display a range of different attitudes towards the post's subject matter. In particular, the manner in which comments are threaded means that it is effectively discrete 'chains' of comments which are being ranked and sorted through the voting system. Where a comment thread involves back-and-forth discussion of an issue between two (or more) users, as they often do, this will inevitably result in a roughly equal share of attention for both sides of the argument.

In the context of political discourse it is important that there is room for the expression of con-

flicting points of view. However, `reddit` has many subreddits which are not dedicated to political discourse or current events. `Reddit`, being a website used by 55 million 'unique users' in the last month and having a robust 'free speech' policy, sees a considerable number of post and comment submissions which are racist, sexist, homophobic or offensive in any number of other ways, some of which have already been considered.

In the course of this research I have been 'casually' observing `reddit`'s default Front page with varying degrees of regularity to get a feel for what it is showing and how this is changing. I do not personally subscribe to many of the default subreddits and do not make any attempt to exhaustively look at the posts on the default Front page (which, by early 2013 on the poorest default subreddits oscillate between cringe-worthy and nauseating) or read their comments pages. Instead I read a few subreddits that I find interesting, collect data and analyse this to address specific questions or areas.

Bearing this in mind, I do not regularly see content which is blatantly racist, sexist or homophobic (if the umbrella is broadened to include anything offensive then the frequency would be much higher) - when I do see this kind of content it is usually in the form of insensitive attempts at humour rather than outright bigorty. However, the presence of some posts which make light of issues like racism or sexism, along with the occasional post about `reddit`'s core demographic (young white US males) being discriminated against or victimised, can make for an uneasy mixture. It must be made clear that I much more frequently see content on the default Front page which admonishes racism, sexism and homophobia or links to some resource which does.

Largely through posts *about* these problems (some of which have been displayed prominently on `reddit`, and some of which cite examples), I am aware that they exist and are in fact quite common in some parts of the site. The most recent post I have seen which raises these issues was submitted by William Shatner to /r/IdeasForTheAdmins (redd.it/18536m) where he asked for a feature which would allow him to easily ignore the swathe of private messages he was receiving once it became widely known that he was *the* William Shatner of Star Trek fame. On this post one user (radii314) commented: "Bill, you mentioned some of the unsavory aspects of `reddit` in an early post somewhere ... I hope you know there is a Dada aspect to this place with the absurd, weird, offensive and strange just chiming in from left field from time-to-time ... there is much of interest to mine here but some bad neighborhoods too" - to which Shatner responded with the following:

> "The unsavory aspects still exist - I am apalled by some of the immature, horrifically racist, sexist, homophobic, ethnic... etc.. posts that are just ignored here. Why are these accounts still active? While `reddit` has done well in getting interest from the mainstream I just wonder if by allowing these children to run rampant and post whatever they feel will cause the most collateral damage if `reddit` is biting off it's own

nose in taking that step to become a mainstream community. That being said, I'm still
new here. That's been my observation in my short time here and I could be wrong.
MBB"

This exchange spawned an /r/bestof post which appeared on the default Front page (redd.it/1876zm)
and was commented on 2,700 times - the highest-scoring comments discuss this aspect of `reddit`
and voice a number of perspectives. The two highest-scoring comments as of March 2013 were

"So because Bill says it it's r/bestof material? People have been saying the same thing
on a daily basis for a long time now. Usually they just get downvoted."

"I've been a software engineer helping to build the internet longer than many of you
have even been alive, and I've been a member of hundreds of online communities. To
those of you saying that this kind of speech is unavoidable because "this is the internet"-
I call bullshit. This is not "the internet", it is a community ON the internet. Peer-
reviewed research has consistently proven that online communities can successfully set
the tone and standards for their members. If `reddit` doesn't even want to attempt
to do that, fine, but don't hide behind saying "this is the internet" and 'free speech'
this and 'free speech' that. And quit pretending that `reddit` doesn't on the one hand
pretend to be progressive while at the same time throwing around words like 'faggot'
every five seconds. Show some guts, quit hiding behind what amounts to flag waving,
and be honest about your own behavior."

The third highest-scoring comment has since been deleted and the fourth is similar in tone to the
second above. The three most dominant perspectives within high-scoring comments were that 1)
the post was only being discussed because it was said by William Shatner, 2) that `reddit`'s users
should do more to supress offensive content by recognising and down-voting it, 3) that `reddit`'s
users have a right to make jokes about prejudice and use slurs in an 'ironic' way, and that if
someone is offended by this they shouldn't use the subreddits where it happens. This third point
is reminiscent of the idea that one doesn't have a 'right to be offended' on the Internet - a position
strongly associated with some of the image boards on `4chan`. With respect to this attitude (and
some other attributes), parts of `reddit` could be described as a 'watered down' version of `4chan`
- although this assertion is based on second-hand information - largely posts where `reddit` users
talk about their perceptions of `4chan` and `reddit` (e.g. redd.it/18jm8j).

In his comment William Shatner seems to be suggesting that users who post offensive comments
should be banned - it is likely this aspect in particular which some users have taken exception to.
Banning users for saying offensive things is not the '`reddit` way' - `reddit`'s philosphy is that this
kind of speech should be down-voted to obscurity and disregarded (provided that the 'community
of users' do not find it to have any merit).

As one delves into `reddit`'s 'early stage' pages or the comments pages for 'young' posts - one will encounter content which is less 'moderated' (because people are yet to vote on it) and more raw. It is in these areas where one is most likely to encounter offensive content - one hopes that as a greater number of users see each item and vote on it that objectionable items will be 'pushed out of view'. This is how Distributed Moderation is intended to work on `reddit` - if hateful speech manages to appear in highly visible locations then this can be seen as a failure of either the voting system or voting users.

While these 'failures' do occur and objectionable content can appear in highly visible locations - this sometimes sparks a discussion about why the content is objectionable and its appearance in a prominent location is harmful. Over the years I have seen a number of posts or comments appearing in high-visibility locations which explain, in simple terms and from the victim's perspective, why the offending content is harmful - what it feels like to be confronted by prejudice.

When `reddit`'s users vote to display an insensitive joke prominently it may be because they do not understand what it is like to be the victim of prejudice - this seems like a reasonable conjecture given that the core demographic is young white males. I would therefore tend to regard the exchanges where a user explains what it is like to be the victim of prejudice as a valuable part of the discourse on `reddit`. If `reddit` were to ban offensive content entirely, ignoring for now the difficulty in doing this fairly, `reddit` would become, in the eyes of some users, just another place where they can't say certain things. By allowing these things to be said they can (usually) be seen to be rejected (through their negative scores) by the community - and sometimes one will receive a lesson in why it is wrong to say these things in general, regardless of whether one is 'allowed to' or not.

Clashes among users over issues like this can also be taken as further evidence that on `reddit` people with different attitudes and opinions come into contact with one another and are exposed to alternative perspectives - a 'public good' in and of itself (Sunstein, 2002). However, it remains to be seen how an increase in users' management of their subreddit subscriptions will affect this - whether users are currently in the process of 'walling off' areas of `reddit` where people with shared ideas about the acceptability of different types of speech congregate.

Section 8.10 considered the prospect of an 'eternal September' for individual subreddits - and there are some subreddits where this phenomon appears to have interacted with group polarization (Myers and Lamm, 1976), driving the subreddit to become an echo chamber. To take the /r/politics subreddit as an example, this always had a liberal slant but has begun hurtling towards 'echo chamber' status recently. Suppose that to begin with `reddit`'s users were slightly more liberal than average, and that over time group polarization began to act on these individuals and their opinions started shifting further in this direction. Once this shift is pronounced enough to cause some users to un-subscribe the 'eternal September' begins - and as waves of more centrist users

leave, the group polarization effect is amplified each time. The result is a subreddit which becomes hopelessly one-sided and where most users no longer look for political news or discussion - a mouthpiece for its own 'hive mind'. In section 8.4 it was shown that /r/atheism and /r/politics were the two default subreddits which new users were most likely to unsubscribe from - and the term 'echo chamber' could be applied to both. [2]

Ghosts in the machine

After reading `reddit` for some time, particularly the popular comments on prominent posts to default subreddits, one begins to develop a sense of what the most popular comments for a post will be like before one has read them. That is not to say that all of these popular comments will be similar, but that each post will have a number of different 'types' of comment - and based on the nature of the post and the subreddit it is submitted to different types of comment are more or less likely to appear prominently. For instance, a popular post on /r/politics will often have some high-scoring comments which attempt to debunk the source, decry it as sensationalist and sometimes criticise the subreddit's users for up-voting the post to prominence. There are usually also high-scoring comments which take the post's content as support for their position on the issue (when these are high-scoring it is often a perspective which many `reddit` users hold). The types of comment which one expects to see at the top ranks varies between subreddits, for many subreddits comments which attempt to be humerous are the norm, with puns, similies and references to well-known `reddit` lore being common.

When one begins to read the comment threads on more serious subreddits these are often back-and-forth discussions between users holding contrasting perspectives - they often appear as a discussion between two individuals and sometimes they are, but upon closer inspection of the user-names of discussants one will sometimes find that the main thread of the discussion actually involves several individuals and merely looks like a discussion between two people.

High-activity areas of the website often seem like they are being produced by a limited selection of '`reddit` personalities' (e.g. the debunker, the explainer, the pun artist, the eye-witness, the libertarian, the anti-capitalist, the `reddit`-critiquer, the small subreddit promoter, etc.) rather than by a very large and disparate set of unique users. This illusion is partially a result of the lack of salience for users' identities, users have unique names but these are displayed somewhat innocuously. Users do not have profile pictures and, aside from the 'OP' (Original Poster) designating [square brackets], or 'flares' on subreddits which offer them, there is no visual marker associated with a user's contributions. This is intentional, content is supposed to be judged on its own merits and its submitter is supposed to be irrelevant - the lack of a recognised problem with 'power users'

---

[2]The /r/politics subreddit was in fact eventually removed from the default set in late 2013, having by this point become so one-sided and ill-tempered that its presence on the default Front page was causing embarrassment.

can be taken as evidence that this approach is working.

My hypothesis for why there is a similar profile to the 'types' of comment which consistently score well on highly active subreddits relates to the fact that up-votes are much more common than down-votes on `reddit`. Suppose that most users who read the comments page for a post will tend to up-vote a certain 'type' of comment, will rarely down-vote comments, and will only read a limited number of the comments which are already high-scoring before they move on. Users will up-vote a few comments of the 'type' which they like and mostly pass over those comments which are not of this type. This would allow for a number of different 'types' of comment to achieve high scores and prominent locations - the 'types' which resonate with the greatest proportions of users who are reading those pages. This could be thought of as users 'rallying around' a comment which matches their perspective and which (by virtue of its submission time and current votes) has a chance to be prominently positioned. As the readership of a given large subreddit is relatively stable from day to day with regard to the 'types' of comment which they look for, the same 'types' of comment will regularly appear in prominent locations in roughly the same proportion.

If, on the other hand, `reddit`'s users were in the habit of voting on everything they read (i.e. down-voting all of the comments which they presently pass over) - this would be much more likely to produce high-scoring comments which were homogenous with respect to their 'type'. In this scenario, it would be much less likely that content espousing 'minority opinions' could appear in high-visibility locations.

The result of the way `reddit`'s users actually vote (i.e. mostly up-votes) leads to the expression of a few of the more common perspectives among users who are active in that area (the 'types' of comment which these users look for) in high-visibility locations - even when these perspectives are contradictory. This makes sense when we consider the number and variety of `reddit`'s users - but to those who perceive a monolithic 'hive mind' which governs `reddit` this must look like a bad case of cognitive dissonance.

G) What becomes of users who disagree with the collective judgments being made through the DM system?

In the early stages of this research `reddit`'s Front page was home to a distinctive mix of content types - and the focusing of user attention on the small sub-set of items which appeared there gave `reddit` a strong sense of identity. Through the voting system `reddit`'s users collectively determine which items appear on the Front page - and it would not be surprising if the mass of voting users were somewhat consistent in this regard. This is one of the tenents of the 'hive mind' concept as applied to `reddit` by some `reddit` users - its compliment being that minority opinions will be routinely down-voted. What becomes of the users who hold these minority opinions? Do they

'drop out' or do they persist in a space which is hostile to their views (and given the nature of DM, which may suppress their views)? This has important implications for the broader potential of DM systems - if users who hold contrary opinions have no incentive to continue their participation then this does not bode well for the potential of DM because a system which tends towards a homogenous user-base is of little value in many domains.

Without access to the participation histories and subreddit subscription choices of users there is little which can be done to empirically address this question. There are however some relevant examples which suggest that some people remain as `reddit` users and un-subscribe from the specific subreddits which are hostile to their views.

`Reddit` has throughout this research been a website where most users identify themselves as atheists or agnostic (see section 6.1) - and since October 2011 /r/atheism has been one of the default subreddits. As a result `reddit`'s Front page frequently displays posts which express irreverence or disrespect towards religion (most commonly Christianity and Islam) - and many of these would likely be perceived as offensive by adherents to these religions. One might therefore expect that Christians and Muslims would be discouraged from using `reddit` at all. However, over the course of this research the /r/Christianity subreddit has grown to over 57,000 subscribers and /r/Islam now has over 10,000 subscribers. This suggests that at least some users are willing to persist on `reddit` despite the prevalence of anti-religious views in /r/atheism and more generally.

Furthermore, there are a number of active and well-known subreddits which are explicitly 'anti-reddit' (e.g. /r/ShitRedditSays, /r/CircleJerk, /r/CircleBroke). This rise of subreddits like this is somewhat surprising because one would imagine that the obvious course of action for someone who doesn't like `reddit` is to stop using it - not to join a community of like-minded individuals *on reddit*. Of course there may be many more individuals who take one look at `reddit`'s Front page and never return, and these people will not be represented in any available data. Nevertheless, for some people it appears that `reddit` is simultaneously both disagreeable and compelling.

H) Do DM systems have strengths and weaknesses in terms of the types of content they process?

Reddit's voting system seems inherently biased towards content which can be quickly and easily consumed, particularly on the highly active default subreddits. The following factors all contribute to this:

- Content which is quickly and easily consumed will naturally convert views into votes at a faster rate than content which takes longer to digest.

- Users can vote as often and as quickly as they please.

- Voting is the dominant force on `reddit` - it determines what is seen and unseen, it gives karma points and it takes them away. If users want to 'have an effect' on `reddit` the obvious approach is to vote quickly and often.

- Voting is entirely anonymous - there is no chance that a user will be 'called out' for up-voting a post which was shown to be nonsense.

- Active areas of the website see a very high rate of incoming submissions - content which requires time and effort to appraise will generally have to compete with content that is much more easily digested or knows how to 'push `reddit`'s buttons' in its title.

- Users who want `reddit` to show them content that is deep, insightful and comprehensive in nature (i.e. longer and taking more time to appreciate) have for some time now been deserting the default subreddits in favour of smaller subreddits.

`Stackoverflow`, on the other hand, serves as an illustration that DM can be calibrated differently to foster a different type of content.

On `Stackoverflow` voting is not the dominant method of interaction, it plays a supporting role and there is not the same feedback loop between voting and attention. `Stackoverflow` also limits the number of votes a user can cast in a day, and relatively few users have ever reached this low daily limit. This suggests that there is less incentive for `Stackoverflow`'s users to vote a lot - probably because on `Stackoverflow` it would be pointless to try and 'game' the voting system.

`Stackoverflow`'s purpose is also more tightly and rigidly defined, if content is not programming-related it doesn't belong there. We can assume that this has two effects - firstly the website's users all share a knowledge of programming (although there are a wide range of ability levels). Secondly, it is easier to make an objective decision about which content should be up-voted because there are more concrete standards for what makes a good post (i.e. questions should usually include snippets of code to demonstrate the problem, answers are often either correct of incorrect, more comprehensive answers are usually better).

In understanding the difference in which types of content fare well on `Stackoverflow` and `reddit` we must also consider what the websites' users are using it for. On `Stackoverflow` many users are programming professionals and `Stackoverflow` is a useful resource in their profession - users are often there as part of their work or to help others with their work-related problems. `Stackoverflow`'s answer-submitting users may even be able to leverage their participation to help them find employment.

On `reddit`, judging from the default Front page, many users are there to be entertained. The provision of 'news' now appears to be a secondary function. Furthermore, on default subreddits which aim to provide news it appears that many users don't thoroughly appraise posts before they

vote on them - with the result that sensationalist or inaccurate posts can often appear in high-visibility locations. On these subreddits the comments pages (once they have received a sufficient level of voting activity) are often more informative than the posts themselves. In terms of news coverage, `reddit`'s strength now appears to be in the democratically mediated commentary whiich accompanies each post, rather than in the content of the posts themselves.

I) Do DM systems fare well with controversial content that divides opinion (e.g. politics); or are they better suited to subjects which are more objective or emotionally neutral (e.g. the best way to solve a programming problem)?

`Reddit`'s DM system has problems with content that divides opinion, but these are not necessarily the problems which I anticipated at the outset of this research. The way in which redditors actually use the voting system (i.e. mostly up-votes) means that minority opinions are not entirely excluded from visible locations - it is possible to discuss controversial issues on `reddit` and to see more than one side of the argument. The problem is with how one interprets the scores that accompany the comments of which a discussion is constituted - are these simply comments which have been judged to contribute most to the discussion (if all users followed the Reddiquette guidelines this would be the case) or do they also tell us something about how `reddit`'s users feel about the topic of the discussion (if users voted to indicate agreement this would be the case)?

`Reddit`'s voting system is one-dimensional, some users may vote according to the guidelines while others may vote according to whether they agree with the content - or they may combine both attributes when they vote. The scores of content on `reddit` undoubtedly confound attention-worthyness with agreement, and this often causes problems when interpreting what the ranked list of comments on a post actually means. `Reddit` has no features which allow for collaboration or which attempt to build consensus, it is by nature a competitive environment. When users submit comments or posts they remain the 'property' of the submitting user. Even when these submissions are similar in nature they are, in so far as the voting system is concerned, competitors - one will be ranked more highly than the other dependant on the votes both receive.

`Stackoverflow` sidesteps the issue of controversial content somewhat because the type of content it deals with is less likely to divide opinions - users are more likely to agree that the answer to a programming question is good or bad than they are when the question concerns the US government's deficit, whether this is too high and how it should be addressed. If we compare the /r/politics subreddit to `Stackoverflow` it is easy to conclude that DM is more suited to programming questions on `Stackoverflow` - but we must recognise that the politics subreddit has an inherently more difficult task and that *it has not been designed with this task in mind.*

J) Do DM systems influence the opinions and perceptions of their users? Can these systems be said to promote groupthink or a 'hive mind'? Do they encourage users to think critically about the content they are voting on? Or do they encourage users to rely on the judgments of other users that have previously been registered through the voting system?

This question must first be broken down into its constituent components, of which two will be addressed. 1) Do DM systems encourage users to think critically about the content they are voting on? 2) Are users influenced by the judgments of other users that have previously been registered through the voting system?

With regard to the first component it is possible that by allowing users to register a vote on any item of content this might encourage them to think more about the merits of the content than if they were a passive observer. When users vote it may be with the mindset of a reviewer rather than a reader. It was in fact with this in mind that the decision was taken to compute and analyse the temporal gaps between users' votes - but when this analysis was conducted it provided evidence in support of a very different conclusion.

The key points on `reddit` are once again that users can vote as quickly and often as they please, and that the votes are effectively being used to sort 'winning' content from 'losing' content. These attributes encourage users to vote quickly in an effort to maximise their influence - there are undoubtedly users who see themselves as reviewers and vote in a considered manner because they want `reddit` to show good content, but these votes are likely drowned out by the quick votes of users who are more interested in making a lot of votes than making the 'right' vote on every piece of content.

With regard to the second component, users may well be influenced by the scores presented alongside content - and such a relationship has been demonstrated for the 'helpfulness' and 'usefulness' ratings of reviews on `IMDB.com` and `Amazon.com` respectively, with users' perception being influenced by these ratings and, perhaps critically, users tended to limit their 'review consumption' to those reviews which had scored well (Otterbacher et al., 2011).

On `reddit` such a relationship would likely be modulated by whether users 'trust' the subreddit which the content appears in. The experiment reported in section 7.3 suggested that the scores presented alongside /r/pics posts did influence users' perception of their quality, but that this influence was subtle. If this experiment were repeated with content drawn from /r/politics or /r/science it is quite possible the effect would disappear because many users do not 'trust' these subreddits to only allow high-quality posts to appear in prominent locations. A Republican party supporter may well experience an opposite effect with the scores of /r/politics posts because they perceive that subreddit as having a strong bias against their views - in this case a high score for a

post might lead to their perceiving it in a negative light.

There is some evidence that users do take the scores of content as indicative of the community's views on it - but that the community itself is often not seen as an impartial arbiter of truth. When one sees a post they do not like in a high-visibility location this may discredit the community on that subreddit in the mind of the reader, rather than influencing their perception of the post. The /r/ShitRedditSays subreddit is a whole community founded on this principle - users submit 'horrible' comments with positive scores and these are taken as ammunution for the lambasting of users on the subreddit where the comment was observed.

One could make an analogy to music charts here - for people who enjoy pop music the fact that a song is 'in the charts' might make them more likely to enjoy it, whereas for people who have decided they don't like 'chart music' this same fact would discourage them from enjoying it. Salganik and Watts (2009) conducted an excellent study of social influence effects in cultural markets which showed that independant groups made different collective judgments about music based purely on how the 'chart' evolved from the beginning of the study.

K) Are there any significant differences between the various forms of Distributed Moderation? Which of these differences are the result of differences in the software being utilised, and which are the result of different user-bases? Are certain features or characteristics of DM software more suited to certain contexts or content types?

There are undoubtedly major differences between Distributed Moderation platforms. At the outset this research aimed to contrast the two Social News websites `reddit` and `Digg` - but just one month after the data collection regime for `Digg` had been completed, `Digg` changed its software infrastructure fundamentally with the infamous 'upgrade' to version 4 (see section 3.1.4), and in the process made the scripts which collected data through its API obsolete. `Digg` failed to recover from the exodus of its users at this time, and therefore no effort was made to write new scripts to collect data from the 'new `Digg`'. `Reddit` and `Digg` have however been compared directly on one important characteristic, the dominance of 'power users', and in this regard `reddit` was found to be superior in that its Front page was not dominated by elite users to the same extent as `Digg`'s.

This left `reddit` and `stackoverflow` as the only two DM websites from which useful data was incoming, and in considering `reddit` and `Stackoverflow` these differences are so great that they make it difficult to conduct any meaningful comparisons. Each DM website has both its own unique software and its own unique user-base - this makes it difficult to seperate effects caused by the software from those caused by differences in user behaviour. If the present research is to consider the technical and social aspects of Distributed Moderation seperately it is therefore better to compare different parts of the same website.

Section 5.1.12 explored the /r/IAmA subreddit - a subreddit which caters to a kind of discourse which is distinct from other subreddits, and which relies on users behaving differently to achieve this. The analysis of /r/IAmA data showed that users did collectively behave in the ways which were required for the subreddit to function as intended, answers by the OP were generally up-voted above comments from other users. The /r/IAmA subreddit is unusual in that there are clearly specific requirements for how users should vote and it is relatively easy to compare collective user behaviour to these requirements. There are many other subreddits which cater to a unique style of discourse and where users operate the standard voting system to achieve different ends, but understanding how these subreddits differ from one another requires closer inspection of the qualities of the content they deal with.

Similarly, the `Stack Exchange` websites share a common software infrastructure but cater to different topics and have independant user-bases - one could also study these in an attempt to seperate the software from the social.

It has been possible to compare `Reddit` and `Digg` directly on one important characteristic, the dominance of 'power users' - and in this regard `reddit` was found to be superior in that its Front page was not dominated by elite users to anywhere approaching the same extent as `Digg`'s.

L) How do Social News websites relate to the 'hybrid media system'? (Chadwick, 2013)

The case studies in Chapter 7 are useful in situating `reddit` within the Hybrid Media System (Chadwick, 2013) (see section 7.2.9 for a summary). In its most essential form `reddit` is an attention 'amplifier' - its posts can link to any resource on the web, and when they score well and appear on the Front page they channel `reddit`'s users' attention towards that resource. The `Wikileaks` case study revealed considerable variety in the types of resource linked to by Front page posts, with both older and newer media being well represented (see figure 7.2).

In the SOPA case study `Reddit` was characterised by an unusual mixture of older and newer media logics - and peformed the dual roles of news broadcaster and collective action platform concurrently.

In its capacity as information broadcaster, `reddit` appears to have played an important part in raising awareness of SOPA - firstly among its own users through Front page posts and then beyond `reddit` through collective action endeavours which deliberately sought (and gained) the attention and coverage of older and newer media elites.

This case study also showed that `reddit` was capable of setting its own agenda, whereas in the previous `Wikileaks` study `reddit` had provided an alternative narrative on a story which was also dominating the conventional news media - this being a common approach among newer media actors, explicitly so in the case of *38 degrees* (Chadwick, 2013). The SOPA case study also

demonstrated that elite actors from older and newer media are paying attention to `reddit` - it may now have earned a place among the social media platforms which professional journalists routinely consult as part of their own work (as described by Chadwick (2013)). `Reddit`'s users are aware that their website has a wide reach and often display this awareness through posts which address or are directed towards external entities who it is assumed can be 'reached' through the Front page. The SOPA case study also documented Front page posts which were attempts to identify and forge alliances with sympathetic entities (corporations, non-profits and politicians) - and later posts which announced/celebrated when these entities took actions which were perceived as 'joining the battle' on `reddit`'s side.

The question of whether `reddit` now had the capacity to effectively 'break' a story was also considered - and as the number of users browsing its Front page increases so too does the chance that `reddit` can present a story to enough people (in the USA at least) that the conventional news media are unable to ignore it - even if they would not have chosen to report the story themselves.

The melding of 'information broadcaster' and 'collective action facilitator' roles appears to have been advantageous in the SOPA case study. Section 7.2.4 describes an attentional feedback loop which was important to the success of the GoDaddy boycott - once the boycott had started and external entities began to take notice their articles/posts about the boycott were fed back into `reddit` as new posts and some of these appeared on the Front page, thereby informing `reddit`'s users that their action was having an affect and probably re-affirming their commitment to it.

The collective action which `reddit` facilitated in opposition to SOPA is similar in many ways to that attributed to *38 degrees* by Chadwick (2013) - both draw from well established set of collective action repetoires (signing petitions, contacting representatives, boycotts) and adopt strategies like seeking to generate coverage in the conventional media and targeting decision-makers directly.

The striking dissimilarity between `reddit` during the SOPA study and *38 degrees* is in how this collective action comes about. *38 degrees* has much more in common with a conventional hierarchical organisation than `reddit` - with a clear distinction between 'leaders' who set the agenda and 'members' who merely choose which 'actions' they will participate in. The collective action which occurred through `reddit` was truly 'grass-roots' in nature and strongly shaped by the voting system, inflecting it with both democratic and anarchic qualities. If there were 'leaders' in these endeavours they were certainly not pre-appointed and will have emerged 'on the spot' by virtue of provding something (be it an idea/resource/service) which the endeavour needed in order to proceed. The accounts I have provided of how some of these endeavours came about also suggest that through the voting system `reddit` has effectively 'crowd-sourced' some of the tasks which are still the purview of leaders in an organisation like *38 degrees*. For example, the democratically mediated discussion which accompanied every popular proposal affords users an opportunity to consider the merits of the proposal collectively, with many being 'abandoned' at an early stage - *38*

degrees leaders perform a similar task by polling their 'members' and piloting actions on a sample of members (Chadwick, 2013).

MoveOn campaigns have been described as "not organising without organisations but organising with different organisations" (Karpf, 2012). I would argue that `reddit` represents a step towards 'organising without organisations', with much of the 'organisation' being implemented in software, leaving users to participate as pseudonymous peers.

M) Can the comments on a Social News (SN) website be thought of as 'deliberative discussions' (Habermas et al., 1989)? Or are they better suited to quickly determining the group's consensus and then broadcasting that message?

`Reddit`'s commenting system represents an unusual method of conducting discussions (although it has at this point been heavily emulated), and it is regrettable that this research has not had time to conduct a thorough investigation of this discourse with respect to its qualities - in particular how `reddit`'s discussions align with the ideals of deliberation (Habermas et al., 1989). This kind of study could be conducted with a similar approach to that of Wright (2007), and the data which would allow one to do so has already been collected. Alas, this is just one of many potentially fruitful uses of this data which the present research has not had time to pursue.

This research does however make frequent reference to comments, and the role of the commenting system has been explored in various contexts (from AMA interviews in section 5.1.12 to collective action in section 7.2). In section 5.1.8 I argued that post voting can be seen as relatively 'shallow' and error-prone (based on analysis of how the voting system works and how individuals use it) - but that a mature comments page has the capacity to display `reddit`'s more 'considered' appraisal of a post. There may be a significance to this configuration which has not yet been explored.

Keen (2007) is concerned that social media threatens conventional media production - and has particular reverence for the gatekeeping roles which are a mainstay of older media. Keen (2007) also seems to have little faith in the capacity for non-experts to assess the quality or reliability of information - and thus a cacophony of un-filtered information is of grave concern.

One of reddit's functions in the Hybrid Media System is that of gatekeeper - more specifically users regulate the content which appears in high-visibility locations through the voting system. This system is not perfect, and it is not unheard of for inaccurate or unreliable posts to appear prominently on subreddits which are intended to provide solid information - the classic example being posts to /r/science which link to a hyperbolic press release and re-iterate its boldest claims in their title. This legitimises Keen (2007)'s concern, but there is perhaps a positive side to this. While post voting allows unreliable posts to score well, the (mature) comments pages for posts which have appeared on the Front page tend to display 'high-quality' comment threads in their

top ranks. Where the post itself is of poor quality these top-scoring comments almost always state this and, crucially, explain *why* the post's content cannot be taken at face value.

This format, whereby posts cannot be assumed to be of high quality (but where they are not there will be comments which explain why) may actually be beneficial in helping readers to develop their own critical information-assessment faculties - a skill which is becoming of central importance for life in the 'information age'. While Keen (2007) describes having to assess the quality of information for onself as a terrible burden, unless one is willing to rely exclusively on the spoon-feeding of validated information from respected sources - it is unavoidable. Furthermore, critical appraisal of information is and always has been of vital importance to a whole range of pursuits, including scientific research.

Depending on the subreddit, its comments pages may be filled with reference-citing answers to scientific (/r/AskScience) or historical (/r/AskHistorians) questions, easily understood answers to questions on a broader range of topics (/r/ExplainLikeIm5) - or inane jokes about a new iteration of image macro meme (/r/AdviceAnimals). Subreddits like /r/AskScience and /r/AskHistorians (see section 8.6) are worthy of singling out as spaces which facilitate a co-mingling of users with an interest in questions on a particular topic and those with the means to provide (often expert) answers. In these spaces the answers go beyond practicality and function (qualities which are dominant on Stackoverflow) but translate their scientific content into language and narratives which are closely tied to the user's question and their perspective. The success of /r/AskScience and /r/AskHistorians in attracting and retaining large numbers of subscribers is strong evidence that these subreddits are successful in imparting their scientific/historical information in a way which is highly accessible to some of `reddit`'s core demographics.

`Reddit`'s users 'collectively' 'figure out' all sorts of things in the comments pages of posts - for many of `reddit`'s (predominantly young) users watching these reasoned discussions (ranging from why the latest 'cure for cancer' is a red herring to why a 'humerous' post is actually subtly re-inforcing racial stereotypes) unfold may be valuable experience. The 'shallow-ness' of post voting, couped with the more reasoned outcomes of comment voting (where 'expert' users are advantaged) - ensures that one of the themes of `reddit`'s Front page content (presuming readers consult the comment pages for posts) is critical appraisal of that very same content. It would be interesting to observe whether `reddit`'s users get better at appraising content over time - but while the user-base continues to grow so rapidly this is not possible to discern.

The investigation of comments on `reddit` highlights key ways in which the site diverges from conventional media and other forms of social media. In comparison to conventional media, `reddit` 'shows its working-out' on the comments pages for posts, whereas conventional media tend to leave such material in the 'news room' and unpublished. In comparison to other forms of Social Media - `reddit` has users with a wide range of expertise on any given topic and *these users actually*

communicate with one another. Users who are truly expert on a given topic may be a vanishingly small minority, but (through comment voting in particular) it appears that contributions from these users can be made highly visible and thus central to the discourse on `reddit`.

Consider one recent example of this phenomenon on a post to the /r/rpg (role-playing games) subreddit (redd.it/1prop9). In this post a user ('UnremarkablePassword') observes that *Hasbro* owns both 'Dungeons & Dragons' and 'Magic: The Gathering' (two highly popular games) and asks why they have not produced a hybrid product which combines both games. The top answer is from a user ('rsdancey') who "was the brand manager for Dungeons & Dragons and the VP of Tabletop RPGs at Wizards of the Coast from 1998 to 2000" - and as such can provide a detailed and robust answer. This answer is not only up-voted to the top of that post's comments page but also submitted to /r/bestof, through which it appeared on `reddit`'s default Front page. This prompts the user 'DonkeyTrain' to state "I love it when someone asks a question on reddit and the most qualified person on earth shows up out of no where to answer."

These two attributes of how `reddit` works in practice (dialogue between experts and non-experts, critical appraisal of post content) may even come to be regarded as some of `reddit`'s legacies in years to come. If we are to have a reliable 'crowd-sourced' news system in the future it will surely owe a debt to `reddit` for pioneering/popularising many aspects of its approach - and maybe also for 'training' its crowd to become more effective appraisers of information.

N) Can the users of DM systems use them to make higher-order decisions about what the endeavour's purpose is or how its users should behave to achieve their collective goals? Are the users of SN websites capable of behaving as a unitary social entity?

Over the course of this research there have been situations where '`reddit`' has appeared to behave as a unitary social entity. These largely occurred in 2009-2011 when there was still a sense of a singular `reddit` community which all users were part of, during these times it was not uncommon to see comments which indicated a sense of pride in `reddit`'s achievements. As `reddit`'s user-base has grown it has become more diverse and there is no longer a sense of a singular `reddit` community - in 2013 one is more likely to encounter subreddits whose users criticise the users of other subreddits, `reddit` has become a place of many communities and these communities do not always get along.

When there was still a common sense of what it meant to be a 'redditor' the community was capable of surprising feats - rallying behing causes and executing endeavours of collective action. `Reddit` was instrumental in bringing about the 'Rally to Restore Sanity and/or Fear', helped `Wikileaks` to maintain a visible web presence when this was under attack, was a hub for spreading awareness of the Stop Online Piracy Act and organising collective action which opposed the legislation. Section

7.4.1 reported analysis which suggested that `reddit`'s users could assess and ratify a suggestion through post and comment voting and then, collectively, behave in a way which executed the proposal (in this case that `reddit` should stop linking to Fox News). The SOPA case study showed that it was possible to 'betray `reddit`'s trust' (Al Franken) and to thoroughly 'piss `reddit` off' (Lamar Smith, GoDaddy), and that there were consequences to doing so (`reddit` was subsequently more likely to prominently criticise, slander or make fun off the offending party). The collective action facilitated by `reddit` during the SOPA study also shared considerable similarities with other more conventional collective action organisations like *38 degrees* despite lacking the hierarchical structure of these organisations.

These situations where `reddit` appeared to behave as a collective social entity share one commonality - a majority of `reddit`'s users were united in their stance on the issue at hand and `reddit`'s collecive stance on the issue was known. In 2013 `reddit` is a more diverse and fragmented place, if there is a large subreddit where users mostly hold a shared perspective there is probably another subreddit whose users see them as being hopelessly mis-lead by the 'hive mind'. Things which happen on or through `reddit` are now perhaps as likely to be mocked as celebrated.

On `reddit`'s default Front page serious content has been displaced by entertainment - with regard to serious topics it is increasingly difficult to determine what `reddit`'s users believe, and without this shared understanding it is probably not possible for `reddit` to throw its weight behind a plan for collective action any more.

While `reddit` is no longer a beacon of optimism for Distributed Moderation, neither is it a reason to dismiss DM as a method of decision-making and organisation for large groups. `Reddit` was not designed with these capabilities in mind, and as the original founders and employees have deserted their replacements do not seem interested in developing a system which allows for these capabilities. `Reddit` has more users now, and probably greater revenues now, than at any point in the past - if commercial success is `reddit`'s goal there is no reason to tinker with it or attempt to broaden the capabilities of its voting system. `Reddit`'s owners seem to be quite happy with the direction the website is heading in.

`Reddit`'s voting system appeared to allow for the expression of collective intelligence when the website had hundreds of thousands of users, despite the fact that it had been designed for a much simpler purpose (aggregating news). `Reddit` has seen a massive increase in user activity but its core mechanisms and structure have not been changed or adapted to accomodate these higher activity levels. That rudimentary collective intelligence and higher-order decision-making was in evidence at all on `reddit` is grounds for optimism about future developments of Distributed Moderation.

O) Are there any ways in which DM systems might be improved or adapted to work well in contexts where the current exemplars of the form struggle?

If one considers current exemplars of Distributed Moderation as early experiments with a novel approach to large-scale discourse, and one is interested in the broad social implications of this method of doing things, then this becomes the most important question of all. What else can we use these systems for, and how does one go about adapting DM for different purposes? The following section considers this question with respect to political discourse, drawing on some analysis of 'Your Freedom' (section 5.3) against the backdrop of what we have learned about DM through the study of reddit and Stackoverflow.

9.2 Adapting Distributed Moderation to Political discourse

The perspective which informs this section is one where the current examplars of 'Distributed Moderation' represent an approach to decision-making and information production which is in its infancy.

As a species we are beginning to experiment with novel approaches to collective decision-making and information production - these include Wikis, Voting systems, Prediction Markets, Tagging systems, Recommendation Engines and Crowd-funding initiatives. 'The Crowd' is proving to be a versatile (but by no means infallible) entity, performing a wide range of tasks which were previously the purview of organisations and experts. Given that the underlying 'Web 2.0' technology making all of this possible has been available for less then ten years it seems safe to assume that there are further ways in which 'the crowd' can be engaged that are yet to be discovered. With this in mind the appropriate attitude when we consider adapting DM to the political is one of optimism or ambition.

The first thing I would advocate is a move away from rating or ranking as the core objective of user activity - and instead consider this as an exercise in Peer Production and collective decision-making. The voting systems which reddit is built on are suited to the production of 'Social News', tending towards content which is easily digested and/or entertaining. Stackoverflow demonstrates that these same systems can be utilised in the production of an altogether different kind of resource, one which does not emphasise the snappy one-liner but the comprehensive and if necessary lengthy. The difference in software is that on Stackoverflow the voting system plays a supporting role to other mechanics from the 'Social Media Toolbox' (e.g. Wikis, tagging). On reddit it stands alone as the website's dominant force, it determines what is seen and not seen - sorting the winners from the losers[3]. This competitive aspect of submitting/voting on reddit would likely prove detrimental

---

[3]Of course the different user-bases and aims of these websites also contribute to their differing styles of content.

to the kind of endeavour I am describing.

Another problem with `reddit`'s uni-dimensional voting system is that it confounds attention with agreement. According to `reddit`'s guidelines votes should reflect whether an item is attention-worthy or not, according to (and bemoaned by) `reddit`'s users, they are often used to express agreement or disagreement with the item being voted on. Any application of DM to politics needs a method through which users can make known their agreement or disagreement - and this should be seperate from the mechanism used to determine which items are attention-worthy. Furthermore, such an endeavour requires a means through which users can *make concrete decisions*, not merely express their feelings in a non-binding way.

I will now outline the kind of system I would personally like to see created as a next step towards applying DM in the domain of politics. I do not mean to suggest that such a system would be perfect or even function well, but I feel that in this area there is more to be learned from the study of an ambitious endeavour that fails than a conservative one that succeeds - provided that the system saw enough user activity to assess its functionality.

The goal I would set for such an endeavour is one of collaboration in the production of a political document, as a starting point this document would be a manifesto of sorts. If we want to test whether an unorganised group of 'peers' can perform the work of a political party or government this is a good place to start. Much of the work of professional politicians involves the negotiation and production of documents (manifestos, white papers, policy documents, legislation).

The basis of such a system would be a Wiki, DM as used by `reddit` would play only a supporting role. Wikis have proven to be a highly effective way for a large number of users to collaborate on the production of a document - but one of their weaknesses is in their inability to deal with issues that divide the community of users. Wikis are a useful tool for the assembly of knowledge which is 'objective' and emotionally neutral, but even the best-known Wiki (`Wikipedia`) has had problems with content that either divides opinion or which pertains directly to the vested interests of some group. There have been several instances where 'turf wars' have broken out on `Wikipedia` as opposing groups tried to gain control of pages important to their cause so that they could use them to shape public opinion in their favour. For instance, it appears that pro-Israel and pro-Palestine groups have in the past organised in attempts to ensure that `Wikipedia` was telling their side of the story on pages relating to controversial issues (e.g. the Palestinian right of return) (Snydar, 2008). `Wikipedia` has also in the past banned whole IP address ranges used by the US House of Congress (WebProNews, 2006) and the church of scientology (Singel, 2009) because editors from these locations were manipulating `Wikipedia` to serve their purposes.

Wikis have no built-in mechanisms to deal with this kind of abuse, when it comes to light the task of addressing it falls, in `Wikipedia`'s case, to the arbitration council - a small group of well-

respected users who have power to take actions not available to other users, including the banning of accounts/IPs and the locking of pages. If a wiki was to be used in producing a political document it would require a means of resolving conflicts democratically. It is inevitable that some users would edit the document so that it reflected their views, even if they knew that they were in a minority. It would be very damaging for such an endeavour if users with special powers were required to address these issues, because this would mean that these users effectively had the final say in any dispute. For users to keep faith with a system like this the voting system(s) would need to be the ultimate arbiter of disputes, the only effective means of decision-making on the site.

The core of such an endeavour would be a 'democratic wiki' - whereby proposed edits to the document would need to be ratified by a voting system before being 'committed' to the actual text. This would take the form of a yes or no vote on the edit which would have to remain open for a specified interval (e.g. 1 day) and would have to receive a certain number of votes to be valid. This would cause the endeavour to operate at quite a slow pace, but in some senses this is a desirable characteristic - it would be beneficial in that users would more easily be able to follow the development of the text and discuss controversial points at length. It would be better for such an endeavour to slowly produce a basic document that most users could support, rather than quickly producing a document which the community of users do not feel that they collectively 'own'.

If such a system saw a considerable level of user activity the number of edits to be voted on would quickly increase and become unmanageable. At this point it would be important to have the facility for an individual or self-determined group to package a selection of similar edits together to be voted on.

This problem of managing a large number of 'edit votes' is also where a Social News type voting system (a 'Reddit for Edits') could be deployed. Within this component each 'edit vote' would be equivalent to a post on `reddit` - it would have an up/down score and this score coupled with a time penalty would be used to create 'hot' pages that showed where the community's attention was currently being directed. For an edit to receive enough votes to be passed it would likely have to be 'up-voted' to a prominent location on the 'Reddit for edits' and thus exposed to a large number of users. These users would vote to accept or reject the edit.

Each edit's page would also have a comments section, similar to `reddit`'s in that it would rank comments by their voting response but with the addition of a second dimension for 'agreement' alongside the primary dimension of 'attention-worthyness'. `Reddit`'s approach to comments has proven to be effective in distilling the opinions of a large number of users in order to see at a glance how the community feels about a subject. This would be useful information for users who compose edits, particularly if a second dimension were added whereby users denoted their agreement with comments. In recording input on two dimensions users would be able to see 1) what the commonly perceived issues around an edit were and 2) how the community of users felt about these issues.

If the requirement for an edit to be approved was a simple majority (i.e. greater than 50%) this could lead to a text which 'flip-flops' between two positions on an issue which divides the community. It would perhaps be better to set more stringent criteria for the ratification of an edit (e.g. approved by at least two thirds of users who vote) in order to produce a more stable text.

In creating such a system one would also need to think carefully about issues surrounding user identity and accountability. Both `reddit` and `Stackoverflow` (and many other on-line platforms) cater to pseudonymous interaction - users have user-names but these are not necessarily tied to their identity off the website, on `Stackoverflow` using one's real name is optional, on `reddit` it is discouraged. Users are therefore free to create multiple accounts. One of the problems which `reddit` has faced as a result has been with users who control many 'sock-puppet' accounts and attempt to use these to manipulate the voting system. `Reddit` has proprietary techniques for identifying and nullifying these attemtpts to cheat. A DM system being used to produce a political document would also need to have an effective means of limiting users to one active account each.

On `reddit` and `Stackoverflow` user voting behaviour is entirely anonymous. In the analysis of back-end voting data from 2009 we also saw that users who submitted posts and voted were largely seperate populations. The result is that users who held a lot of influence over the collective decisions being made on `reddit` were effectively invisible to other users.

Anonymity is realted to poorer quality contributions in online communities of practice (Kilner and Hoadley, 2005) - but it is not clear whether this relationship will hold true for voting systems, where no public trace of the user's contribution exists. The low 'quality' of anynomous (or pseudonymous) contributions is closely tied to a tendancy in some users to deliberately seek a negative reaction from others (trolling) (Suler, 2004) - and voting represents a very indirect way of provoking this reaction (although comments on `reddit` in which users ask why they are being down-voted suggest that reactions are sometimes elicited). In relation to voting, the complete anonymity of the voting system and limitless nature of votes likely increase the chance that users will vote in a 'careless' manner, in addition to raising the possibility that they could vote in a manner which is deliberately anti-social or manipulative without fear of being detected.

The anonymity/pseudonymity of participants is a characteristic which one might well change in applying DM to the production of a political document. When elected representatives vote on legislation the manner in which they vote is generally public knowledge. There might well be advantages to making users' voting records public - this would add some accountability for voting behaviour and may encourage users to vote in a more considered manner. If users knew their voting record would be publicly displayed this might be something they could take pride in. It would also allow users to know when there were individuals who were voting frequently and/or indiscriminately.

The presence of user 'profiles' on `reddit` and `Stackoverflow` seems important in motivating some users to participate - both websites keep a record of user's submissions, a tally of their reputation points, and award badges for certain behaviours. By carefully designing the reward systems, and by incorporating voting within a user's history and the reward system, it may be possible to encourage the kinds of user attitude and behaviour which such an endeavour would require to be successful.

Finally, there is much to be learned about 'the crowd' by observing its behaviour in a scenario where votes result in concrete decisions (in this case that edits are committed to the text). This aspect would give users a greater level of collective responsibility for the resultant text than is the case for existing DM voting systems (i.e. if one is 'ashamed' by a the appearance of a post on `reddit`'s Front page it is easy to dismiss this as being the fault of other users, and in any case it will have disappeared from the Front page by the next day). If we were to afford 'the crowd' a much greater level of responsibiltiy for the resource they are producing (individually, by making their votes public knowledge; collectively, by having committed edits to the text remaining as part of the text unless there is a successful vote to remove or change them), would 'the crowd' collectively rise to this challenge and behave in more 'responsible' ways? Would users learn, as a collective, from their experience of producing this text - and get progressively better at performing this kind of task?

I for one would be very keen to observe both the process and outcome of this kind of endeavour.

9.3   Conclusions

This research has attempted to give a description of how some Distributed Moderation systems work - how they aggregate the contributions of their users to produce a resource, how and why their users participate. It has also considered the social implications of Distributed Moderation as a means of production and 'organisation' - in the here and now and also in the potential for its future development.

The research has also had a strong methodological component. It has probed the utility of publicly accessible data in understanding how these websites work and what they do. This data is rich and ecologically valid, but it also often requires extensive interrogation and assessment before yielding any useful findings. The data-collection scripts for `reddit` which were developed as part of this research will be provided online at the researcher's blog (www.researchmills.com) and will be linked to whenever work related to this thesis is discussed. These scripts may be useful to others who wish to collect data from reddit, but given the researcher's lack of a programming background at the outset of this research, the scripts themeselves are rudimentary (with the possible exception of comment records for posts) and could be easily surpassed.

Of greater interest is the 'data collection and transformation regime' within which these scripts are nested. The data-sets which have been analysed in this thesis were not retrieved through the API, they were contructed, with the data returned from API calls serving as the raw materials of their construction. To illustrate this point - many of the analyses presented herein concern reddit posts, and while it is easy to retrieve data on a post through the reddit API, the data returned describes only the moment in which the query is made (score, comments, etc.). The data-sets created for Chapter 5 involved not observing specific items of content but *locations* (pages). Where a data-set for posts is created it is based on tens or hundreds of sightings of the post in various locations and points in time - through these sightings a detailed history of where a post appeared and how it was performing is created. It is through this history that questions of how many votes/comments a post would receive by virtue of appearing in a particular place can be addressed. It is also possible to extract other meaningful variables like a post's score the first and last time it was observed on the Front page, the number of posts which appeared on a subreddit's Main page but ended with a negative score, etc. This can be taken further still by adding periodic queries of the post's comments page to the data collection regime, allowing for study of the evolution and movement of comment threads over time (and in relation to the post's own performance over time). Another illustration of the importance of data-set *construction* to this research can be seen in Chapter 6, where all analyses of back-end data presented rely on data-sets which were all constructed from a simple list of 3.5 million votes with their timestamp, valence, user and post ID.

This is, to my knowledge, an unusually 'involved' form of data assembly. One gets the impression that in many studies where an API is used to collect data for social science research online, the researchers have a firm idea at the outset of which questions they wish to address and what forms of data they can collect to address these questions. In the case of this research, only the source of data/subject of research was firmly established at the outset of data collection. The researcher's ignorance of APIs in general and reddit's in particular meant that ideas about what form the data would take or which questions could be addressed with it were necessarily loose. Extensive periods of time spent working with data collection scripts and the data itself allowed for a much deeper understanding of both reddit and the available data to develop. This understanding spurred the evolution of the 'data collection and transformation regime' and this greater understanding of the available data facilitated realisations about which questions could be addressed as part of the research.

This 'backwards' approach to the research also led to the collection of data-sets which are unusually 'broad' - and these data-sets themselves are likely to be of great value to any researcher with an interest in reddit, or more specifically its history between 2010-2013. As part of this research around 200Gb of data on `reddit` has been collected, including but not limited to: observations of the Front page and pages like New, Rising and main for a large selection of subreddits at 30-minute

intervals, and records of comments (their text, timing, score, rank, author) for all posts to specified subreddits over short intervals. The research presented in this thesis barely scratches the surface in terms of the level and variety of analyses which these data-sets would sustain - there are whole databases representing months of data which have not been 'augmented' in any way and have only been considered in the shallowest ways possible (i.e. longitudinal analysis in Chapter 8). There are also records representing a detailed (in some cases 'complete') history of subreddits (some which fizzled out, some which went on to attract a lot of subscribers, some which attempted to use DM in novel ways), which, due to time and space constraints are not even mentioned in the thesis. One promising approach which has not been pursued relates to the study of memetics (Marsden, 1998) - there are data-sets which could be utilised to investigate the evolution of popular language on `reddit`, and data-sets concerning subreddits which are closely linked to the emergence and proliferation of 'Image Macro Memes'.

All of these data-sets are not only conveniently accessible (they have already been extracted through the API and stored in a local `MySQL database`) - they are functionally irreplaceable. As reddit continues to grow in usage and public awareness, so too does the potential motivation to study the history of its development. However, reddit provides extremely poor access to its own historical records (as evidence one might wish to try searching for a post which they saw on reddit last week). For example, it would be difficult to determine whether `reddit` covered a particular story one month ago unless one witnessed this occur - it is impossible to go back and piece together what this coverage would have looked like (i.e. which posts appeared on the Front page, when did they appear there, what kinds of comments did these posts receive initially and did this change over time).

All of the case studies presented in Chapter 7 relied upon live data being collected on a longitudinal basis (instigated before the researcher had an inclination that some of the posts appearing on the Front page would come to constitute distinct cases worthy of study), and these case studies could not have been executed in this manner without such data. Further support for this approach to data collection can be gleaned from research published by others as the present research has been underway which emphasises the importance of 'high temporal resolution', data which allows one to piece together a minute-by-minute account of how a series of events unfolded (Chadwick, 2013; Elmer, 2013). Data collection continues, example snippets of data will be posted on the researcher's blog, and efforts will be made to make other researchers aware of the availability of this data.

In comparing the present research to that of, for example, Chadwick (2013) (which provides detailed accounts of how some stories unfolded through the Hybrid Media System), it diverges in ways which could be construed as both strengths and weaknesses. The weakness of the present research is that it has focused almost entirely on one site (`reddit`) and this limits the degree to which it can speak authoratively to `reddit`'s place in the 'Hybrid Media System'. Where Chadwick provides detailed

accounts of how information flowed between realms like the blogosphere, Twitter and conventional news, here I have largely relied upon `reddit` to inform me of occasions when it has 'made an impact' in the conventional media or blogosphere. In the SOPA case study this can actually be construed as a strength - the detail of how that story developed has been pieced together almost exclusively from information extracted from records of `reddit`'s Front page. If the case study provides a useful narrative on the SOPA story this implies that `reddit` itself produced a useful narrative on its Front page.

The strength of focusing on a single component of the Hybrid Media System is that it has allowed for a much 'deeper' understanding of how this entity operates. Returning to Chadwick (2013)'s treatment of the Hybrid Media System - Social Media entities' roles in these stories are largely recounted in terms of the items which ultimately had an impact on the story, the 'winners' in the competition for attention so central to Social Media. In this research we have looked beyond `reddit`'s 'winners' at the processes which produce these (Chapter 5) and at the broad range of 'entrants' from which the 'winners' were selected (`Wikileaks` case study). Each entity in the Hybrid Media System has its own internal methods of determining salience (e.g. voting on `reddit`, followers/re-tweeting on `Twitter`) - if we are to understand this system more fully we need to understand the internal logics of each Social Media platform and also how these may be interacting with each other 'beneath the radar' of the eventual winners - which are by definition highly visible and which have tended to draw the attention of researchers seeking to understand the social impact of these platforms.

As `reddit` has grown in scale and complexity the task of understanding what's happening there and how its working has become increasingly unmanageable for an inidividual or small group. One productive approach may be to enlist the help of `reddit`'s users in this task. For instance, one way of using publicly-accessible data is to scrutinise and classify it along certain criteria (as was conducted in the '`Wikileaks`' case study) - this is a labour-intensive task and one which could conceivably be 'crowd-sourced' if a suitable mechanism could be devised and implented and if willing volunteers were available.

As has been discussed, there are already active communities of users with an interest in understanding how `reddit` works and a willingness to do some work towards building this understanding (e.g. /r/TheoryOfReddit). The /r/TheoryOfReddit subreddit is itself a useful resource simply because it brings together the observations of a large number of individuals with an interest in `reddit` - this group can collectively observe things which an individual could easily miss and through voting the most interesting observations are raised to prominence for the subreddit's subscribers.

Several years ago this kind of interest was characteristic of `reddit`'s global community - as evidenced by the appearance of user-produced surveys on the Front page and their accompanying discussions. However as `reddit`'s global community fractured and disappeared this interest ap-

pears to have been lost - now one rarely sees posts which treat `reddit` itself as their subject on the default Front page. It was this interest and enthusiasm which I hoped to tap with the content-rating experiment - although by the time this was implemented and launched `reddit`'s users were arguably in the midst of losing their collective fascination with how the website worked.

In this research we have also encountered many of the limitations of publicly accessible data - questions about which it can tell us nothing or which could be much more thoroughly addressed with full access to procedural data. Even *with* full access to procedural data there are limits to what one can do in a purely observational capacity.

If we are to further our understanding of Distributed Moderation and its broader potential it may prove highly beneficial for social scientists to become much more directly involved in the design, implementation and running of these systems. Such involvement would grant much greater access to the data which one needs to understand how the system works, and even the possibility of controlling and manipulating aspects of the system in an experimental manner to observe the results - feeding enhancements in one's understanding of the system directly back into its design to improve its utility. If there are going to be attempts at adapting Distributed Moderation to other domains, as there have already been in the political domain, then could these attempts not also benefit from the presence of social scientists who work in relevant areas from their earliest design stages through to their execution?

If the academic community does not become more directly involved in the production of these systems then our understanding of them will likely develop as follows: 1) wait to see what the next generation of upcoming web developers produce and which instances people use, 2) request access to their data for research purposes, 3) if the request is denied (where the endeavour is already 'commercial' in nature it probably will be), hope that they provide a way to extract some useful data on which an understanding of how the system works can be based, or hope that the system's administrators can describe this accurately and honestly.

There is a further problem with waiting for Distributed Moderation to 'naturally' develop in a domain like politics. The active and successful forms of Distributed Moderation which the present research has considered share a common origin - they were produced by people who we might retrospectively describe as entrepreneurs. The endeavours have, from almost their outset, required a way to generate revenue and meet their costs. From their creators' perspective they are a product which, once sufficiently developed, can be sold at a high price - provided that they have at least *the potential* to generate a profit. In `reddit`'s case it was sold to the publisher Conde Nast at a relatively early stage, and has since been spun out into its own seperate company by Advance Publications, Conde Nast's parent company. One could argue that over the years the outlook of `reddit`'s administrators (a group which no longer includes any of the original founders or employees) has become more profit-oriented (e.g. increased presence of advertisements, the sale

of a premium 'Reddit Gold' service and subsequent efforts to make the 'gifting' of `reddit` Gold to another user a normative way of expressing recognition for their contributions).

The adaptation of Distributed Moderation to politics is not compatible with a profit motive, and this would likely make it an unappealing target for an upcoming entrepreneurial web developer. In any case we should hope that this is true, because a successful adaptation of Distributed Moderation to political policy *which aimed to make a profit from users' participation* would potentially be a very dangerous entity. In order for such an endeavour to be successful and pro-social it will require the financial backing of an organisation which does not aim to profit from its investment. In this sense it is somewhat encouraging that established political parties and governments have shown a willingness to invest in this kind of facility (`Your Freedom`, `AmericaSpeakingOut`), although the manner in which they have spent this money (i.e. out-sourcing the task to a company which provided a sub-par system with no ambition beyond replicating some recognisable DM features) is disappointing.

There are however small groups of altruistic individuals who are attempting to create DM systems geared to politics - for the greater good rather than in the expectation that they will personally profit. One such group is working on a project called Votorola, describing itself as follows: "Votorola is social software in support of public autonomy. We develop the communication tools to enable a radically free, self-determined society, one in which plans of action are coordinated by discussion aimed at mutual understanding and consensus." (Votorola, 2013)

Votorola's software has been in development for three years, and can actually trace its origins back to a post appearing on `reddit`'s Front page (redd.it/arzh7) in January 2010. This post was titled "America, we need a third party that can galvanize our generation. One that doesn't reek of pansy. I propose a U.S. Pirate Party." and submitted by user levmyshkin. The post was very popular, appearing on the Front page for an extended period of time and receiving almost 2,500 comments. As observed with many subsequent collective action endeavours popularised through `reddit`'s Front page - this one was quickly outfitted with its own subreddit, wiki, forum and web site(s) adopting the name 'American Pirate Party'. The de facto leaders of the endeavour (i.e. people who contributed a lot in the early stages) organised meetings on IRC in which they intended to discuss and decide on the nascent party's policies and executives. These meetings were attended by hundreds of people and IRC was not an appropriate venue for a discussion of this nature on this scale. It was this environment which precipitated the Votorola project, a small group of participants set themselves the task of producing a piece of software which would allow all of the people who wanted to be involved in the American Pirate Party to collectively determine what its policies would be.

In the absence of such a system, and in keeping with subsequently observed collective action endeavours launched through `reddit`'s Front page, interest in the American Pirate Party waned

and eventually died off. Members of the Votorola project had already realised that the software they intended to produce would be of utility to a broad range of organisations, and as such the Votorola project had already become independent of the American Pirate Party. Work on the development of Votorola's software tools persists to this day. A list of similar projects is available at: `metagovernment.org/wiki/active_projects`.

The idea that large-scale user voting behaviour can be the driving force behind an endeavour remains an intriguing possibility. `Reddit` was at one point a demonstration of what was possible with a simple vote up/down mechanic and some ranking algorithms - in observing its development we have also witnessed the limitations of this approach. However, there are limitless possibilities to how one could go about harnessing large-scale voting behaviour to make concrete decisions. We are now beginning to experiment with these possibilities.

For now, it is easy to dismiss the idea of instituting a more direct or participatory form of governance through on-line voting systems as being an impractical pipe-dream. However, if the short history of Internet communications is any indication, dismissing this possibility out of hand may turn out to be a short-sighted perspective. The Internet is littered with examples of concepts that could have easily been dismissed as unworkable *until the point where one instance established that a new form of communication or production did in fact 'work', and through its ongoing popularity was in fact 'here to stay'.*

Bibliography

Adamic, L. and Glance, N. (2005). The political blogosphere and the 2004 US election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, page 43. ACM.

Adamic, L. and Huberman, B. (2000). The nature of markets in the World Wide Web. *Quarterly Journal of Electronic Commerce*, 1(1):5–12.

Adamic, L., Zhang, J., Bakshy, E., and Ackerman, M. (2008). Knowledge sharing and yahoo answers: everyone knows something. In *Proceeding of the 17th international conference on World Wide Web*, pages 665–674. ACM.

Adamic, L. A. and Adar, E. (2003). Friends and neighbors on the web. *Social Networks*, 25(3):211–230.

Albert, R., Jeong, H., and Barabási, A.-L. (1999). Internet: Diameter of the world-wide web. *Nature*, 401(6749):130–131.

Alexa (2010). Alexa.com: The web information company. `http://www.alexa.com`.

Anderson, C. (2008). *The long tail: Why the future of business is selling less of more*. Hyperion Books, New York.

Anstead, N. O'Loughlin, B. (2011). The emerging viewertariat and bbc question time: Television debate and real time commenting online. *Press/Politics*, 16(4):440–462.

Antin, J. and Churchill, E. F. (2011). Badges in social media: A social psychological perspective. In *CHI 2011 Gamification Workshop Proceedings (Vancouver, BC, Canada, 2011)*.

Arrington, M. (2006). Troubles in diggville. `http://techcrunch.com/2006/09/06/troubles-in-diggville/`.

Atwood, J. (2009). A theory of moderation. `http://blog.stackoverflow.com/2009/05/a-theory-of-moderation/`.

Awan, A. (2007). Virtual jihadist media: Function, legitimacy and radicalizing efcacy. *Journal of Cultural Studies*, 10(3):389–408.

Baker, K. J. M. (2012). How to shut down reddits creepshots once and for all: Name names. `http://jezebel.com/5949379/naming-names-is-this-the-solution-to-combat-reddits-creepshots`.

Barabasi, A. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286:509–512.

Barabási, A., Albert, R., and Jeong, H. (2000). Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A: Statistical Mechanics and its Applications*, 281(1-4):69–77.

BBC (2011). Godaddy faces boycott over sopa anti-piracy law support. `http://www.bbc.co.uk/news/technology-16320149`.

Beaumont, C. (2010). Twitter users send 50 million tweets per day. `http://www.telegraph.co.uk/technology/twitter/7297541/Twitter-users-send-50-million-tweets-per-day.html`.

Bell, D. (1977). Teletext and technology: new networks of knowledge and information in postindustrial society. *Encounter*, 48(6):9–29.

Benen, S. (2010). Checking in with america speaking out. `http://www.washingtonmonthly.com/archives/individual/2010_06/024361.php`.

Benkler, Y. (2002). Coase's Penguin, or, Linux and "The Nature of the Firm". *The Yale Law Journal*, 112(3):369–446.

Benkler, Y. (2006). *The Wealth of Networks: How Social Production Transforms Markets and Freedom.* Yale University Press, New Haven.

Bimber, B., Flanagin, A., and Stohl, C. (2012). *Collective action in organizations: Interaction and engagement in an era of technological change.* Cambridge University Press.

Bohman, J. (2004). Expanding dialogue: The internet, the public sphere and prospects for transnational democracy. *The Sociological Review*, 52:131–155.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and Isdn Systems*, 30(1-7):107–117.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. (2000). Graph structure in the web. *Computer networks*, 33(1-6):309–320.

Calhoun, C. (1998). Community without propinquity revisited: Communications technology and the transformation of the urban public sphere. *Sociological Inquiry*, 68(3):373–397.

Castells, M. et al. (2000). *The rise of the network society.* Blackwell Oxford.

Chadwick, A. (2011). The political information cycle in a hybrid news system: The british prime minister and the bullygate affair. *The International Journal of Press/Politics*, 16(1):3–29.

Chadwick, A. (2013). *The Hybrid Media System: Politics and Power*. Oxford University Press.

Change.org (2010). Change.org. `http://www.change.org`.

Chen, A. (2012). Unmasking reddits violentacrez, the biggest troll on the web. `http://gawker.com/5950981/`.

Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-law distributions in empirical data. *Siam Review*, 51(4):661–703.

Coleman, S. and Gotze, J. (2001). *Bowling together: Online public engagement in policy deliberation*. Hansard Society London.

Collins, R. (1979). *The credential society: An historical sociology of education and stratification*. Academic Press New York.

Corrado, A. and Firestone, C. M. (1996). *Elections in cyberspace: Toward a new era in American politics*. Aspen Institute of Human Studies.

Dahlgren, P. (2005). The internet, public spheres, and political communication: Dispersion and deliberation. *Political Communication*, 22(2):147–162.

Daisley, B. (2010). Understanding the social media explosion. In *Proceedings of the Association of Survey Computing Conference: Embracing the Social Media Revolution*.

Davis, J., Hirschl, T., and Stack, M. (1997). *Cutting edge: technology, information capitalism and social revolution*. Verso Books, New York.

DiMaggio, P., Hargittai, E., Neuman, W. R., and Robinson, J. P. (2001). Social implications of the internet. *Annual Review of Sociology*, 27:307–336.

Dimiero, B. (2012a). Report: News networks ignore controversial sopa legislation. `http://mediamatters.org/blog/201201050008`.

Dimiero, B. (2012b). Study: Sopa coverage no match for kim kardashian and tim tebow. `http://mediamatters.org/blog/201201130015`.

Dover, D. (2008). How reddit ranking algorithms work. `http://www.seomoz.org/blog/reddit-stumbleupon-delicious-and-hacker-news-algorithms-exposed`.

Editorial, L. T. (2011a). Piracy vs. an open internet. `http://www.latimes.com/news/opinion/opinionla/la-ed-newpiracy-20111125,0,519794.story`.

Editorial, N. Y. T. (2011b). Going after the pirates. `http://www.nytimes.com/2011/11/27/opinion/sunday/going-after-the-pirates.html`.

Eisenstein, E. (1979). *The printing press as an agent of change*, volume 2. Cambridge University Press.

Elmer, G. (2013). Live research: Twittering an election debate. *New Media & Society*, 15(1):18–30.

Eriksen, E. (2005). An emerging european public sphere. *European Journal of Social Theory*, 8(3):341–363.

Facebook and Youtube (2010). Digital debate. `http://apps.facebook.com/digitaldebate/`.

Farrell, D. and Petersen, J. (2010). The Growth of Internet Research Methods and the Reluctant Sociologist*. *Sociological Inquiry*, 80(1):114–125.

Fenn, J., Raskino, M., and Gammage, B. (2009). Gartner's Hype Cycle Special Report for 2009. *Retrieved August*, 13:2009.

Flake, G., Lawrence, S., Giles, C., and Coetzee, F. (2002). Self-organization of the web and identification of communities. *Communities*, 35(3):66–71.

Friedman, M. (2010a). $240 to place story on digg front page: One marketing firms pitch. `http://newsfeed.time.com/2010/08/30/digg-users-lash-out-at-new-format-join-forces-with-reddit/`.

Friedman, M. (2010b). Reddit campaign for colbert rally breaks donation record. `http://newsfeed.time.com/2010/09/14/reddit-campaign-for-colbert-rally-breaks-charity-records`.

Gaffney, D. (2010). # iranelection: quantifying online activism. `http://devingaffney.com/-iranelection-quantifying-online-activism`.

Ganley, D. and Lampe, C. (2009). The ties that bind: Social network principles in online communities. *Decision Support Systems*, 47(3):266–274.

Ghosh, R. (2005). Understanding free software developers: Findings from the FLOSS study. *Perspectives on free and open source software*, pages 23–45.

Gilbert, E. (2013). Widespread underprovision on reddit. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 803–808. ACM.

Giles, J. (2005). Wisdom of the crowd. *Nature*, 438:281.

Gillmor, D. (2004). *We the media*. Financial Times Prentice Hall, New Jersey.

Glasgow University Media Group (1976). *Bad news*, volume 1. Routledge, Oxford.

Goldhaber, M. (1997). The attention economy and the net. *First Monday*, 2(4-7).

Goldstein, M. L., Morris, S. A., and Yen, G. G. (2004). Problems with fitting to the power-law distribution. *European Physical Journal B*, 41(2):255–258.

Gonzalez-Bailon, S. (2009). Opening the black box of link formation: Social factors underlying the structure of the web. *Social Networks*, 31(4):271–280.

Gordon, H. S. (1954). The economic theory of a common-property resource: the fishery. *The Journal of Political Economy*, 62(2):124–142.

Graham, T. Wright, S. (2013). Discursive equality and everyday talk online: The impact of superparticipants. *Journal of Computer-Mediated Communication*, pages 1–18.

Green, P. J. and Yandell, B. S. (1985). *Semi-parametric generalized linear models.* Springer, New York.

Greenberg, A. (2007). Digg's drm revolt. `http://www.forbes.com/2007/05/02/digital-rights-management-tech-cx_ag_0502digg.html`.

Grossman, L. (2009). Iran protests: Twitter, the medium of the movement. *Time Magazine*, 17.

Guardian (2010a). Election debate: Youtube and facebook put your questions. `http://www.guardian.co.uk/media/pda/2010/apr/29/youtube-facebook-election`.

Guardian (2010b). Nick clegg enlists public to help battle red tape and invasions of privacy. `http://www.guardian.co.uk/politics/2010/jul/01/nick-clegg-red-tape-privacy`.

Guardian (2012a). Exposing reddit's purveyors of 'creep shots'. `http://www.guardian.co.uk/commentisfree/2012/oct/11/reddit-creep-shots`.

Guardian (2012b). Reddit blocks gawker in row over 'creepshot' photos. `http://www.guardian.co.uk/media/2012/oct/12/reddit-blocks-gawker-creepshot-photos`.

Guardian, T. (2012c). Leveson inquiry: Rupert murdoch appears, day one. `http://www.guardian.co.uk/media/2012/apr/25/rupert-murdoch-live-blog`.

Gulli, A. and Signorini, A. (2005). The indexable web is more than 11.5 billion pages. In *Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 902–903. ACM.

Gutenberg, B. and Richter, C. (1944). Frequency of earthquakes in California. *Bulletin of the Seismological Society of America*, 34(4):185.

Habermas, J., Burger, T., and Lawrence, F. (1989). *The structural transformation of the public sphere.* MIT Press Cambridge, MA.

Halavais, A. (2009). DO DUGG DIGGERS DIGG DILIGENTLY? *Information, Communication & Society*, 12(3):444–459.

Halavais, A. (2012). A genealogy of badges. *Information, Communication & Society*, 15(3):354–373.

Halliday, J. (2010). Digg investigates claims of conservative 'censorship'. `http://www.guardian.co.uk/technology/2010/aug/06/digg-investigates-claims-conservative-censorship`.

Hargittai, E. (1999). Weaving the Western Web: Explaining differences in Internet connectivity among OECD countries. *Telecommunications Policy*, 23(10-11):701–718.

Harper, F., Moy, D., and Konstan, J. (2009). Facts or friends?: distinguishing informational and conversational questions in social q&a sites. In *Proceedings of the 27th international conference on Human factors in computing systems*, pages 759–768. ACM.

Herman, E. and Chomsky, N. (1988). *Manufacturing consent*. Pantheon Books New York.

Hertel, G., Niedner, S., and Herrmann, S. (2003). Motivation of software developers in Open Source projects: an Internet-based survey of contributors to the Linux kernel. *Research policy*, 32(7):1159–1177.

Hess, C. and Ostrom, E. (2003). Ideas, artifacts, and facilities: information as a common-pool resource. *Law and contemporary problems*, 66(1/2):111–145.

Hogg, T. and Lerman, K. (2007). Stochastic models of user-contributory web sites. In *AAAI Conference on Weblogs and Social Media*.

Hogg, T. and Szabo, G. (2009). Diversity of user activity and content quality in online communities. In *Proc. of the Third International Conference on Weblogs and Social Media (ICWSM2009)*, pages 58–65.

Huberman, A. H. and Adamic, L. A. (1999). Growth dynamics of the world wide web. *Nature*, 401:131–132.

Huberman, B., Loch, C., and Oncular, A. (2004). Status as a valued resource. *Social Psychology Quarterly*, 67(1):103.

Huberman, B., Romero, D., and Wu, F. (2009). Crowdsourcing, attention and productivity. *Journal of Information Science*, 35(6):758.

Jane, M. (2011). *Reality Is Broken, Why Games Make Us Better and How They Can Change the World*. Penguin Press, New York.

jenakalif (2011). Who in the world is reddit? `http://blog.reddit.com/2011/07/who-in-world-is-reddit.html`.

Jenkins, H. (2006). *Convergence culture: Where old and new media collide*. NYU Press, New York.

Jordison, S. (2012). From civilization to big brother: how a game recreated orwell's 1984. `http://www.guardian.co.uk/books/booksblog/2012/jun/13/civilization-ii-big-brother-orwell-1984`.

Karpf, D. (2012). *The MoveOn effect: The unexpected transformation of American political advocacy.* Oxford University Press.

Keen, A. (2007). *The Cult of the Amateur: How Todays Internet is Killing Our Culture and Assaunting our Economy.* Nicholas Brealey Publishing, London.

ketralnis (2011). Nerd talk: The tale of the life of a link on reddit, told in graph porn. `http://blog.reddit.com/2011/07/nerd-talk-tale-of-life-of-link-on.html`.

Kilner, P. G. and Hoadley, C. M. (2005). Anonymity options and professional participation in an online community of practice. In *Proceedings of th 2005 conference on Computer support for collaborative learning: learning 2005: the next 10 years!*, pages 272–280. International Society of the Learning Sciences.

Kim, S. and Oh, S. (2009). Users' relevance criteria for evaluating answers in a social q&a site. *Journal of the American Society for Information Science and Technology*, 60(4):716–727.

Kittur, A. and Kraut, R. (2008). Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 37–46. ACM.

Kreiss, D., Finn, M., and Turner, F. (2011). The limits of peer production: Some reminders from Max Weber for the network society. *New Media & Society*, 13(2):243.

Kumar, R., Novak, J., Raghavan, P., and Tomkins, A. (2005). On the bursty evolution of blogspace. *World Wide Web*, 8(2):159–178.

Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. (1999). Trawling the Web for emerging cyber-communities. *Computer Networks*, 31(11-16):1481–1493.

Kunegis, J., Lommatzsch, A., and Bauckhage, C. (2009). The Slashdot Zoo: Mining a social network with negative edges. In *Proceedings of the 18th international conference on World wide web*, pages 741–750. ACM.

Kvalseth, T. O. (1989). Note on cohen's kappa. *Psychological reports*, 65(1):223–226.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM.

Lampe, C. and Garrett, R. (2007). It's all news to me: The effect of instruments on ratings provision. In *System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on*, pages 180b–180b. IEEE.

Lampe, C. and Johnston, E. (2005). Follow the (slash) dot: effects of feedback on new members

in an online community. In *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, pages 11–20. ACM.

Lampe, C. and Resnick, P. (2004). Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 543–550. ACM.

Lampe, C., Wash, R., Velasquez, A., and Ozkaya, E. (2010). Motivations to participate in online communities. In *Proceedings of the 28th international conference on Human factors in computing systems*, pages 1927–1936. ACM.

Landis, J. and Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Lardinois, F. (2010a). Digg redesign tanks: Traffic down 26 percent. `http://www.readwriteweb.com/archives/digg_redesign_tanks_traffic_down_26.php`.

Lardinois, F. (2010b). Digg user rebellion continues: Reddit now rules the front page. `http://www.readwriteweb.com/archives/digg_user_rebellion_reddit_on_front_page.php`.

Lawrence, S. and Giles, C. (2000). Accessibility of information on the web. *intelligence*, 11(1):32–39.

Leadbeater, C. and Powell, D. (2008). *We-think*. Profile, London.

Lerman, K. (2006). Social networks and social information filtering on digg. *Arxiv preprint cs/0612046*.

Lerman, K. (2007a). Dynamics of collaborative document rating systems. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 46–55. ACM.

Lerman, K. (2007b). Social information processing in news aggregation. *IEEE Internet Computing*, pages 16–28.

Lerman, K. and Galstyan, A. (2008). Analysis of social voting patterns on digg. In *Proceedings of the first workshop on Online social networks*, pages 7–12. ACM.

Lerman, K. and 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*, pages 621–630. ACM.

Linden, A. and Fenn, J. (2003). Understanding Gartner's hype cycles. *Paper on the Gartner Groups web site, at http://www. gartner. com/pages/story. php. id*, 8795:8.

Long, P. T., Van Dong Anh, N., Vi, N. T. T., Quoc, L., and Thang, H. Q. (2011). A meaningful model for computing users' importance scores in Q&A systems. In *Proceedings of the Second Symposium on Information and Communication Technology*, pages 120–126. ACM.

Lu, E. and Hamilton, R. (1991). Avalanches and the distribution of solar flares. *Astrophysical Journal*, 380:L89–L92.

Mamykina, L., Manoim, B., Mittal, M., Hripcsak, G., and Hartmann, B. (2011). Design lessons from the fastest q&a site in the west. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, pages 2857–2866. ACM.

Margolis, M. and Resnick, D. (2000). *Politics as usual.* Sage Publications, California.

Marres, M. and Weltevrede, E. (2013). Scraping the social. *Journal of Cultural Economy*, 6(3):312–334.

Marsden, P. (1998). Memetics and social contagion: Two sides of the same coin. *Journal of Memetics-Evolutionary Models of Information Transmission*, 2(2):171–185.

Masnick, M. (2009). Reddit, sears, grills that cook babies... and the streisand effect. `http://www.techdirt.com/articles/20090821/0213555950.shtml`.

Masnick, M. (2011a). NY Times LA Times Both Come Out Against SOPA & PIPA. `http://www.techdirt.com/articles/20111126/23150116902/ny-times-la-times-both-come-out-against-sopa-pipa.shtml`.

Masnick, M. (2011b). Sopa supporters learning (slowly) that pissing off reddit is a bad idea. `http://www.techdirt.com/articles/20111222/13292217173/sopa-supporters-learning-slowly-that-pissing-off-reddit-is-bad-idea.shtml`.

Mattson, K. (1998). *Creating a democratic public: The struggle for urban participatory democracy during the progressive era.* Pennsylvania State University Press.

Meeker, M. (2009). Economy and Internet Trends Report. `http://www.slideshare.net/misteroo/economy-tech-trends-in-2009-by-mary-meeker-morgan-stanley`.

Michels, R. (1915). *Political parties: a sociological study of the oligarchical tendencies of modern democracy.* Hearst's International Library Company, New York.

Milbank, D. (2010). Republicans new web site not exactly what they hoped it would be. `http://www.washingtonpost.com/wp-dyn/content/article/2010/05/25/AR2010052504396.html`.

Miller, L. (2009). E-petitions at westminster: the way forward for democracy? *Parliamentary affairs*, 62(1):162–177.

Mitzenmacher, M. (2004). A brief history of generative models for power law and lognormal distributions. *Internet mathematics*, 1(2):226–251.

Montopoli, B. (2010). Jon stewart rally attracts estimated 215,000. `http://www.cbsnews.com/8301-503544_162-20021284-503544.html`.

Munroe, R. (2009). reddit's new comment sorting system. `http://blog.reddit.com/2009/10/reddits-new-comment-sorting-system.html`.

Munson, S. A. and Resnick, P. (2010). Presenting diverse political opinions: how and how much. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1457–1466. ACM.

Murthy, D. (2008). Digital ethnography: An examination of the use of new technologies for social research. *Sociology-the Journal of the British Sociological Association*, 42(5):837–855.

Myers, D. G. and Lamm, H. (1976). The group polarization phenomenon. *Psychological Bulletin*, 83(4):602.

Naisbitt, J. (1991). *Megatrends 2000*. William Morrow, New York.

Newman, M. (2001). Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E*, 64(1):16131.

Newman, M. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary physics*, 46(5):323–351.

Newman, M., Forrest, S., and Balthrop, J. (2002). Email networks and the spread of computer viruses. *Physical Review E*, 66(3):35101.

Noam, E. (2003). The Internet: Still wide open and competitive? *OII Issue Brief*, 1.

Norris, P. (2001). *Digital divide: Civic engagement, information poverty, and the Internet worldwide*. Cambridge University Press.

Oktay, H., Taylor, B., and Jensen, D. (2010). Causal discovery in social media using quasi-experimental designs. In *Proceedings of the First Workshop on Social Media Analytics*, pages 1–9. ACM.

Oleoleolson (2010). Massive censorship of digg uncovered. `http://blogs.alternet.org/oleoleolson/2010/08/05/massive-censorship-of-digg-uncovered/`.

O'Reilly, T. (2005). What is web 2.0. *Design patterns and business models for the next generation of software*, 30:2005.

Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge university press.

Otterbacher, J., Hemphill, L., and Dekker, E. (2011). Helpful to you is useful to me: The use and interpretation of social voting. *Proceedings of the American Society for Information Science and Technology*, 48(1):1–10.

Pareto, V. (1896). *Cours d'economie politique*. Librairie Droz.

Poor, N. (2005). Mechanisms of an online public sphere: The website slashdot. *Journal of Computer-Mediated Communication*, 10(2).

Reddit-Blog (2010a). Everything went better than expected. `http://blog.reddit.com/2010/08/everything-went-better-than-expected.html`.

Reddit-Blog (2010b). "experts" misunderestimate our traffic, and we don't know why. `http://blog.reddit.com/2010/07/experts-misunderestimate-our-traffic.html`.

Reddit-Blog (2011a). reddit: billions served. `http://blog.reddit.com/2011/02/reddit-billions-served.html`.

Reddit-Blog (2011b). reddit levels up with three new programmers. `http://blog.reddit.com/2011/06/reddit-levels-up-with-three-new.html`.

Reddit-Blog (2011c). Saying goodbye to an old friend and revising the default subreddits. `http://blog.reddit.com/2011/10/saying-goodbye-to-old-friend-and.html`.

Reddit-Blog (2012). 2 billion & beyond. `http://blog.reddit.com/2012/01/2-billion-beyond.html`.

Reddit-Blog. (2012). Potus iama stats. `http://blog.reddit.com/2012/08/potus-iama-stats.html`.

Reese, B. (2013). *Infinite Progress: How the Internet and Technology Will End Ignorance, Disease, Poverty, Hunger, and War*. Greenleaf Book Group, Austin.

Rheingold, H. (1993). *The Virtual Community: Homesteading on the Electronic Frontier*. Addison Wesley Publishing Company, Boston.

Rheingold, H. (2003). *Smart Mobs: The Next Social Revolution*. Basic Books, New York.

Richardson, L. (2007). Beautiful soup documentation. `http://www.crummy.com/software/BeautifulSoup/bs4/doc/`.

Robinson, J., Kestnbaum, M., Neustadtl, A., and Alvarez, A. (2002). Information technology and functional time displacement. *IT & Society*, 1(2):21–36.

Rogers, E. (1966). Diffusion of innovations (1962). *Translated by A. Fujitake, BAIFUKAN*.

Rotham, W. (2010). 'cigar guy' is internet antidote to sad keanu. `http://www.msnbc.msn.com/id/39522903/ns/technology_and_science-tech_and_gadgets/`.

Ruggiero, T. E. (2000). Uses and gratifications theory in the 21st century. *Mass communication & society*, 3(1):3–37.

Salganik, M. and Watts, D. (2009). Web-Based Experiments for the Study of Collective Social Dynamics in Cultural Markets. *Topics in Cognitive Science*, 1(3):439–468.

Salihefendic, A. (2010). How reddit ranking algorithms work. `http://amix.dk/blog/post/19588`.

Schroer, J. and Hertel, G. (2009). Voluntary engagement in an open web-based encyclopedia: Wikipedians and why they do it. *Media Psychology*, 12(1):96–120.

Shirky, C. (2008). *Here comes everybody: The power of organizing without organizations*. Penguin Press, New York.

Silverstein, C., Marais, H., Henzinger, M., and Moricz, M. (1999). Analysis of a very large web search engine query log. In *ACM SIGIR Forum*, volume 33, pages 6–12. ACM.

Singel, R. (2009). Wikipedia bans church of scientology. `http://www.wired.com/business/2009/05/wikipedia-bans-church-of-scientology/`.

Snydar, T. (2008). Latest front in mideast wars: Wikipedia. `http://www.thejewishweek.com/news/international/latest_front_mideast_wars_wikipedia`.

Solon, O. (2011). Rome, sweet rome: movie that started as a reddit comment bought by warner bros. `http://www.wired.co.uk/news/archive/2011-10/17/reddit-movie`.

Stack Overflow (2011). Stack overflow creative commons data dump. `http://blog.stackoverflow.com/2009/06/stack-overflow-creative-commons-data-dump/`.

Stephen, A. T. and Toubia, O. (2009). Explaining the power-law degree distribution in a social commerce network. *Social Networks*, 31(4):262–270.

Stone, B. and Cohen, N. (2009). Social networks spread defiance online. *New York Times*.

Stvilia, B., Twidale, M., Smith, L., and Gasser, L. (2008). Information quality work organization in Wikipedia. *Journal of the American Society for Information Science and Technology*, 59(6):983–1001.

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & behavior*, 7(3):321–326.

Sunstein, C. (2002). *Republic. com*. Princeton University Press.

Surowiecki, J. (2005). *The Wisdom of Crowds*. Random House, New York.

Szabo, G. and Huberman, B. (2010). Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88.

Tajfel, H. (2010). *Social identity and intergroup relations*, volume 7. Cambridge University Press.

Tapscott, D. and Williams, A. D. (2008). *Wikinomics: How mass collaboration changes everything*. Penguin, New York.

Telegraph (2010a). Debate 2010: General election policy debate 2010. `http://debate2010.force.com/`.

Telegraph (2010b). General election 2010: first 'digital debate' on facebook and youtube. `http://www.telegraph.co.uk/news/election-2010/7577518/General-Election-2010-first-digital-debate-on-Facebook-and-YouTube.html`.

The Point (2010). The point: Make something happen. `http://www.thepoint.com`.

Thomas, O. (2008). Diggs secret editors. `http://gawker.com/346263/diggs-secret-editors`.

Tomlin, J. (2003). A new paradigm for ranking pages on the world wide web. In *Proceedings of the 12th international conference on World Wide Web*, page 355. ACM.

Townsend, A. (2010). Drink up nyc: Meet the tiny crustaceans in your water. `http://newsfeed.time.com/2010/08/31/drink-up-nyc-meet-the-tiny-crustaceans-not-kosher-in-your-tap-water/?hpt=C2?`

Trueswell, R. (1969). Some Behavioral Patterns of Library Users: The 80/20 Rule. *Wilson Library Bulletin*.

Turkle, S. (2010). *Alone Together: Why We Expect More from Technology and Less from Each Other*. Basic Books, New York.

Vaccari, C. (2012). *Social Media and Democracy: Innovations in Participatory Politics*, chapter The News Media as Networked Political Actors: How Italian Media are Reclaiming Political Ground by Harnessing Online Participation. Routledge, York.

Van Dijck, J. (2009). Users like you? Theorizing agency in user-generated content. *Media, Culture & Society*, 31(1):41.

Venables, W. N., Ripley, B. D., and Venables, W. (1994). *Modern applied statistics with S-PLUS*, volume 250. Springer-verlag New York.

Virasoro, D., Leonard, P., and Weal, M. (2011). An analysis of social news websites.

Voltier Creative (2010). Reddit's astonishing altruism. `http://voltiercreative.com/2010/11/reddits-astonishin-altruism/`.

Voltier Creative (2011). Reddits astonishing altruism: Part two. `http://voltiercreative.com/2011/04/reddits-astonishing-altruism2/`.

Votorola (2013). Votorola home page. `http://zelea.com/project/votorola/home.html`.

Warren, J. and Jurgensen, J. (2007). The wizards of buzz. `http://online.wsj.com/news/articles/SB117106531769704150`.

Wauters, R. (2009). Url shortening wars: Twitter ditches tinyurl for bit.ly. *Retrieved May*, 16:2010.

Wauters, R. (2010). Why is reddit all over digg right now? `http://techcrunch.com/2010/08/30/why-is-reddit-all-over-digg-right-now/`.

WebProNews (2006). Wikipedia bans congress. `http://www.webpronews.com/wikipedia-bans-congress-2006-01`.

Wilkinson, D. (2008). Strong regularities in online peer production. In *Proceedings of the 9th ACM conference on Electronic commerce*, pages 302–309. ACM.

Willinger, W. and Paxson, V. (1998). Where mathematics meets the Internet. *Notices of the American Mathematical Society*, 45(8):961–971.

Willis, J. and Yule, G. (1922). Some Statistics of Evolution and Geographical Distribution in Plants and Animals, and their Significance. *Nature*, 109:177–179.

Wilson, E. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212.

Wise, K., Hamman, B., and Thorson, K. (2006). Moderation, response rate, and message interactivity: Features of online communities and their effects on intent to participate. *Journal of Computer-Mediated Communication*, 12(1):24–41.

Woyke, E. (2010). $240 to place story on digg front page: One marketing firms pitch. `http://blogs.forbes.com/elizabethwoyke/2010/11/10/240-to-place-story-on-digg-front-page-one-marketing-firms-pitch/`.

Wright, S. (2006). Government-run online discussion forums: moderation, censorship and the shadow of control. *British Journal of Politics and International Relations*, 8(4):550–568.

Wright, S. (2007). A virtual european public sphere? the futurum discussion forum. *Journal of European public policy*, 14(8):1167–1185.

Wright, S. (2012). Politics as usual? revolution, normalization and a new agenda for online deliberation. *new media & society*, 14(2):244–261.

Wright, S. and Street, J. (2007). Democracy, deliberation and design: the case of online discussion forums. *New Media & Society*, 9(5):849–869.

Wu, F., Wilkinson, D., and Huberman, B. (2009). Feedback loops of attention in peer production. In *2009 International Conference on Computational Science and Engineering*, pages 409–415. IEEE.

Yang, D. (2012). Techdirt 2011: The numbers. `http://www.techdirt.com/articles/20120104/23151517284/techdirt-2011-numbers.shtml`.

Yi, K. (2012). Harnessing collective intelligence in social tagging using delicious. *Journal of the American Society for Information Science and Technology*, 63(12):2488–2502.

YourFreedom (2010). Your freedom archive. `http://webarchive.nationalarchives.gov.uk/20100709093439/http://yourfreedom.hmg.gov.uk`.

Zeileis, A., Kleiber, C., and Jackman, S. (2008). Regression models for count data in r. *Journal of Statistical Software*, 27(8):1–25.

Zipf, G. (1949). *Human behavior and the principle of least effort.* Addison-Wesley, Reading.