The following working paper contains excerpts from a chapter of my thesis. I am in the process of writing up aspects of this work for publication and have made this raw text from my thesis available to serve as the basis for discussions.

The purpose of this document is to develop an understanding of how Reddit 'works' based chiefly on data collected through its API.

If you have happened upon this work by chance please feel free to direct any comments to my e-mail address: r.mills@lancs.ac.uk

Richard Mills

1 The functionality of Reddit's Distributed Moderation systems

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 (this is a short-form url which links directly to a post's comments page on Reddit). In this post one user is querying why a story about North and South Korea exchanging artillery fire, which was submitted to the worldnews sub-reddit, 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 sub-reddit. 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.

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 natural 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.com/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 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, 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 1.2 and 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

What is reddit?

reddit is a source for what's new and popular on the web.
Users like you provide all of the content and decide, through voting, what's good (↑) and what's junk (↓).
Links that receive community approval bubble up towards #1, so the front page is constantly in motion and (hopefully) filled with fresh, interesting links.

How is a submission's score determined?

A submission's score is simply the number of upvotes minus the number of downvotes. If five users like the submission and three users don't it will have a score of 2. Please note that the vote numbers are not "real" numbers, they have been "fuzzed" to prevent spam bots etc. So taking the above example, if five users upvoted the submission, and three users downvote it, the upvote/downvote numbers may say 23 upvotes and 21 downvotes, or 12 upvotes, and 10 downvotes. The points score is correct, but the vote totals are "fuzzed".

- **Vote.** The up and down arrows are your tools to make reddit what you want it to be. If you think something contributes to conversation, upvote it. If you think it does not contribute to the subreddit it is posted in or is off topic in a particular community, downvote it.

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

access them (e.g. a 'Reddit Gold' sub-reddit 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 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 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. Section 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 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 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 sub-reddits which *rely on people adopting sub-reddit specific voting norms* if they are to function. 'Ask Me Anything' iterviews (on the /r/IAmA sub-reddit) require users to vote on comments along criteria which are distinct from the other sub-reddits. It is possible to test whether users' collective behaviour on this sub-reddit meets the specific requirements for an 'Ask Me Anything' interview to function. Section 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.

1.1 A power law 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 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 2 gives an initial impression that the distribution of votes between posts may follow a power law. The power law was introduced in section **??**. 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 value k;

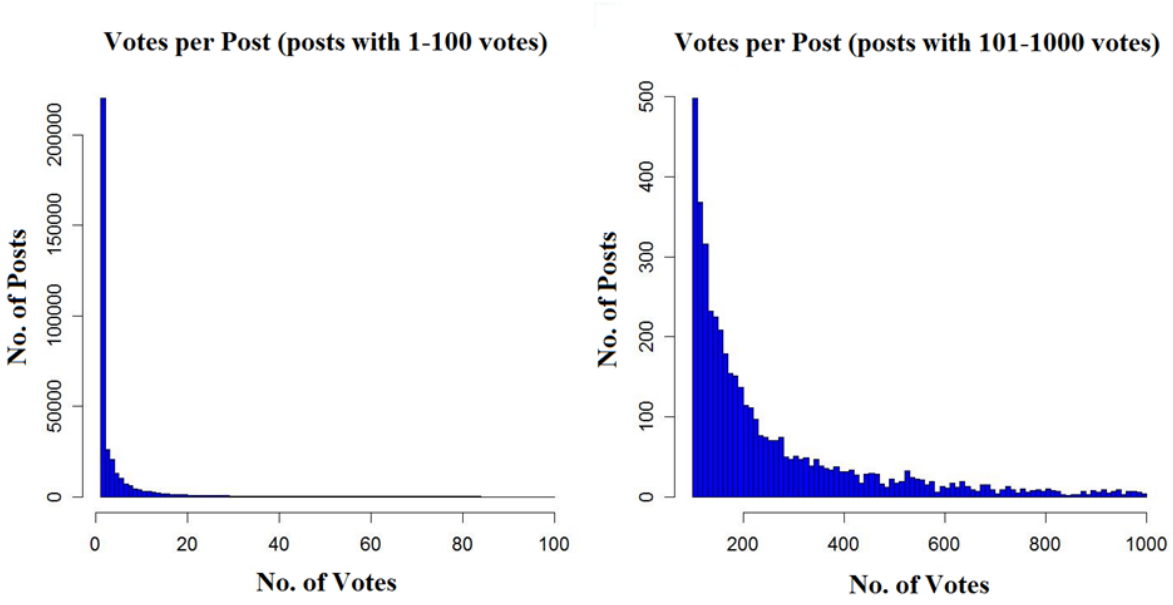

Figure 2: The raw distribution of votes between posts for two sub-sets of posts

- γ is the power-law exponent;
- $\zeta(\gamma)$ is the Riemann zeta function defined as $\sum_{k=1}^{\infty} k^{-\gamma}$.

The method of Clauset et al. (2009) was adopted to test the power law's goodness-of-fit for this data. The first step in doing so was to generate plots of the data on logarithmic axes (see figure 3). The top-left pane of this graph shows raw data over the full 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) than would be expected based on the rest of the distribution.

Separate plots similar to 3 were produced for the 40 most active sub-reddits, 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 sub-reddits separately because in many respects they behave as separate entities. At the time when the procedural data were collected there were twelve default sub-reddits 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 sub-reddits on their Reddit front page. All of the page types employed by Reddit (e.g. new, rising, hot) exist both for individual sub-reddits and in aggregated form (i.e. compiled from all of the sub-reddits 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

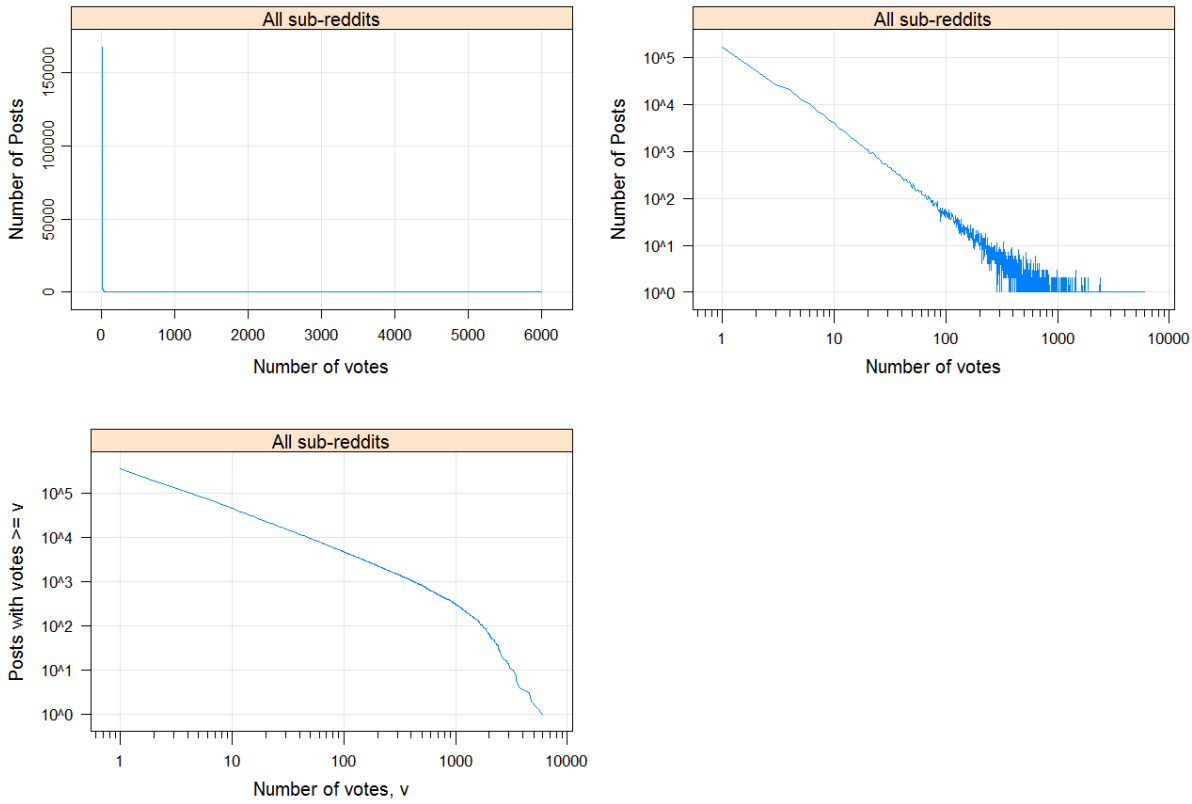

Figure 3: Plots of the frequency distribution of votes across posts

into accounts can easily manage their sub-reddit subscriptions. For this reason it is prudent to treat sub-reddits 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 sub-reddits, as well as to the full data-set incorporating all sub-reddits.

Table 1 contains estimated parameters for fitting a power law to a selection of the most active sub-reddits. The columns on the right of the table relate to the power law's fit; Clauset's fitting procedure involves establishing a minimum cut-off point for the distribution below which the counts are not considered, denoted here as Vmin - all possible values of Vmin are fitted and the exponents are generated, with the value of Vmin which minimises the Kolmogorov-Smirnov goodness-of-fit statistic being selected. γ is the exponent of the power law, D is the Kolmogorov-Smirnov goodness-of-fit statistic and ntail is the number of posts with votes greater than Vmin.

Assessing whether the power law provides a plausible fit for the data is not straightforward; even when the underlying distribution is in fact a power law observed data points will rarely follow this form precisely. To assess whether the power law provides a plausible fit samples are taken from a synthetic data-set which does follow the power law, and the fit of these synthetic samples is compared to the empirical data in question. This procedure was carried out 1,000 times for each of the sub-reddit power law fits in order to generate a p-value that should be reliable. These p-values are reported in table 1.

It can be seen in table 1 that the power law provides a plausible fit for some but not all of the sub-

sub-reddit	Posts	Posts with 1 vote	Votes	Vmin	γ	D	ntail	p
reddit.com	150041	104503	516775	17	1.91	0.013	2042	0.043
pics	16511	5714	315569	10	1.85	0.012	3190	0.073
politics	16627	2396	302025	23	1.95	0.014	1640	0.284
funny	10035	2877	195500	9	1.81	0.022	1712	0.007
WTF	10410	2855	185315	14	1.82	0.021	1279	0.016
worldnews	10834	5479	140648	10	1.91	0.017	1644	0.08
technology	11875	6247	138776	9	1.82	0.021	1265	0.09
programming	4273	615	132914	5	1.73	0.042	2780	0.000

Table 1: Frequency counts and power law fit paramaters for the most active sub-reddits. A p-value of greater than 0.05 means that the power law is *not* a probable fit for the distribution of votes on the sub-reddit.

reddits tested. There is no discernible reason why the power law fits for certain sub-reddits but not for others. One working hypothesis is that the discrepancy arises due to the extent that individual sub-reddits operate in isolation from the rest of the website. For example, the sub-reddit for which a power law offers the best fit is 'politics'. Observation of Reddit suggests that the politics sub-reddit may be quite independent as compared to other large sub-reddits - users often comment on the 'liberal' bias of /r/politics and it is often cited (along with /r/atheism) as a sub-reddit to un-subscribe from. However, the sub-reddit for which the power law offers the worst fit is 'programming', and this sub-reddit likely also has this same characteristic of being independent of the other default sub-reddits.

Likelihood-ratio tests were conducted to compare the goodness of fit of the power law distribution with that of some alternative distributions. For all sub-reddits the power law distribution offered a better fit than the Poisson or Exponential distributions. It is possible to pursue this further, and indeed there is evidence that a power law distirbution with an exponential cut-off would offer a better fit than a pure power law. However, taking such a step adds unnecessary complexity in the current context because there is little to be gained in understanding how Reddit functions. The important point is that the distribution of votes between posts is highly skewed and approximates a power law - estimating the power law's parameters is of utility because other researchers have on occasion published articles on the power law's fit for data from other websites (e.g. Digg, Youtube).

These analyses highlight the importance of Reddit's layered and multi-threaded structure. Despite the website's over-arching visual theme the sub-reddits it is comprised of are far from homogeneous. Rather, each sub-reddit is better thought of as its own distinct entity or process. There are thousands of these separate processes operating in tandem, with several hundred (in 2009, this figure would be much larger in 2012) 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 sub-reddit subscription system individual users can choose which of these 'output streams' they find useful or interesting, allowing these to feed into their Front page. When an individual sub-reddit 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 sub-reddit is going in can unsubscribe, or the administrators can choose to remove it from the list of default sub-reddits.

This kind of compartmentisation 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 'Reddit.com' sub-reddit; a legacy from a time when sub-reddits did not exist, and by far the most active sub-reddit; actually surpasses this limit and has too many participants. The row of table 1 representing this sub-reddit suggests that this may be the case; of 150,042 posts submitted to this sub-reddit 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 'Reddit.com' sub-reddit 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 national 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 sub-reddit being added to the list of default sub-reddits; where the level of attention and voting activity accruing to submitted items shows an instant and substantial increase.

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 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 sub-reddits they subscribe to. For users not signed into an account these posts are drawn from the sub-reddits 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 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.

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 sub-reddits. In aggregated form these tabs display content from all of the sub-reddits a user has subscribed to, users who are not signed into an account are shown content from the default sub-reddits.

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 sub-reddit is referred to here as the sub-reddit's 'Main page' (because it is displayed upon navigating to the sub-reddit). 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 we need 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 we observe a hypothetical post it is on the New page and it has 3 up-votes, 0 down-votes and 1 comment. The second time we observe the post it appears on both the New and Rising pages and it has 5 up-votes, 1 down-vote and 2 comments - so we record 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 account of 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.

In this section we consider 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 we were looking at the number of new up-votes per observation we might find that at one 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. For our purposes this proved highly problematic as it is a source of considerable noise in the data.

To circumvent this problem we 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 sub-reddits. 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 us to fit 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).

As we are relying on the Front, Main, New and Rising pages for all of our information on a post's progress 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, when we consider the page(s) a post appeared on we initially only consider the first 'leaf' - for a post to be counted as appearing on a page it must appear there with a rank in the top 25.

1.3.1 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' sub-reddit 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 $V_i \sim \text{Negative Binomial}(\hat{V}_i, \theta)$

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 2 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 sub-reddit 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 2: 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 sub-reddit - 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 sub-reddit and how it is received when it appears there. This is a question which will be addressed further in section 1.5.

1.3.2 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 3) and comments (table 4).

	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 3: 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.

	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 4: 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.

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 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 sub-reddit.

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.

1.3.3 '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 sub-reddit 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 sub-reddit during this window it appears that there were never more than 25 posts classified as Rising at any observation point.

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. 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 sub-reddit'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

	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: 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

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.

1.3.4 Comparing other default sub-reddits

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 sub-reddits. There is a high degree of consistency in most of the patterns in voting activity between these sub-reddits, suggesting that these patterns are universal on Reddit amongst the

default sub-reddits.

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 sub-reddits 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 sub-reddit. Shortly after the data analysed here was collected a website was launched (stattit.com) which displays activity metrics for sub-reddits - /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 sub-reddits). If we compare the data collected in this observation period from the /r/worldnews and /r/funny sub-reddit pages there is a clear difference in the number of posts being submitted - there are records of 9598 individual posts to /r/worldnews and 124,367 individual posts to /r/funny. As an aside, casual observation of /r/worldnews at this stage indicates that the sub-reddit is in decline - this will be investigated further in Chapter 8 dealing with longitudinal trends on Reddit.

	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 6: Showing model parameters for a negative binomial regression model of number of new up-votes by page (and leaf thereof) for the /r/funny sub-reddit - 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 sub-reddit is presented in table 6. 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 sub-reddit has much higher activity rates on pages which are specific to the sub-reddit, but when posts from these sub-reddits 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 sub-reddit 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 surprassed /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 plenty of 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 sub-reddit is in a much better state of 'health' than the /r/worldnews sub-reddit. 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 sub-reddits 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 sub-reddit, 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 sub-reddit 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 sub-reddits 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 sub-reddits). 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 sub-reddit) 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 sub-reddit 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.

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. As such they serve to focus the attention of their users and visitors on the best content (as judged by Reddit users 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' sub-reddit 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 7.

This model suggests that an item submitted to the 'pics' sub-reddit 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' sub-reddit 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

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

which appeared on the Rising page. The reason for this is suggested by figure 5, showing the lifespan of a single post on Reddit.

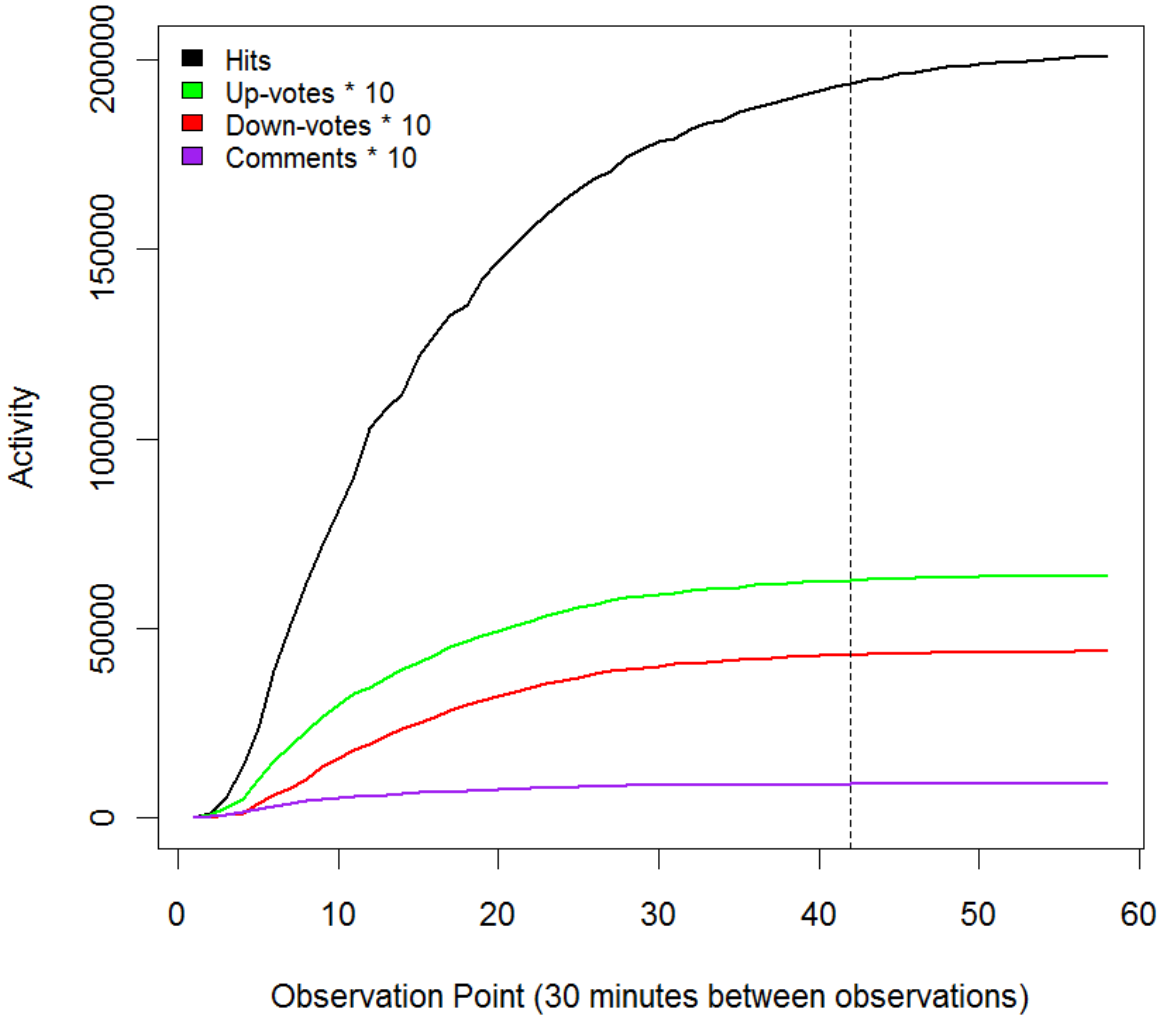

Figure 5: Showing cumulative Hits, Votes and Comments for a single /r/pics Front page Post throughout its lifespan on Reddit. This post appeared on the Front page first at observation point 3 and remained there until observation point 42 (represented by dashed vertical line).

Figure 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 sub-reddits 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 sub-reddit and 0.13 votes per minute for a post on the Rising page of its sub-reddit. Unfortunately shortly after the data on hit-rates were collected Imgur fundamentally 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.

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' sub-reddits.

In section 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 sub-reddits). 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 sub-reddit it is submitted to. Beyond the New page placement on any other page must be earned through the accumulation of votes. 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 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 sub-reddits) 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 1.3 we saw a considerable difference in the 'depth' of voting activity on the /r/funny and /r/worldnews sub-reddits (/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 sub-reddit which reached each page in the sequence. Figure 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 6 - sub-reddits which show the expected drop in posts which reach each page in the sequence (e.g. /r/funny) and sub-reddits on which very few posts have been excluded from appearing on the Rising and Main pages (e.g. worldnews).

For sub-reddits 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 sub-reddits, whereas the Rising page seems to display a certain percentage (70-80%) of posts which appeared on the New page. Therefore whether a sub-reddit exhibits the 'hurdling' effect appears to be governed by its rate of post submissions.

Looking at the sub-reddits which do not exhibit the 'hurdling' effect (movies, science, technology, worldnews) reveals two things. First, these sub-reddits 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 sub-reddits that exhibit the 'hurdling' effect. This second point in particular suggests that the New and Rising pages for these sub-reddits are not contributing much to the decision about which posts go on to the Front page - most of the posts submitted to these sub-reddits 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 sub-reddits 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 sub-reddits leads to a better decision about which posts should progress in the sequence and appear on the Main page - as compared to sub-reddits 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. 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

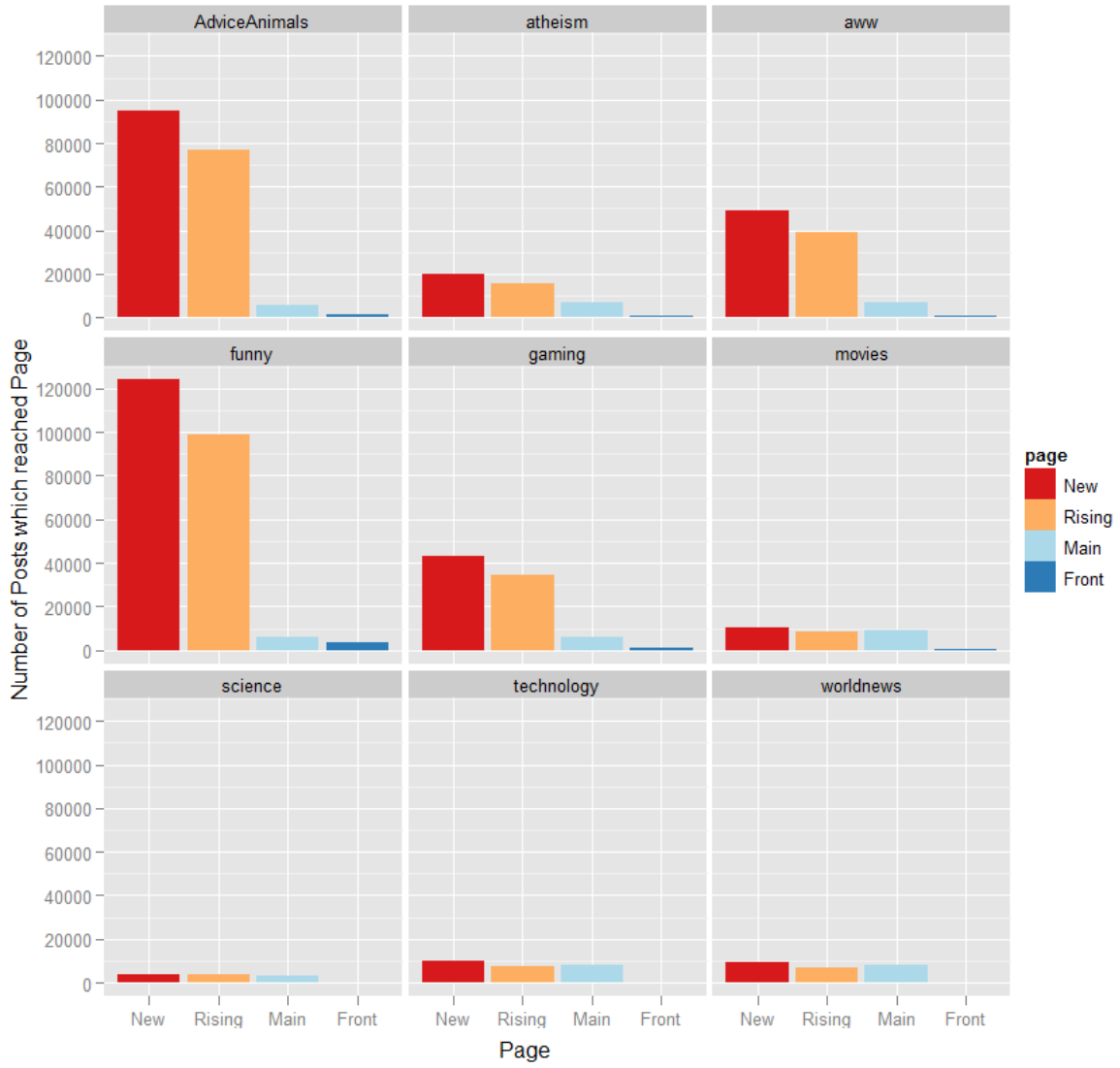

Figure 6: Showing the number of unique posts which appeared on the four observed page types for a number of sub-reddits - August 20th to September 27th 2012

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 sub-reddits 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 sub-reddits as compared to posts which appear on the

Main pages of sub-reddits 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 sub-reddits that exhibit hurdling.

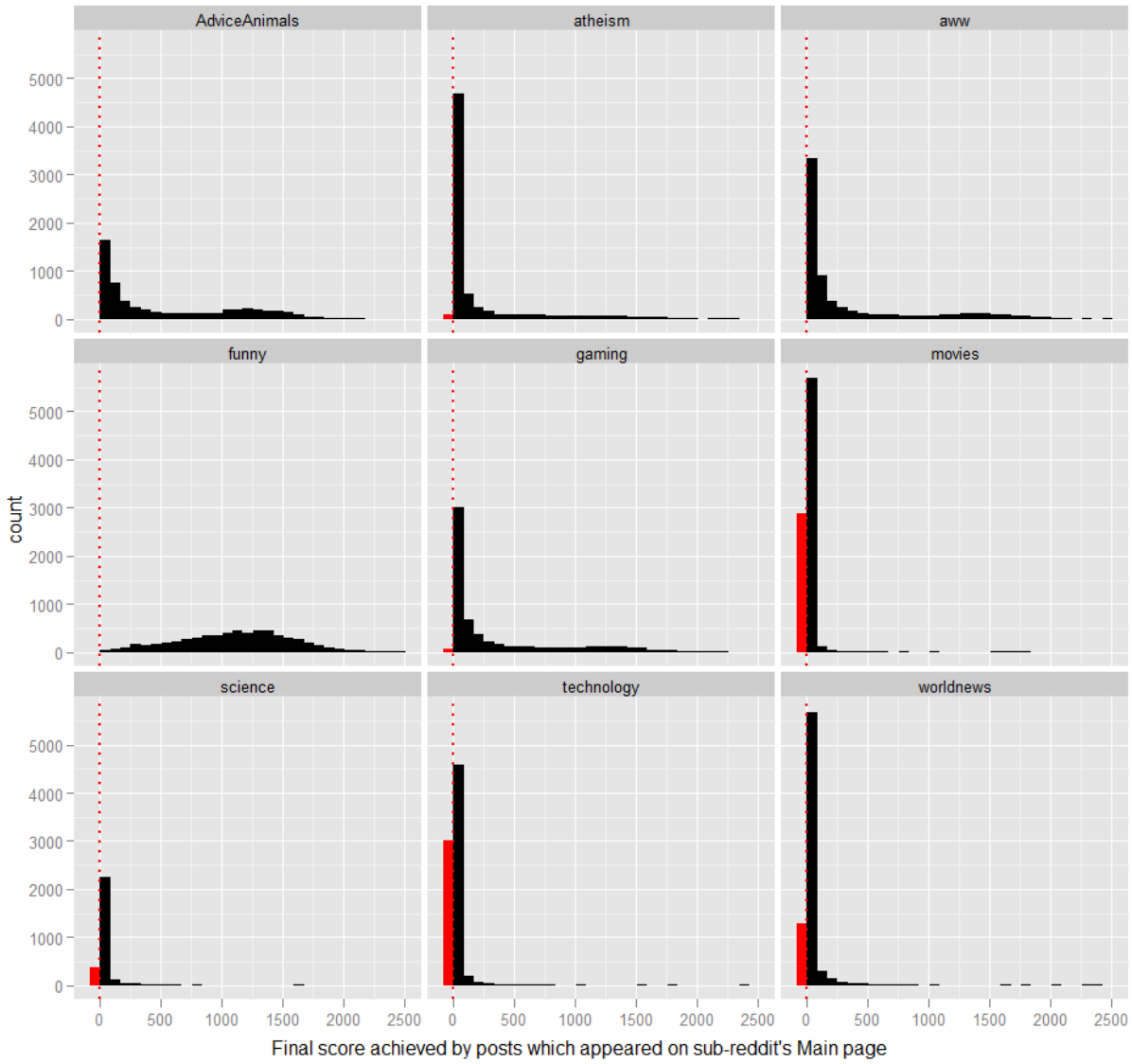

Figure 7: Showing the final scores achieved by posts which appeared on the Main page (leafs 1-4) of their sub-reddit. Excludes 132 posts with a final score of greater than 2500. Bars representing posts with a score of zero or less are coloured red.

Figure 7 reveals exactly such a trend. Sub-reddits 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 sub-reddit with the hurdling effect and then went on to finish with a negative score are very rare. In contrast this happened regularly on sub-reddits 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 sub-reddit that end with a negative score but table 8 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.

Sub-reddit	% 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 8: Showing the percentage of all posts, and of posts which appeared on the Main page for their sub-reddit, which had a negative score at final observation

Returning to figure 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 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: sub-reddits 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 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 sub-reddit's Main page. This is in contrast to a sub-reddit 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 sub-reddits likely results in a distribution of scores which is less 'stretched' towards the high end of the distribution.

1.5.1 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 sub-reddits under consideration - the effect was not significant for the /r/worldnews sub-reddit.

The effect was strongest for the /r/funny sub-reddit, 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 sub-reddits with the 'hurdling' effect and a relatively high voting rate on New and Rising, and weaker for sub-reddits 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 sub-reddits 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 sub-reddits (e.g. this model for /r/funny accounts for 14% of the deviance in whether a post reached the Front page).

Table 9 shows the model coefficients for these logistic regression models. There is quite a strong degree of similarity between the sub-reddits in terms of the effect of a post's score at first observation being negative, neutral or positive. For all sub-reddits 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 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.

Sub-reddit	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 9: 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.

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 sub-reddits). 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. Sub-reddits 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 sub-reddits 'bad' posts are filtered out before they appear on the Main page of the sub-reddit - and on the Main page posts of reasonably high quality compete to reach the Front page. For a number of the less active default sub-reddits 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 sub-reddits. 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 sub-reddits 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 sub-reddit (a sub-reddit 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 sub-reddit (a sub-reddit 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 sub-reddits means that posts from light-hearted or entertaining sub-reddits are not competing directly with those from more serious sub-reddits - the top-ranking post

for /r/worldnews always has a spot on the Front page but there is some 'meta-competition' between sub-reddits whereby sub-reddits 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 sub-reddits 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, it doesn't take 100x longer to vote on an article as compared to a joke. Part of the problem for these sub-reddits, 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 sub-reddits (e.g. /r/worldevents) over time in response to a perceived decline in the quality of the sub-reddit.

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 their own right in the Chapter XX.

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'.

1.7.1 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).

1.7.2 '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 sub-reddit 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 sub-reddit 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 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.

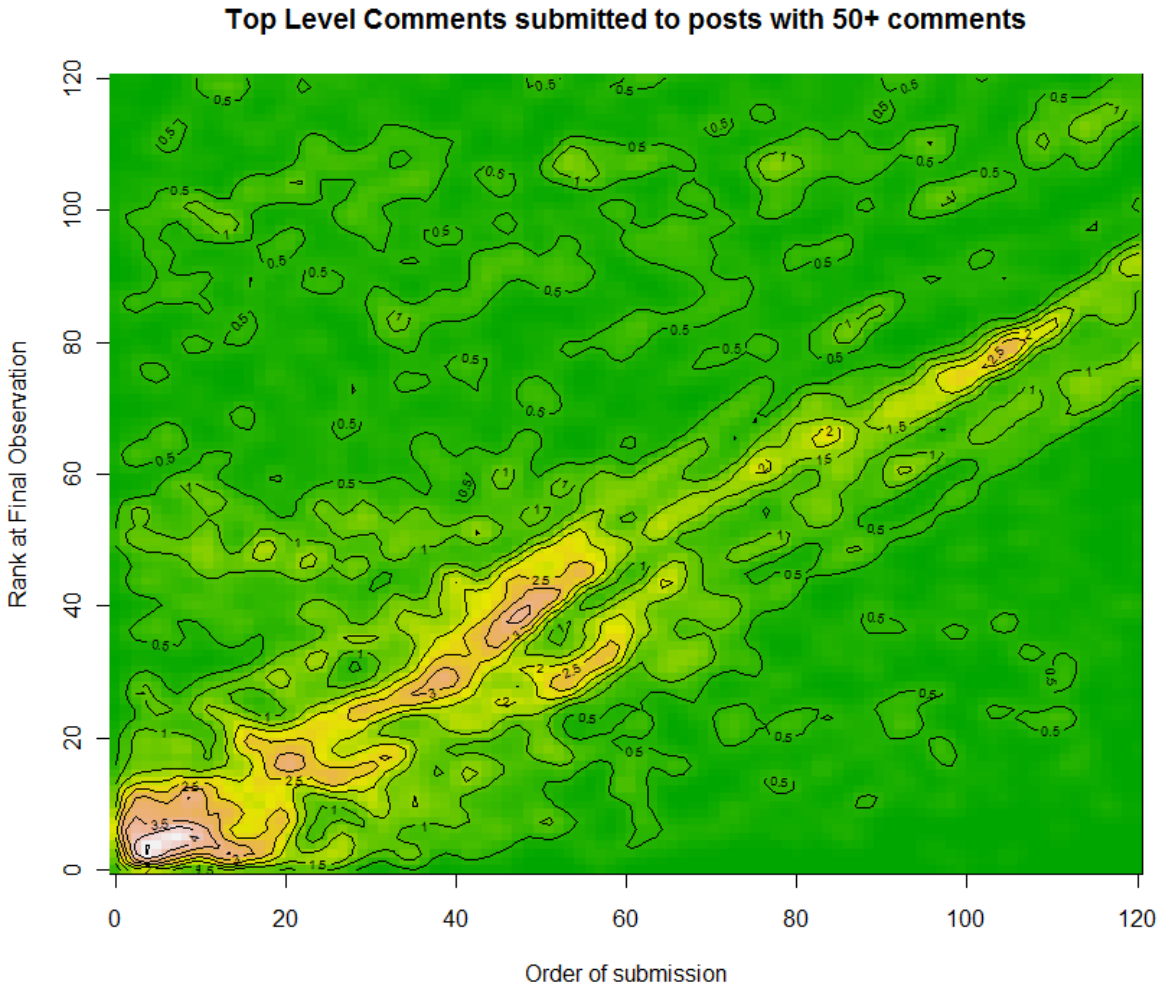

Figure 8: A kernel density plot showing submission order against final observed rank

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.

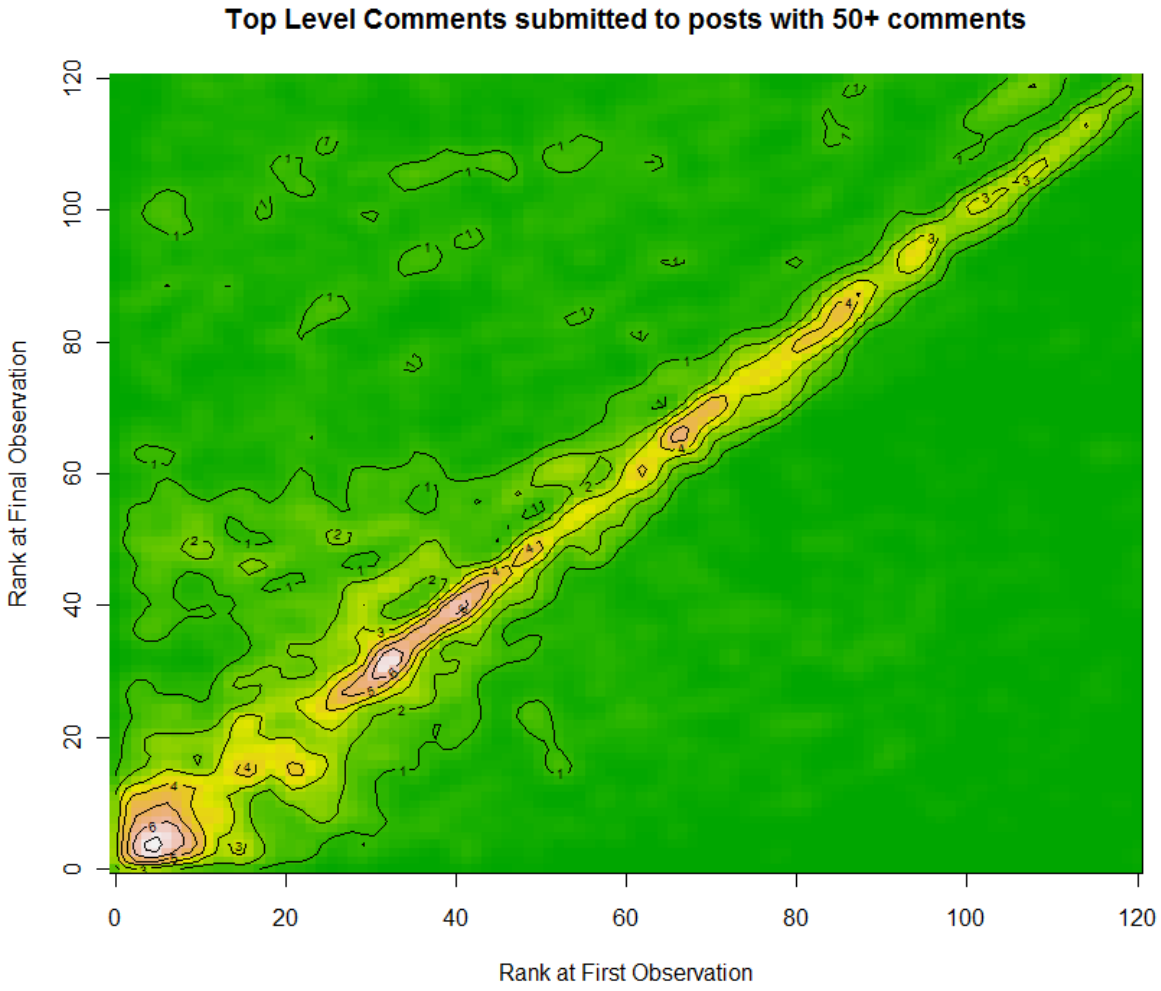

Figure 9: A kernel density plot showing first observed rank against final observed rank

If we instead compare the first observed rank for these comments with their final observed rank a very different picture emerges. Figure 9 suggests much lower mobility for comments between their 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 8 and 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 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 8 and 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.

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.

1.9 The comments page for a high-activity post (redd.it/x8kra)

Figure 10 shows the number of new comment votes by comment level over time for a single post to the politics sub-reddit (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

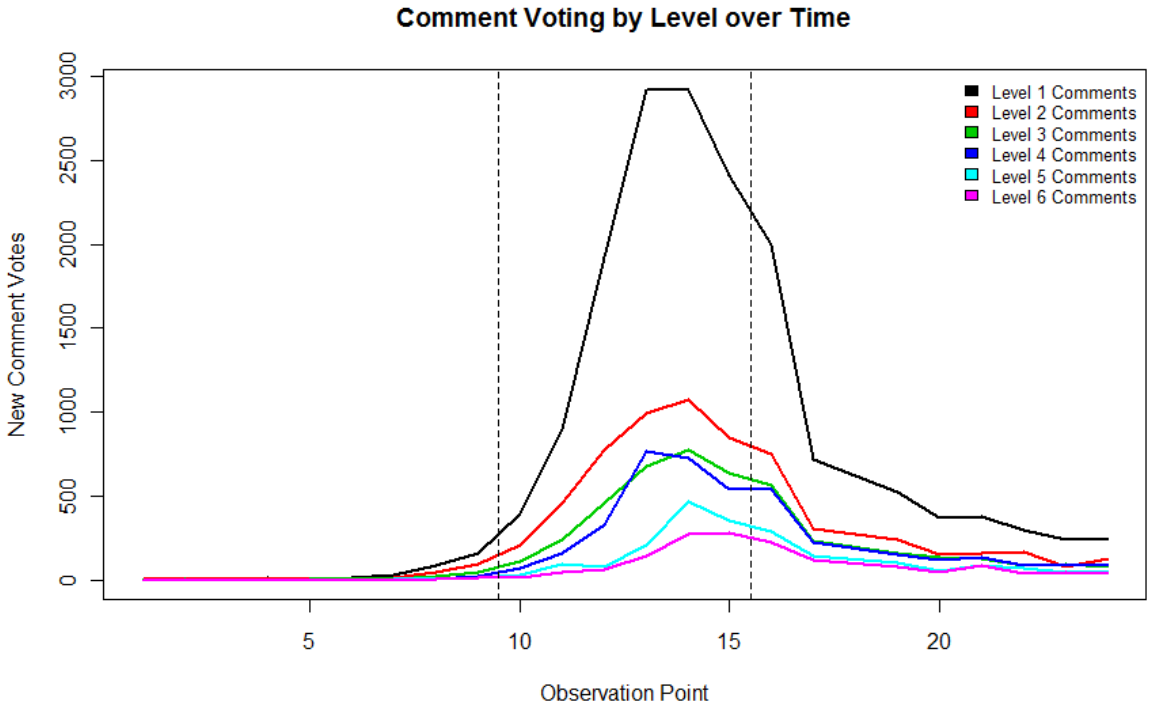

Figure 10: Showing new votes by comment level over time for post redd.it/x8kra - dashed vertical lines represent the approximate points at which the post appeared on the Front page for the first and last time

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. 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 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 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

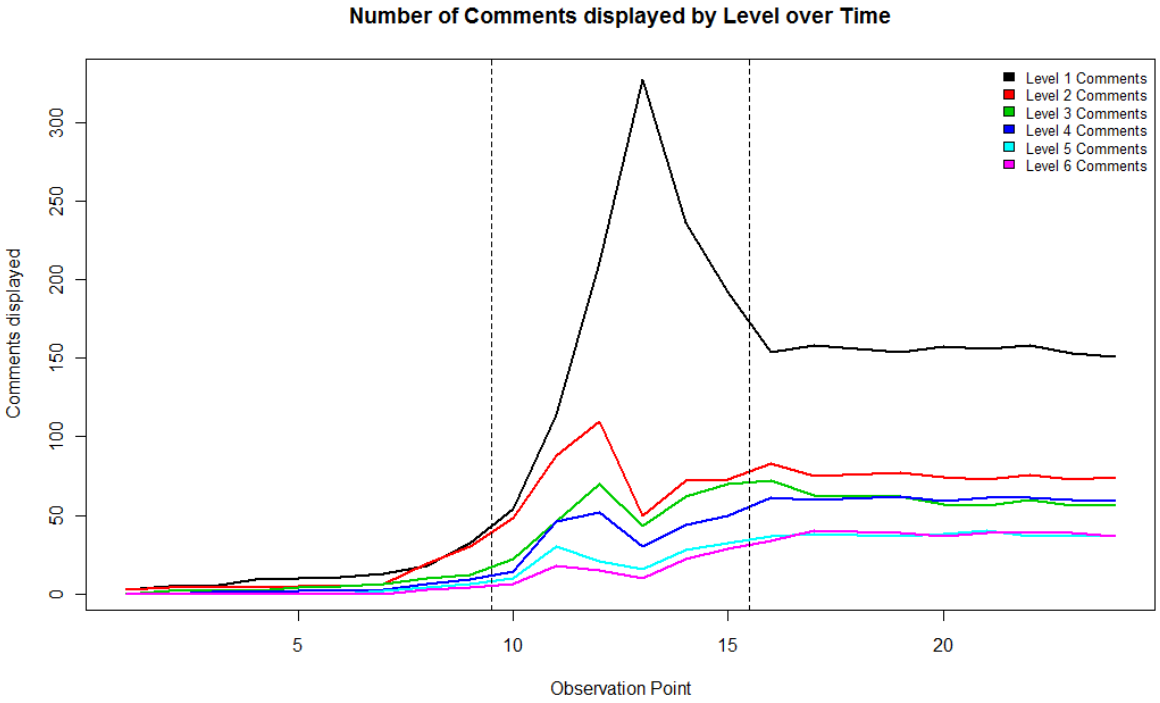

Figure 11: Showing number of comments displayed by level over time for post redd.it/x8kra - dashed vertical lines represent the approximate points at which the post appeared on the Front page for the first and last time

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).

The Kernel density plot of first observed rank against last observed rank (Figure 9) suggested that comment ranks tend to be quite stable between first and last observation. This is somewhat at odds with figure 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 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

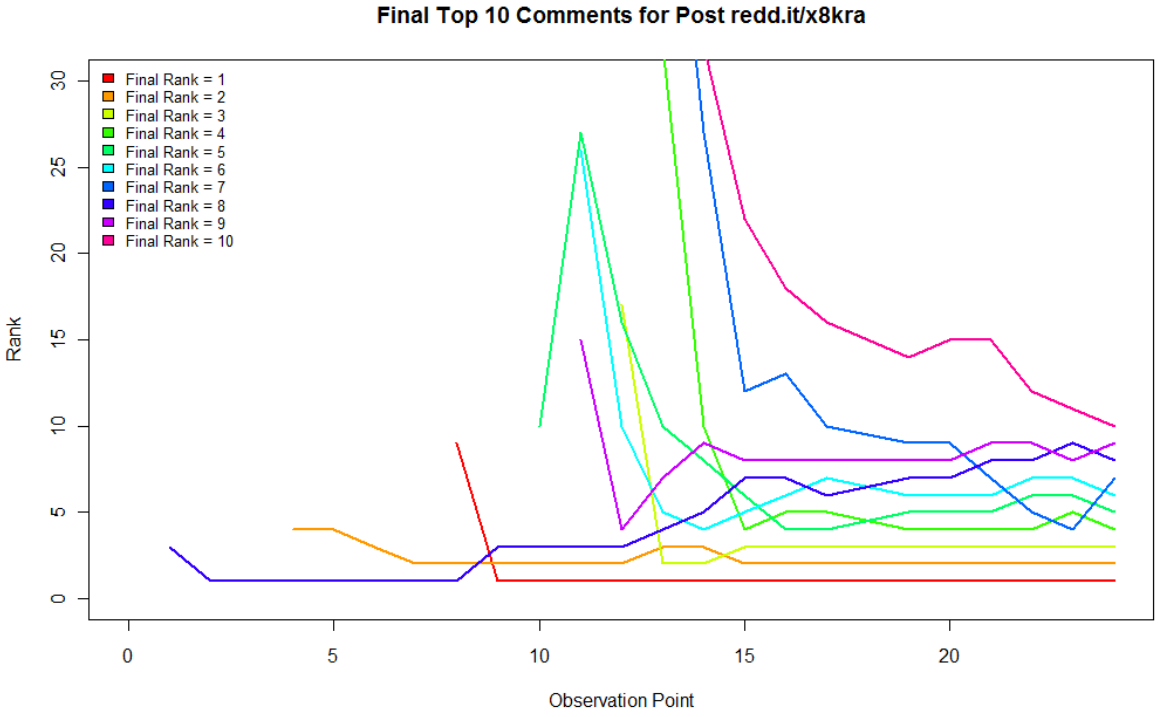

Figure 12: Showing the ranks over time for the top-level comments on post redd.it/x8kra with final rank 10 or better

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 top 20 positions.

However, Figure 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 12 with figure 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.

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 sub-reddits 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 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 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 sub-reddit). 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.

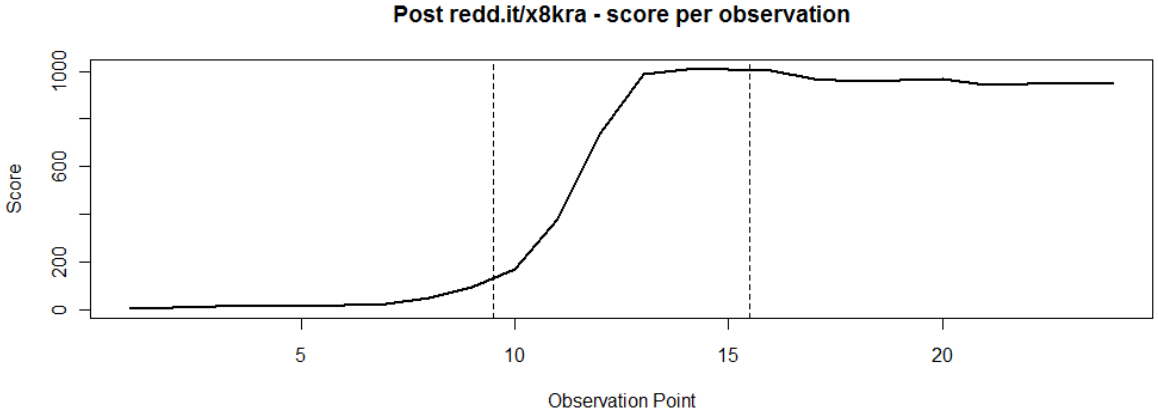

Figure 13: Showing the score over time for the post redd.it/x8kra - dashed vertical lines represent the points where the post appeared on the Front page for the first and last times.

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 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 XX.

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 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).

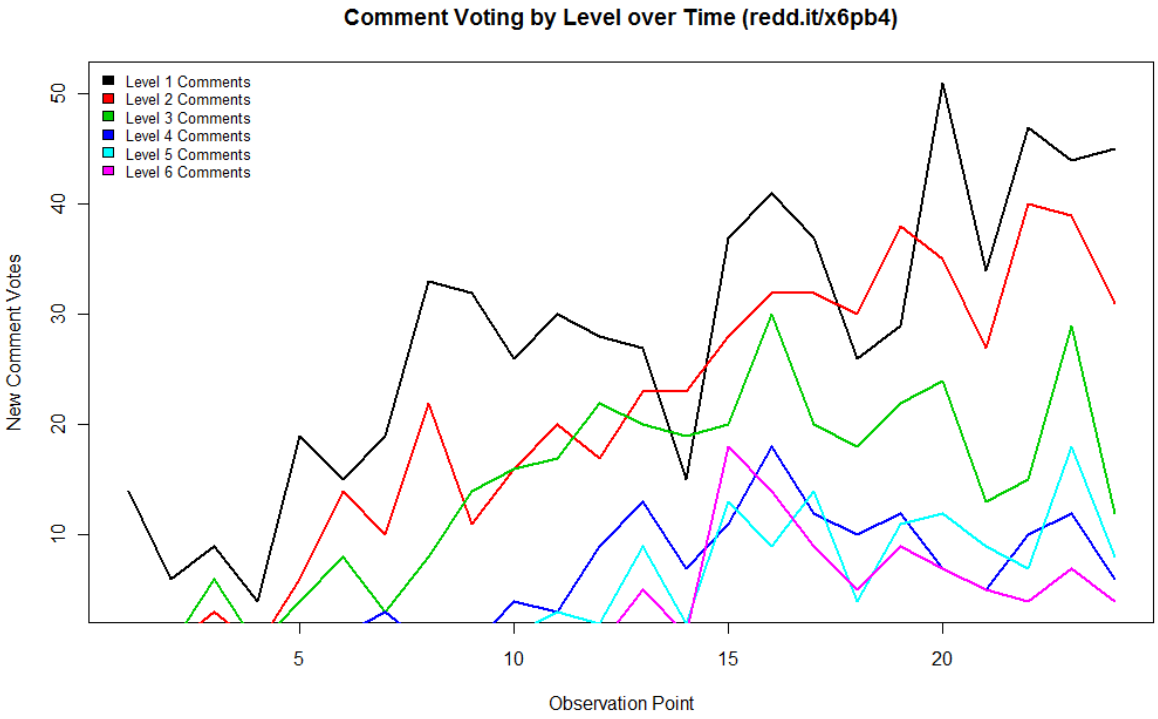

Figure 14: Showing new votes by comment level over time for post redd.it/x6pb4

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 15). This post did not experience the surge 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 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.

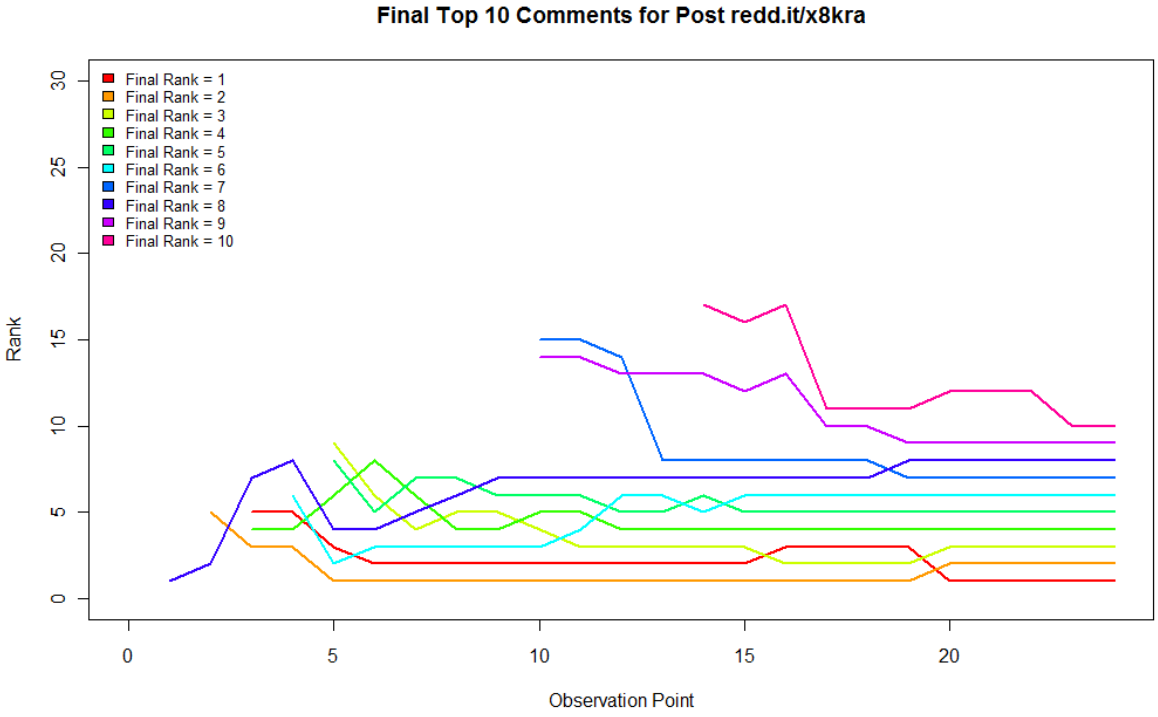

Figure 15: Showing the ranks over time for the top-level comments on post redd.it/x6pb4 with final rank 10 or better

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

The /r/IAmA sub-reddit 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 tha 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 sub-reddit 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 sub-reddit'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 sub-reddit'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 sub-reddit, 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 sub-reddit 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 sub-reddits' 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 sub-reddit 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 is at that stage *not* one of the top-scoring comments - then it is hypothesised that the top-level comment which was repled 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 sub-reddit - 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.

1.12.1 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).

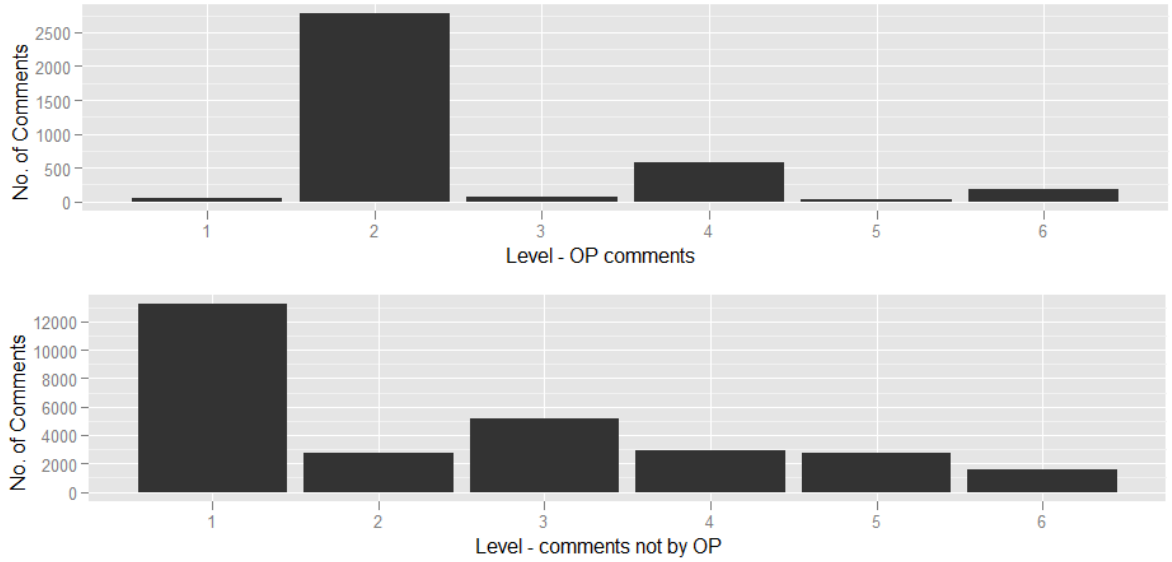

Figure 16: Showing the level of comments submitted to Front page IAmA posts - for comments by the OP and by other users.

Figure 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 sub-reddits - 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 our hypotheses relates to this. Where there are comments from non-OP users at level 2 we would

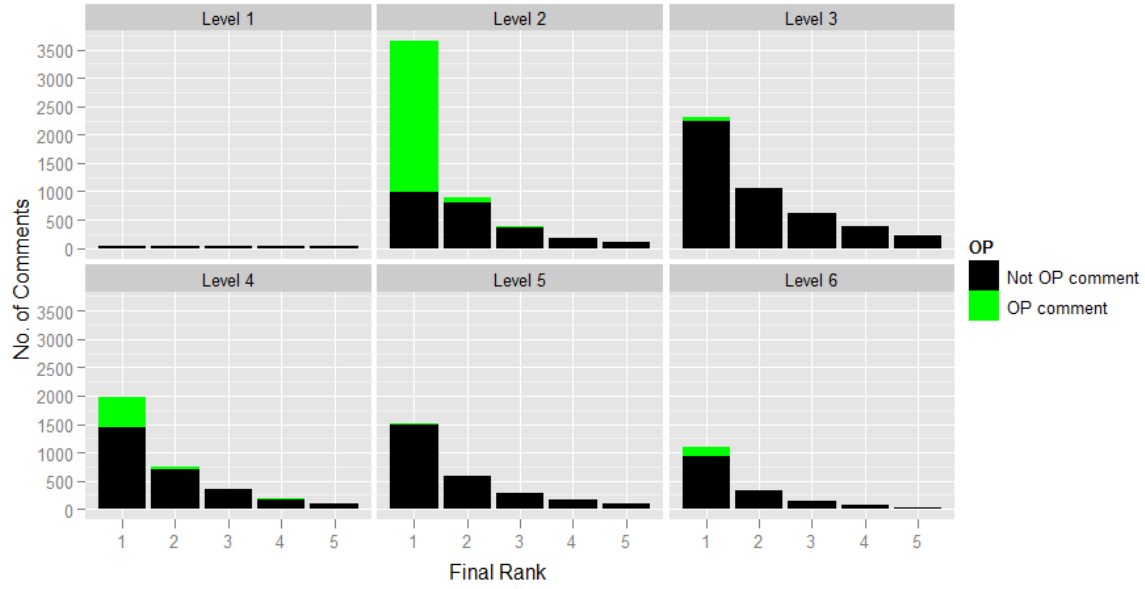

Figure 17: Showing the final ranks of comments submitted to Front page IAmA posts by level. Comments with final rank greater than 5 have been excluded, N = 17,736. The bars at level one are all equal to 42 because a given post can only have one comment at each rank, whereas at level 2 there are as many comments at rank 1 as there are level 1 comments which have been replied to.

expect the OP comments to be ranked most highly. Figure 17 suggests that this is almost 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.

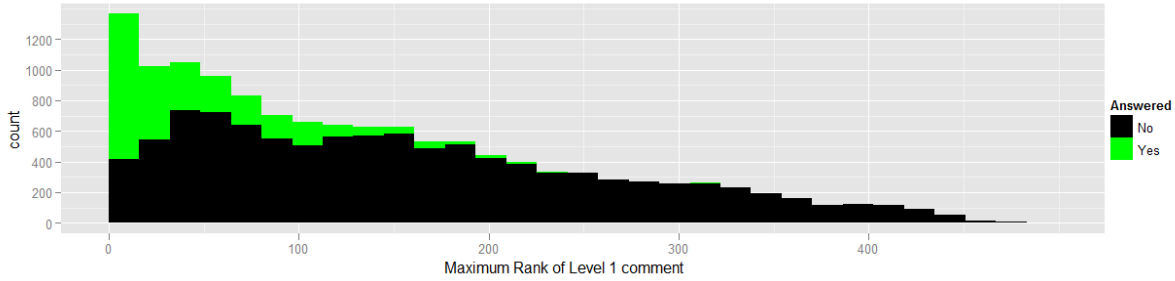

Figure 18: Showing the maximum ranks of level 1 comments and whether they received a response from the OP

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 - instead we will only consider the presence or absence of a response from the OP, where a response is present this will be

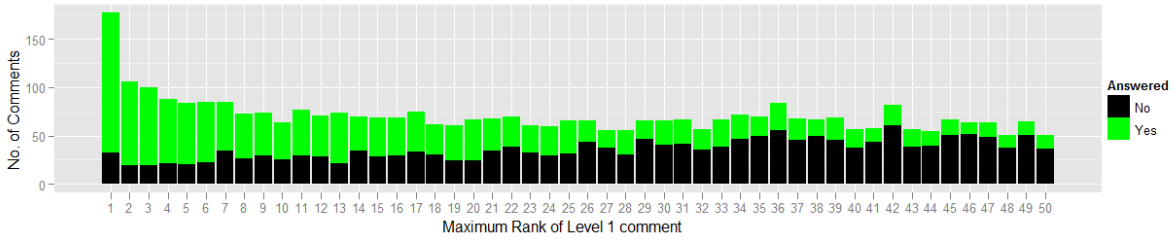

Figure 19: Showing the maximum ranks of level 1 comments which appeared at some stage in the top 50 comments - and whether they received a response from the OP

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 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 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 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.

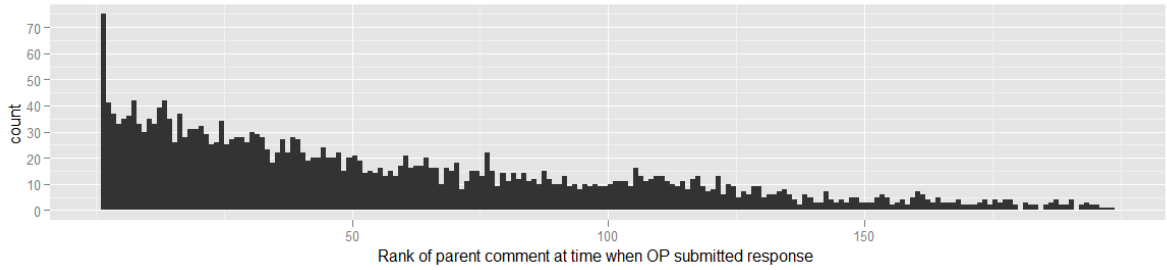

Figure 20: Showing the ranks of level 1 comments at the observation point before they had a response from the OP

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 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 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 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 5974 observations (which come from 1198 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 10 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 observation than the scores of comments ranks 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

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 10: 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.

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).

1.12.2 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

sub-reddits 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 sub-reddit'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 21, 22 and 23 suggest that the same trends are present and perhaps even stronger than in the Front page data.

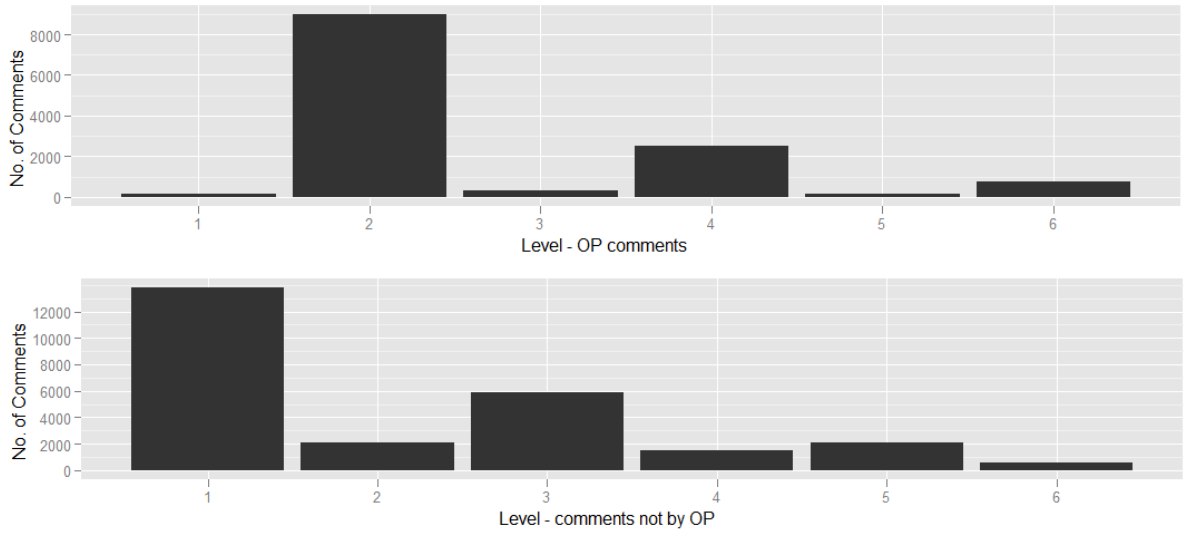

Figure 21: Showing the level of comments submitted to Main page IAmA posts - for comments by the OP and by other users.

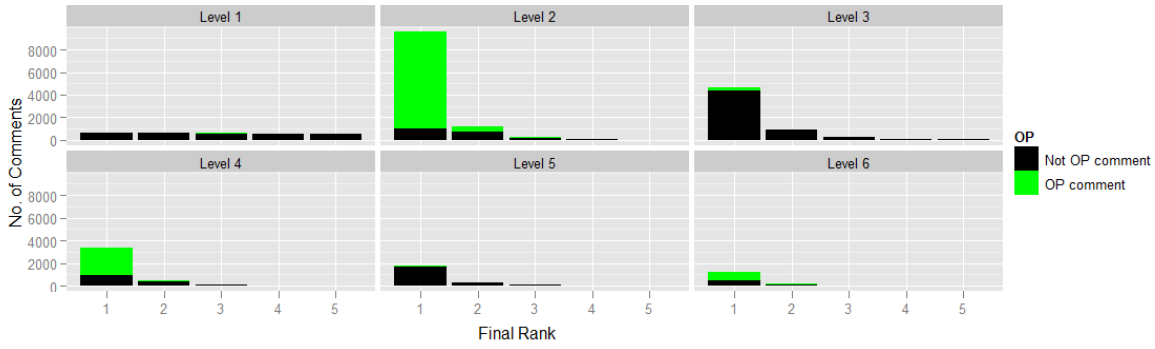

Figure 22: Showing the final ranks of comments submitted to Main page IAmA posts by level. Comments with final rank greater than 5 have been excluded, N = 12,875.

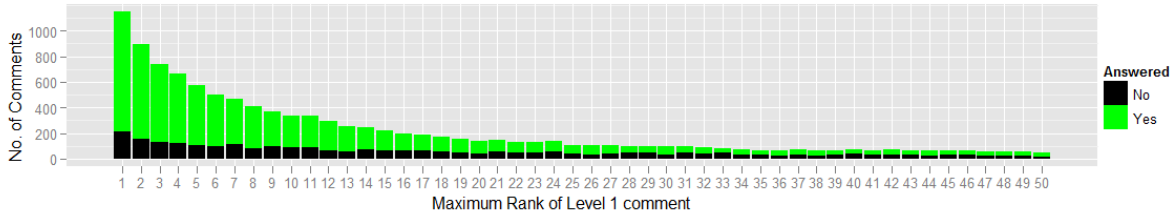

Figure 23: Showing the maximum ranks of level 1 comments which appeared at some stage in the top 50 comments - and whether they received a response from the OP

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 attempts 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 sub-reddit. As technical component is held constant across all sub-reddits, these distinct patterns of behaviour can only be the result of different social norms about how one votes on this sub-reddit. 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 sub-reddits 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.

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 sub-reddits) 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 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 probably also more posts to vote on). There have also been direct interventions from the moderators of some sub-reddits (/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 **??**.

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 sub-reddit subscriptions and to unsubscribe from many of the default sub-reddits. The evidence for this is chiefly observed in popular comments where users advise others to un-subscribe from default sub-reddits (particularly /r/politics and /r/atheism) and that Reddit is best when one subscribes to smaller sub-reddits which match one's interests - the opinion that default sub-reddits have been declining in quality is also regularly voiced. Figure 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 sub-reddits in descending order. As expected, the top 20 sub-reddits 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 sub-reddit (/r/lifeprotips).

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 24: A screen capture from redditmetrics.com taken on 14th December 2012 - showing the number of new subscribers for sub-reddits in descending order

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.

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 sub-reddits 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 sub-reddits). This was found to produce a 'hurdling' effect on the highly active sub-reddits whereby posts had to progress through several pages (New -> Rising -> Main -> Front) and many were filtered out at each stage. The hurdling effect was much weaker for less active sub-reddits and this resulted in a larger number of 'poor' (i.e. negatively scoring) posts appearing on the Main page for these sub-reddits.

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 **??** 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 whose behaviour can be accurately predicted.

References

A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *Siam Review*, 51(4):661–703, 2009. Times Cited: 14.

Danny Dover. How reddit ranking algorithms work. web, Jun 2008. URL `http://www.seomoz.org/blog/reddit-stumbleupon-delicious-and-hacker-news-algorithms-exposed`.

ketralnis. Nerd talk: The tale of the life of a link on reddit, told in graph porn. web, Jul 2011. URL `http://blog.reddit.com/2011/07/nerd-talk-tale-of-life-of-link-on.html`.

Randall Munroe. reddit's new comment sorting system. web, Oct 2009. URL `http://blog.reddit.com/2009/10/reddits-new-comment-sorting-system.html`.

Amir Salihefendic. How reddit ranking algorithms work. web, Nov 2010. URL `http://amix.dk/blog/post/19588`.

E.B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

A. Zeileis, C. Kleiber, and S. Jackman. Regression models for count data in r. *Journal of Statistical Software*, 27(8):1–25, 2008.