Widespread Underprovision on Reddit

Eric Gilbert
School of Interactive Computing & GVU Center
Georgia Institute of Technology
gilbert@cc.gatech.edu

ABSTRACT

Many online communities ask their members to do work for the good of everyone on the site. On social voting sites like Reddit, this means that users judge a stream of incoming links by voting them up or down. The links with the most up-votes bubble up to the main page, pointing everyone toward the best content. A threat to all sites designed this way, however, is underprovision: when too many people rely on others to contribute without doing so themselves. In this paper, we present findings suggesting that widespread underprovision of votes is happening on Reddit, arguably the internet's largest social voting community. Notably, Reddit overlooked 52% of the most popular links the first time they were submitted. This suggests that many potentially popular links get ignored, jeopardizing the site's core purpose. We conclude by discussing possible reasons behind it, and suggest future research on social voting sites.

Author Keywords

social navigation; reddit; voting; free riding; underprovision

ACM Classification Keywords

H5.3. Group and Organization Interfaces; Asynchronous interaction; Web-based interaction.

INTRODUCTION

Reddit, one of the most visited sites on the web[1], is a social voting site that calls itself the "voice of the internet." In the site's own words,

> 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. [23]

In short, since you can't follow everything published on the web, you call upon the aggregate wisdom of Reddit's votes to find places on the web worth visiting.

Many other online communities also employ this design pattern, known in the academic literature as "social navigation" [7, 8]. Sites like Digg [17] and Hacker News [11] have affordances nearly identical to Reddit's. On Slashdot, members vote to moderate comments [16]. For every product review it receives, Amazon asks its customers, "Was this review helpful to you?" The reviews with the most helpful votes jump to the top so that other customers see them first [10]. On eBay, buyers rate sellers (and vice versa) in the hope of separating the reputable ones from people you shouldn't trust [13]. These votes are public goods, provided for the benefit of everyone in the community [19].

However, what if too few people vote? On Reddit, for example, people might choose to simply visit the popular content without voting on the stream of new links pouring into the site. It takes time and energy to wade through all those new links. After all, if most of them were good, we wouldn't need Reddit. If too many people free ride off of everybody else, the quality of the site will eventually degrade, perhaps to the point where Reddit can no longer surface the best content submitted to the site.

All social production communities experience this situation to an extent [4], a concept known as underprovision. Despite the threat of underprovision quickly turning into a "tragedy of the commons" [12], many online and offline communities overcome widespread underprovision via a combination of social norms, repeated interaction and reputation mechanisms [5, 20, 26]. Studies of Wikipedia [21], Slashdot [16] and eBay [25], for example, all depict flourishing communities that experience some underprovision, yet don't allow it overwhelm the community.

However, discussions among redditors (the name Reddit community members give to themselves) alerted us to possible, significant underprovision on their site. Reddit is primarily a link-sharing site, but also has sections devoted purely to discussion, such as the popular *AskReddit* subreddit (i.e., a sub-community). Redditors expressed concern that their links go completely unnoticed by the community. Meanwhile, someone else submitting the same link later would go to the front page. As Reddit runs on a reputation currency called "karma," redditors complained that this was unfair. For example, one redditor writes:

> yes. this is what's frustrating. an innocuous, earnest submission with absolutely no activity whatsoever is just . . . vexing. it feels like . . . if you delete it and do it again later, it might work . . . I guess I'm just grateful when it works. it's such an incredible resource when the comments are flowing, but if your post gets buried for whatever reason, it's painfully anti-climactic.

[1] `http://www.alexa.com/siteinfo/reddit.com`
At the time of this writing, Alexa ranks Reddit as the 121st most visited site worldwide and the 54th most visited in the United States.

CSCW'13, February 23–27, 2013, San Antonio, Texas, USA.

Comments like this one inspired the statistical work we present here. In this paper, we present a study of underprovision on Reddit. Examining both page view data and duplicate submissions, we arrive at the conclusion that widespread underprovision of votes is likely happening on the site. Notably, Reddit overlooked 52% of the most popular links the first time they were submitted. This suggests that many potentially popular links (i.e., ones the Reddit community would value) are ignored, jeopardizing Reddit's core purpose.

We conclude by discussing possible reasons behind the underprovision of votes we observe on Reddit. In particular, we discuss candidate answers to the following question: Which design elements, if any, invite the underprovision we see? While we can offer no conclusive answers with this short paper, we hope to invite lines of work that can investigate this deeper research question.

METHOD

In this paper, we present two types of statistical evidence addressing underprovision on Reddit: page view data and an analysis of duplicate submissions.

Page Views

To gain precision around issues of underprovision, we first performed an analysis of page view data we collected from Reddit. We wanted to understand how many people actually look at the new links flowing into the site. As noted earlier, Reddit organizes its design around the most popular content; you have to seek out the queue of brand new submissions. The ratio of new submission page views to popular content page views would give us a first-order view of where attention flows on Reddit.

While sites like Alexa [2] can provide aggregate traffic for a site, fine-grained page-view data among subsections of a site is hard to come by unless you have access to site logs. However, we employed a workaround in focusing on a particular subcommunity within Reddit: the *pics* subreddit. Reddit's second-largest subcommunity by membership [24], *pics* lets users share pictures from around the internet. *pics* has more than 2.1M subscribers, accounting for roughly 3.6% of Reddit's total subscriptions, and its contributions very often make their way to site's most popular page. While the image content varies widely, redditors share the majority of their images using the image-sharing site `imgur.com` (a statistic we derived from our data). While Reddit does not provide page view data, `imgur.com` shows page views to users.

In this paper, we compare the page views that new *pics* images receive with how many the most popular *pics* images receive. While imperfect—it only measures a segment of Reddit—this result will give us insight into the site's distribution of attention. Between April 25, 2011 and May 11, 2011, we crawled the most popular *pics* images and the *pics* new queue every 10 seconds, recording every `imgur.com` link. Our dataset consists of page view statistics for the 14,864 images submitted to *pics* and the 648 most popular images during these 17 days. These particular 17 days were not a significant aspect of the study; rather, they allowed us to study Reddit over time while collecting enough data to draw conclusions.

Other referring sites. One potential confound with the approach just described is that people can visit an imgur picture by referral from any site on the web, or simply by typing its URL directly into their browser. Without access to imgur's server logs (and their associated HTTP referrer data), it is impossible to say how much of an image's traffic originates from Reddit. However, it is important to consider that a Redditor created imgur expressly for the Reddit community to share images, calling it "My Gift to Reddit" [18]. While imgur has branched out to other online communities since then, in a recent interview given to *The Wall Street Journal*, imgur's founder stated that Reddit referrals still dominate imgur's traffic [9]. Therefore, while we cannot achieve the kind of precision we would normally like, we see it as reasonable to examine page views of imgur images linked from Reddit for a first-order estimate of where attention flows on the site. We argue that despite noise at the individual image level (i.e., it would be hard to answer "Did this image get popular because of Reddit?"), this noise washes out between groups at the scale of tens of thousands of data points.

Duplicate Submissions

If enough people monitor the Reddit new queue, then truly good content should rarely go overlooked. (This is a variant of *Linus's Law*: "Given enough eyeballs, all bugs are shallow" [22].) In the second phase of our work, we crawled Reddit every ten seconds between April 25 and May 11, acquiring over a gigabyte of text. This time, however, we crawled the Reddit main page and its overall new queue. Our dataset consists of the 172,030 links submitted to Reddit and the 9,370 links that appeared on the most popular page during these 17 days. We will search this dataset for links which ultimately became popular, but were submitted earlier by someone else. These earlier links—good by definition because they ultimately became very popular—went overlooked by the community.

Ground truth data. This method allows us to construct ground truth data: links that ultimately became popular (even if they were overlooked the first few times) are precisely what Reddit values. However, our approach disregards links that could have been popular if only they had attracted Reddit's attention. Future work might consider overcoming this limitation by sending new Reddit links en masse to Mechanical Turk, for example. For the time being, we argue that the approach presented here will provide the research community with a conservative estimate of the proportion of valuable, yet overlooked content.

RESULTS

The median page views for an image put on the *pics* new queue (but one which did not end up becoming popular) is 557. The median page views for a popular image, on the other hand, is 148,911. Put another way, an image which ends up on the most popular page receives greater than two orders of magnitude more views than one that does not, Wilcoxon W = 9,008,436, $p < 10^{-15}$. As is typical for quantitative social data, these p-values hold less meaning than the magnitude of the differences between groups. Figure 1 illustrates this finding graphically, showing the log-scale distributions of both groups induced via Gaussian kernel density estimates.

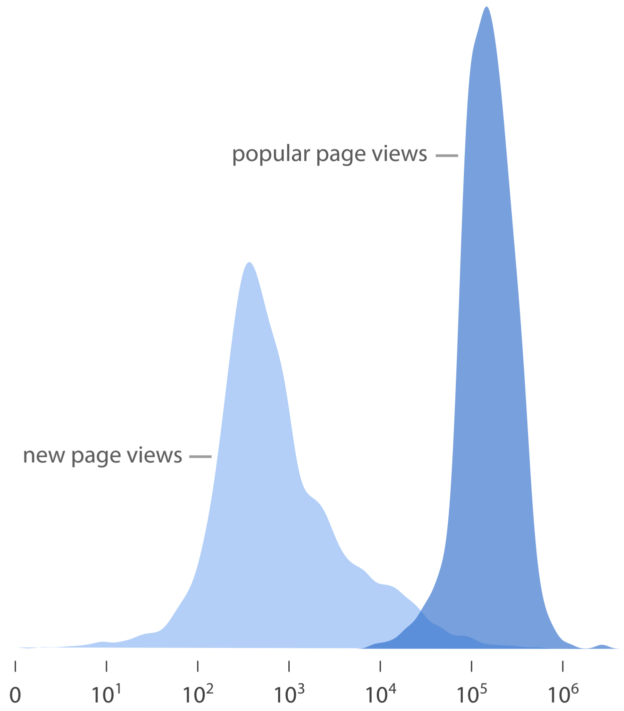

Figure 1. Page view distributions for new *pics* images versus most popular images. On average, the most popular images receive two orders of magnitude more page views than images on the new queue. (The distributions were induced via Gaussian kernel density estimates.)

Gaussian kernel density estimates smooth datasets in an attempt to expose a more continuous, more realistic picture of the underlying distribution [6].

Duplicate Submissions

Of the 9,370 most popular links we collected, we were able to identify when 5,186 first appeared in the new queue. (Since it takes time for a link to become popular, our process of crawling the new queue requires time to find ultimately popular links.) Of these 5,186 most popular links, someone else had submitted the same link earlier in 2,672 cases. In other words, 51.52% of the links destined for the most popular page had been submitted by someone else earlier (within our 17-day window). Those previous links went overlooked by nearly the entire Reddit community.

Figure 2 presents the distribution of times a most popular link had been submitted earlier by someone else. For instance, in 1,717 cases (33.1%) exactly one person submitted the same link earlier. A single link in our dataset had 54 prior submissions before finally finding popularity, probably the result of a bot slipping through Reddit's automated defenses. Figure 2, while only showing cases zero through five, seems to exhibit the long-tail form (e.g, log-normal, power-law, stretched exponential, etc.) characteristic of social data. This formulation of the data is too sparse, however, to confirm which exact shape it takes.

DISCUSSION

By combining these statistical approaches, we see a picture of underprovision emerge. Our *Duplicate Submissions* data present perhaps the most direct evidence. If you submit a great link to Reddit, more than half the time someone else will get the karma associated with the upvotes.

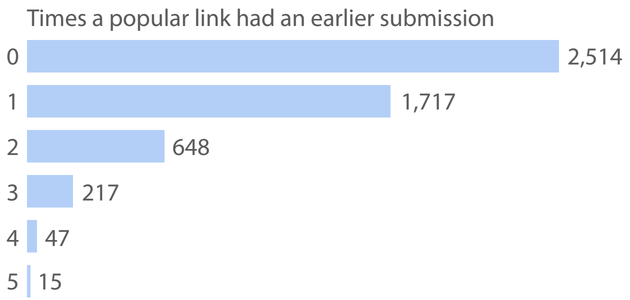

Figure 2. The distribution of how many times 5,186 most popular links had been submitted earlier by someone else. For example, in 648 cases two other people had submitted the same link earlier. Less than half of Reddit's most popular links (2,514/5,186 = 48.48%) get noticed on their first submission to the site.

We find it surprising that such a stark difference exists between attention paid to the new queue and attention paid to the most popular content. This remains true when you sum over both groups: the total page views given to the most popular images is more than four times that given to new images, despite the fact that this dataset contains 22.9 times more new images than most popular images.

Redditors often lament the presence of bots in the community, something we shouldn't ignore. Bots upvote links they favor and downvote ones they do not, en masse. Of course, any large online community will have to contend with them. Yet, one interpretation of the bot problem comes from the work presented here: Reddit may be susceptible because too few people look at new content, like a market with too few competitors. We know from previous work, for example, that a small number of colluding cliques can drive what emerges on Digg's front page, another social voting site [17].

We examined links which ultimately ended up popular to construct ground truth. We needed to conclusively call the link "good." However, our findings imply that many links that could reach popularity never get noticed, beyond the duplicate submissions examined here. Perhaps community members want multiple submissions to infer the worth of a link. In other words, redditors may monitor every link, yet purposely withhold their vote until they see a link for the second or third time. While we see this as a less viable explanation, we should investigate the impact this practice could have on those that get ignored in the early stages of a link's submission.

Design Implications

These early data suggest that widespread underprovision of votes is happening on Reddit. We next turn to possible explanations. What features of Reddit's design could generate widespread underprovision of votes? Is it specific to Reddit or something found more generally among social navigation systems? Each possibility below suggests a different design intervention, yet we caution that this space needs deeper empirical work before targeting specific design strategies. We instead use our data, along with theory and design criticism, to introduce starting points for new design-oriented research questions about social voting sites.

Could Reddit's design disincent searching the new queue? It seems axiomatic for a social voting site: most of the content submitted to it isn't any good. Redditors have a strong incentive to drop in on the site for its most popular content and ignore the firehose of new submissions. Most of the time, they will find something funny, interesting or inspiring. They would find exactly the opposite on the new queue: mostly bad jokes, uninteresting blog posts and uninspiring stories. Imagine the wasted effort a redditor would have to expend wading through all the noise flowing into Reddit.

In our dataset, 9,370 of the 172,030 links submitted to Reddit became popular (5.4%). Even if the true rate of good content is twice what we see currently on the site (i.e., roughly 11%), the enterprising redditor would face a deluge of bad content. Perhaps Reddit's fundamental design idea—showing by default the good content that others promote from the firehose of the internet—leads to underprovision. Perhaps the firehose is simply too much for too many people.

Perhaps because the voting mechanism isn't social? In previous work, researchers have found many large, successful online communities that overcome underprovision problems. However, earlier research has focused on designs that include conversations among community members: discussions on Slashdot or reviews on eBay, for instance. The Reddit voting mechanism isn't conversational. You vote with a click and go away. Perhaps this difference explains the discord with existing online communities research.

Could the subcommunity architecture reward poachers? Subreddits let users find smaller communities more in tune with their interests. The subreddits vary in size and topic, ranging from hundreds of thousands of subscribers to just a handful [24]. While this design follows best practices [14], as it shields redditors from the parts of the internet they don't care about, it also makes poaching possible. For example, imagine that a redditor posts a great link to a subreddit with 15 subscribers. Someone else in bigger subreddit can then crosspost it. The link submitted to the bigger subreddit probably has a greater chance of success.

Does someone else sell it better? When you post a link to Reddit, you also give it title. The site suggests that the title reflect where the link points, but it also gives the redditor room to comment. Titles are often witty, timely and sometimes reflect Reddit's idiosyncratic values. Perhaps the second or third submitter of the same link comes up with a better title, piquing everyone's interest in a way the original submitter did not.

Do redditors vote on the underlying link or on how well someone sells the title? In our data, while most resubmitted links have different titles, we find that most contain a majority of words from the original submission. If re-titling significantly explains our findings, then the mechanism is probably subtle. This question deserves investigation, as its answer could underpin how much voting mechanisms hinge on subtle, small design choices like letting user provide titles. Other sites, such as Hacker News, often moderate submissions whose titles do not mirror the target's title tag.

Or, do some redditors post at the wrong time? It seems likely that during certain times of day or under certain conditions (e.g., during periods of peak usage), the likelihood of having your link noticed could go down. In our data, we find effects for time-of-day, with popular links posted in the morning hours much more likely to get ignored than those in the afternoon and evening. At the same time, many fewer links are posted during morning hours, so it remains unclear how much simple external forces like time-of-day account for popular submissions going unnoticed on Reddit.

Theoretical Implications

Above, we explore design-oriented research questions that may help shed light on our findings. We believe the present research also suggests new directions for CMC theory. Researchers have consistently found vastly skewed participation curves in social systems (e.g., power-law, stretched exponential, log-normal, etc.), both offline and online [1, 15]. Sometimes this is framed as a result of a process of preferential attachment happening on underlying social networks [3]. Consider that voting participation patterns skew this way, too. Do different skews (i.e. exponents, in the power-law framing) lead to different proportions of overlooked, yet valuable content on social voting sites?

On Reddit, we see roughly half of all valuable content overlooked on its first submission. Researchers might leverage this work to contrast this finding with other social voting sites, like Digg and Hacker News. How does this "overlooked proportion" vary with broader participation skews, presuming different ones on different sites? Moreover, researchers can take our results as a point of departure for community-wide studies. Does such a high-rate of overlooked, valuable content result in high turnover among newcomers? How do people gauge when and what they should submit in the face of such high rates of failure?

CONCLUSION

Large online communities often overcome the threat of underprovision via a combination of social norms, repeated interaction and reputation mechanisms. However, in this paper we document a case of widespread underprovision within a major online community.

We hope that researchers will see directions for new research in this work, perhaps developing interventions to solve underlying design problems. For example, if the social voting design truly disincents searching the new queue, a strategy of investable reputation (i.e., something akin to buying stock in a great, but unknown company) might resolve the problem. We look forward to future work addressing these issues.

ACKNOWLEDGEMENTS

We would like to thank the Social Spaces group at Illinois, as well as the comp.social lab at Georgia Tech. Tony Bergstrom and Tanushree Mitra contributed valuable feedback on early drafts. Google helped support this work.

REFERENCES

1. L. Adamic and B. Huberman. Zipf's law and the Internet. *Glottometrics*, 3(143–150), 2002.
2. Alexa. http://alexa.com. Retrieved June 1, 2012.
3. A. Barabási and R. Albert. Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512, 1999.
4. Y. Benkler. *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. Yale University Press, 2006.
5. F. Berkes, D. Feeny, B. McCay, and J. Acheson. The Benefits of the Commons. *Nature*, 340(6229):91–93, 1989.
6. R. Cao, A. Cuevas, and W. Gonzalez Manteiga. A Comparative Study of Several Smoothing Methods in Density Estimation. *Computational Statistics & Data Analysis*, 17(2):153–176, 1994.
7. A. Dieberger, P. Dourish, K. Höök, P. Resnick, and A. Wexelblat. Social navigation: techniques for building more usable systems. *Interactions*, 7(6):36–45, 2000.
8. P. Dourish and M. Chalmers. Running Out of Space: Models of Information Navigation. In *Proc. CHI*, 1994.
9. L. Gannes. Interview: Imgur's Path to a Billion Image Views Per Day. http://dthin.gs/JgKPqL. Retrieved August 27, 2012.
10. E. Gilbert and K. Karahalios. Understanding Deja Reviewers. In *Proc. CSCW*, pages 225–228, 2010.
11. Hacker News. http://news.ycombinator.com. Retrieved June 1, 2012.
12. G. Hardin. The Tragedy of the Commons. *Science*, 162(3859):1243–1248, 1969.
13. T. Khopkar, X. Li, and P. Resnick. Self-selection, Slipping, Salvaging, Slacking, and Stoning: The Impacts of Negative Feedback at eBay. In *Proc EC*, pages 223–231, 2005.
14. A. Kim. *Community building on the web*. Peachpit Press, 2000.
15. J. Laherrere and D. Sornette. Stretched exponential distributions in nature and economy: "fat tails" with characteristic scales. *The European Physical Journal B-Condensed Matter and Complex Systems*, 2(4):525–539, 1998.
16. C. Lampe and P. Resnick. Slash(dot) and Burn: Distributed Moderation in a Large Online Conversation Space. In *Proc. CHI*, pages 543–550, 2004.
17. K. Lerman and A. Galstyan. Analysis of Social Voting Patterns on Digg. In *Proc. WOSN*, pages 7–12, 2008.
18. My Gift to Reddit. http://redd.it/7zlyd. August 27, 2012.
19. M. Olson. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Harvard University Press, 1974.
20. E. Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, 1999.
21. R. Priedhorsky, J. Chen, S. Lam, K. Panciera, L. Terveen, and J. Riedl. Creating, destroying, and restoring value in Wikipedia. In *Proc. GROUP*, pages 259–268, 2007.
22. E. Raymond. *The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary*. O'Reilly & Associates, Inc., 2001.
23. Reddit. http://www.reddit.com/help/faq#Whatisreddit. Retrieved June 1, 2012.
24. Redditlist. http://redditlist.com. Retrieved June 1, 2012.
25. P. Resnick and R. Zeckhauser. Trust Among Strangers in Internet Transactions: Empirical Analysis of eBay's Reputation System. *Advances in Applied Microeconomics*, 11:127–157, 2002.
26. M. Smith and P. Kollock. *Communities in Cyberspace*. Psychology Press, 1999.