JIS

Crowdsourcing, attention and productivity

Bernardo A. Huberman

Social Computing Lab, HP Laboratories, Palo Alto, CA, USA

Daniel M. Romero

Center for Applied Mathematics, Cornell University, Ithaca, NY, USA

Fang Wu

Social Computing Lab, HP Laboratories, Palo Alto, CA, USA

Abstract.

We show through an analysis of a massive data set from YouTube that the productivity exhibited in crowdsourcing exhibits a strong positive dependence on attention, measured by the number of downloads. Conversely, a lack of attention leads to a decrease in the number of videos uploaded and the consequent drop in productivity, which in many cases asymptotes to no uploads whatsoever. Moreover, short-term contributors compare their performance to the average contributor's performance while long-term contributors compare it to their own media.

Keywords: crowdsourcing; social attention

1. Introduction

We are witnessing an inversion of the traditional way by which content has been generated and consumed over the centuries. From photography to news and encyclopaedic knowledge, the centuries-old pattern has been one in which relatively few people and organizations produce content and most people consume it. With the advent of the web and the ease with which one can migrate content to it, that pattern has reversed, leading to a situation whereby millions create content in the form of blogs, news, videos, music, and so on, and relatively few can attend to it all. This phenomenon, which goes under the name of *crowdsourcing*, is exemplified by websites such as Digg, Flickr, YouTube, and Wikipedia, where content creation without the traditional quality filters manages to produce sought out movies, news and even knowledge that rival the best encyclopaedias. That such content is valued is confirmed by the fact that access to these sites accounts for a sizable percentage of internet traffic. For example, as of June 2007 YouTube, which in many ways can be considered a media company that outsources the production of its content to millions of users, comprised approximately 10% of all traffic on the internet [1].

Correspondence to: Bernardo A. Huberman, 1501 Page Mill Road, Palo Alto, CA 94304, USA. Email: bernardo.huberman@hp.com

What makes crowdsourcing both interesting and puzzling is the underlying dilemma facing every contributor, which is best exemplified by the well-known tragedy of the commons [2]. In such dilemmas, a group of people attempts to provide a common good in the absence of a central authority. In the case of crowdsourcing, the common good is in the form or videos, music, or encyclopaedic knowledge that can be freely accessed by anyone. Furthermore, the good has 'jointness' of supply, which means that its consumption by others does not affect the amounts that other users can use. And since it is nearly impossible to exclude non-contributors from using the common good, it is rational for individuals not to upload content and free ride on the production of others. The dilemma ensues when every individual can reason this way and exploit the efforts of others, making everyone worse off [3–7].

And yet paradoxically, there is ample evidence that while the ratio of contributions to downloads is indeed small, the growth in content provision persists at levels that are hard to understand if analysed from a public goods point of view. One possible explanation for this puzzling behaviour, which we explore in this paper, is that those contributing to the digital commons perceive it as a private good, in which payment for their efforts is in the form of the attention that their content gathers in the form of media quotes, downloads or news clicked on.

As has been shown, attention is such a valued resource that people are often willing to forsake financial gain to obtain it [8]. In the world of academia, for example, attention is the main currency, for we publish to get the attention of others, we cite so that other researcher's work gets attention, and we cherish the prominence of great work because of the attention it gathers [9]. Similarly, within online communities, status and recognition have been shown to be very important motivators for contributing [10]. Another important instance is open source software development. Although several studies have shown that open source projects are characterized by a very small core of contributors [11] where the free-riding problem is not acute, the idea of prestige seeking in the open source community has been explored. Raymond describes it as a 'gift culture' in which participants compete for prestige by giving time, energy and creativity away [12].

2. Results

This paper explores the conjecture that attention is an important driver of contributions to the digital commons. If this is indeed the case, one should be able to observe a correlation between the rate at which content is generated and the number of downloads. And if in addition a causal relation between the two does exist, we expect that those contributors that have a high level of downloads will continue to contribute, whereas those who see a decline in the attention that their content is receiving will decrease their productivity.

In order to investigate this conjecture we collected data from YouTube, a popular website that allows its users to upload, view and share video clips. After a YouTube user uploads a video, a 'view count' number is immediately displayed next to the video's title, which measures how many times it has been watched. Our dataset contained 9,896,816 videos submitted by 579,471 users by 30 April 2008. For each video upload we obtained its datestamp, the uploader's id and the final view count. Since older videos will naturally have more views, we needed to detrend the final view count data. Previous research has shown that the view count of YouTube videos does not saturate with time [13] and Figure 1 shows that this trend is linear with very high correlation. Therefore we performed a linear regression of $v \approx a\tau + b$ where v is each video's final count and τ is the time when each video was uploaded (see Table 1). The result is $a = -28.80$ and $b = 404{,}650$. For the rest of the paper we present the results obtained using the detrended values of v, i.e. $v_d = v - (a\tau + b)$. However, the tests were also conducted using the actual values of v and we found that all the results hold in both cases.

To study the dynamic interplay between productivity and attention, we partitioned time into two-week periods, starting when a user's first video was uploaded and ending when they uploaded their last one. A common pattern we observed is that most periods between a contributor's first and

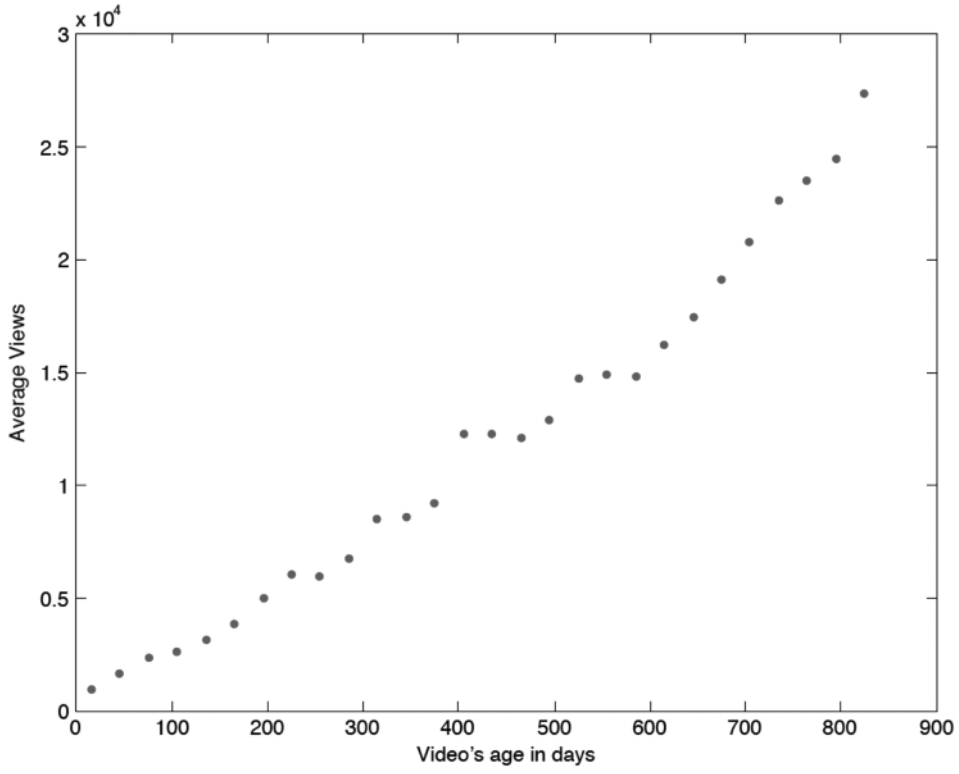

Fig. 1. Average number of views vs age of videos in days. The number of views linearly increases as the age of the video increases with very high correlation: 0.98.

Table 1
Definition of variables

Variable	Definition
v	Views count of a video
τ	Time at which a video was uploaded
v_d	Detrended v
t	Active period
n_t	Videos uploaded during period t
v_t	Average view count of videos uploaded during period t
T	Total active periods
v^-	10,000 (for first test)
v^-	median $\{v_t\}_{t=1}^{T-1}$ (for second test)

last uploads contain no uploads at all (on average, 66% of these periods are empty), indicating an intermittent productivity. Because of the sporadic nature of our data, we considered only the 'active' periods for each contributor (i.e. periods containing at least one upload), and labelled them as $t = 1, 2,....$

We measured the productivity of each contributor by the number of videos n_t uploaded during the tth active period, and the attention received by the average number of views v_t of the n_t videos (see Table 1). In other words, we wanted to establish how v_t affects n_{t+1}, n_{t+2},..., which provides dynamical information on how each contributor responds to different amounts of attention.

2.1. Linear regression of attention vs productivity

We first conducted a robust linear regression $\{n_{t+1}\}_{t=1}^{T} \sim \alpha\{log_{10} v_t\}_{t=1}^{T} + \beta$ for each contributor who was active for $T > 10$ periods [14]. (Because the view counts varied over many orders of magnitudes, it made sense to consider $\log_{10} v_t$ instead of v_t.) We thus collected 76,462 α values and conducted a t-test of the null hypothesis H1: 'The α values come from a normal distribution with non-positive mean'. The resulting p-value is less than 0.001, suggesting that H1 can be rejected. We also conducted the same test with different choices of T, and observed that as long as $T > 10$ the p-value was always less than 0.001. Hence, for those contributors who were active for a minimum number of periods, the more views they received in one period, the more videos they uploaded during the following period.

2.2. Productivity after different attention levels

A more direct approach to test our conjecture is to measure the change in each contributor's productivity at different attention levels. For each contributor who was active for at least two different periods, define $\bar{v}$ = median $\{v_t\}_{t=1}^{T-1}$ as the median received attention, where T is the number of active periods. According to this definition, all periods can be divided into two groups of equal size, $\lfloor (T-1)/2 \rfloor$: the 'good periods' in which she receives higher than usual attention ($G = \{S: v_s < \bar{v}\}$), and the 'bad periods' in which she receives lower than usual attention ($B = \{S: v_s < \bar{v}\}$).

Let

$$n^G = \frac{1}{\lfloor (T-1)/2 \rfloor} \sum_{s \in G} n_{s+1}$$

denote the average productivity following a good period, and similarly define

$$n^B = \frac{1}{\lfloor (T-1)/2 \rfloor} \sum_{s \in B} n_{s+1}$$

as the average productivity following a bad period. With these definitions the difference $\Delta = n^G - n^B$ measures the change of a contributor's productivity between different attention levels. If $\Delta > 0$ contributors upload more videos after obtaining more views, and if $\Delta < 0$ the opposite is true.

One can also test whether a contributor's productivity increases as they outperform the average contributor in the general population. To do so, we measured the average view count of all videos in our dataset, which is given by $\bar{v} = 10{,}000$ and used it to measure the productivity difference between good periods (more than 10,000 views on average) and bad periods (less than 10,000 views on average) through the quantity $\Delta = n^G - n^B$. For this test we only consider contributors that had at least one good period and one bad period.

We divided the contributors into several different groups depending on their number of active periods, and for each group tested the null hypothesis H2: 'The Δ values come from a normal distribution with non-positive mean'. Table 2 shows the results from these tests when $\bar{v} = 10{,}000$, including the number of contributors considered in each subgroup, the mean of the Δ values, and the p-values of H2. Notice that the p-values are very small for most groups, which supports our hypothesis that a competitive factor enters into the productivity of contributors. Also note that the mean of Δ decreases as the number of active weeks increases, indicating that those people who made relatively few contributions care more about their relative performance against others.

Figure 2 shows the histogram of the different 20,061 Δ values for the group of contributors who were active for two to nine periods when $\bar{v} = 10{,}000$. A t-test of the null hypothesis that $\Delta \leq 0$ yields a p-value less than 0.001, leading to rejection of the null hypothesis. Thus on average each contributor uploads more videos after a good period than a bad period. This indicates that each contributor tends to become more productive after receiving a number of views that exceeds the average contributor's performance.

For comparison purposes we also tested the same null hypothesis for $\bar{v}$ = median $\{v_t\}_{t=1}^{T-1}$ (i.e. the median view count of each contributor) which is not constant but varies from contributor to

Table 2
Test of the null hypothesis $\Delta \leq 0$, where $\Delta = n^G - n^B$ measures the productivity difference between a contributor's good periods (in which contributions received more than 10,000 views on average) and bad periods (less than 10,000 views on average). As the number of active weeks increases, the mean of Δ decreases

Active weeks	Contributors	Δ-Mean	*p*-Value
2–9	20,061	0.65	<0.001
10–19	24,517	0.53	<0.001
20–29	7789	0.38	<0.001
30–39	2153	0.20	0.18
40–70	515	0.09	0.50

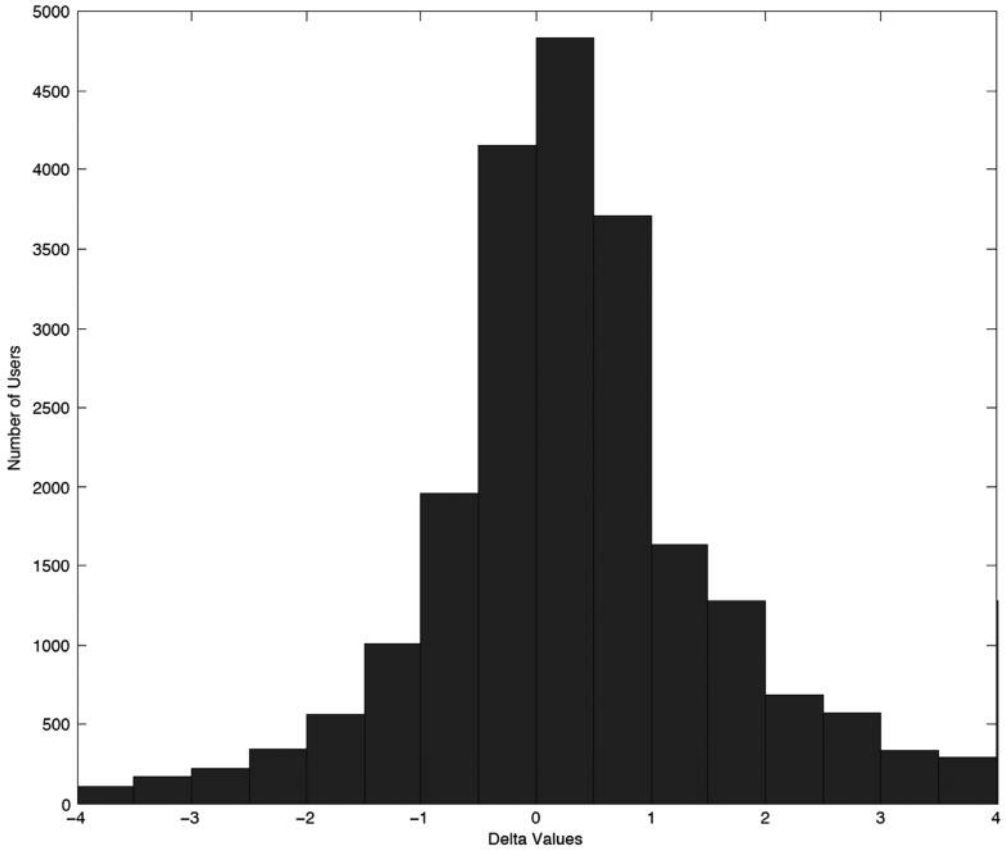

Fig. 2. Histogram of contributor's Δ values for contributors who were active from two to nine weeks. Notice that the maximum of the histogram is shifted to the right of the origin. The null hypothesis that data come from a normal distribution with non-positive mean can be rejected with *p*-value less than 0.001.

contributor. The results are listed in Table 3. We see that in this case the mean of Δ increases as the number of active weeks increases, indicating that the productive ones care more about how they have improved their own performance, rather than comparing with the rest of the community.

2.3. Power law of contributors

Tables 2 and 3 suggest that the number of active periods follows a power law distribution. We would like to determine if this is in fact true. Since the number of active periods and the number of contributions are highly correlated, it suffices to determine if the number of contributions follows a

Table 3
Tests of the null hypothesis $\Delta \leq 0$, where $\Delta = n^G - n^B$ measures the productivity difference between a contributor's good periods (in which contributions received more than the median view count) and bad periods (less than the median view count). As the number of active weeks increases the mean of Δ increases

Active weeks	Contributors	Δ-Mean	*p*-Value
2–9	85,949	0.05	0.15
10–19	68,317	0.20	<0.001
20–29	14,757	0.23	<0.001
30–39	3303	0.30	<0.001
40–70	673	0.43	<0.01

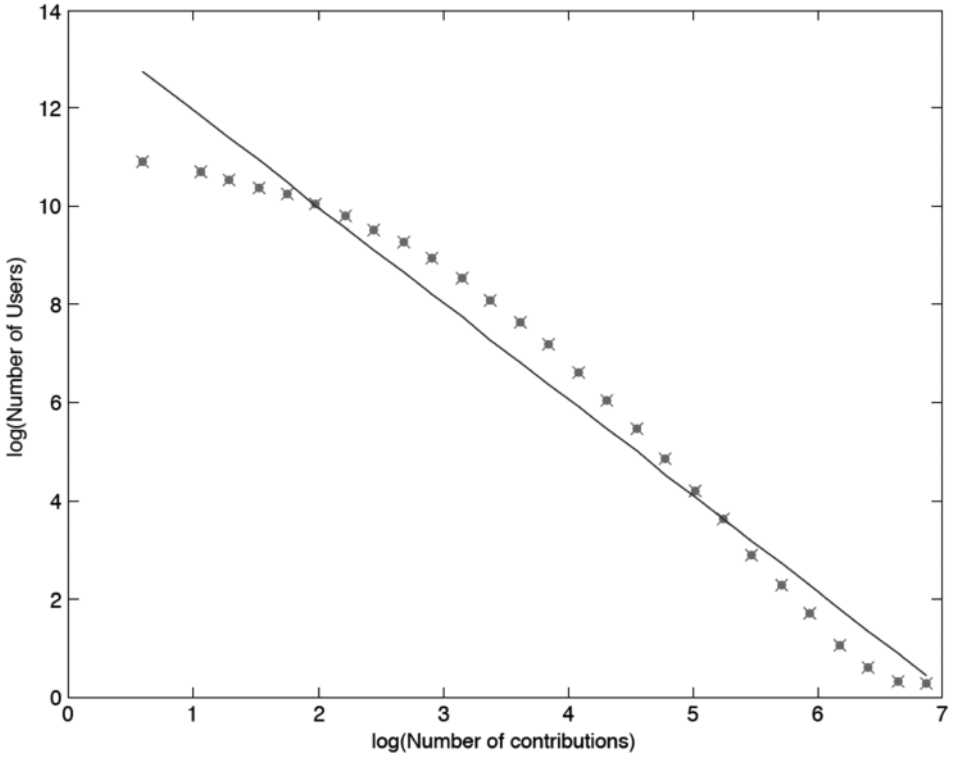

Fig. 3. log(Number of users) vs log(Number of contributions). This graph suggests a power law distribution with exponent of 2.46.

power law distribution. This has been shown to be the case in other peer production systems such as Bugzilla, Essembly, Wikipedia, and Digg with power law exponents of 1.98, 2.02, 2.28, and 2.4 respectively [15]. The same study also showed that the power law exponent is proportional to the effort required to contribute in the given system. We found that the power law also holds for YouTube with an exponent of 2.46 (see Figure 3), which indicates that the effort required to contribute in YouTube is high compared to other systems.

2.4. Evidence of causality

While the observed correlations between attention and productivity suggest a trend, they do not imply a causal relation between them. In fact, it is not clear whether an increase in attention causes productivity as a whole to grow or vice versa. In order to clarify this issue we used a Granger causality test, which is a statistical tool that determines causality in terms of prediction accuracy [16].

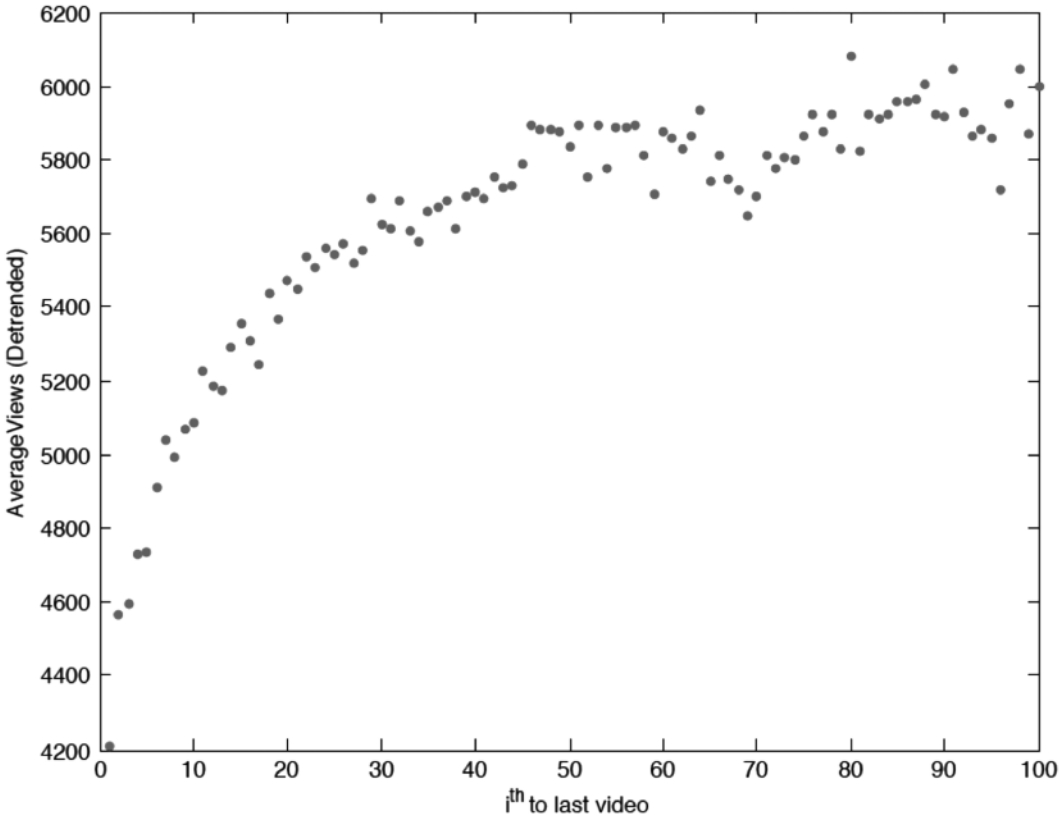

Fig. 4. Average number of views vs *i*th to last video. The origin represents the last video. The average number of views decreases as contributors approach their last video. Only videos with at most 10,000 views were used in this figure.

Given two signals X_1 and X_2, we say that X_1 G-causes X_2 if past values of X_1 contain information that helps predict future values of X_2. It is important to note that Granger causality is meaningful if only found in one direction, i.e. X_1 G-causes X_2 but X_2 does not G-cause X_1. If on the other hand Granger causality is found in both directions it is likely that X_1 and X_2 are only correlated and that the correlation is caused by a third signal.

In order to determine the causal relation between attention and productivity, we defined $\bar{v}_t$ to be the average of the all contributor's views during their *t*th active period, and similarly we let $\bar{n}_t$ be the average of all contributor's videos uploads during their *t*th active week. We then conducted a Granger causality test of the hypothesis that $\bar{v}_t$ G-causes $\bar{n}_t$, which resulted in a *p*-value of 0.01, and of the hypothesis that $\bar{n}_t$ G-causes $\bar{v}_t$, which gave a *p*-value of 0.61. This result shows that attention plays a determinant role in the productivity of those uploading videos.

2.5. Why do users stop uploading?

Finally, since it is a common observation that many contributors stop uploading videos, we decided to test if this behaviour was due to the small number of downloads their videos receive. To do so we considered all the contributors in our dataset that had not uploaded any videos during the previous four months to the date the data were collected.

Figure 4 shows the number of average views as a function of the *i*th to last video. As can be seen, as contributors approach their last video upload at the origin, the average number of previous views of their videos exhibit a marked decrease. This confirms our conjecture that decreasing attention leads to a lack of productivity, in this case to the point of making contributors stop uploading any videos. Furthermore, the asymptotical effect in Figure 4 shows that contributors who ultimately

stop uploading videos start out receiving a constant amount of attention, which eventually drops dramatically right before they quit. In other words, these contributors manage to maintain the amount of attention they receive constant for a while but never manage to increase it; on the contrary, it decreases to the point that the contributors cease trying.

3. Conclusions

By analysing a massive data set from YouTube we have shown that the productivity exhibited in crowdsourcing exhibits a strong positive dependence on attention. Conversely, a lack of attention leads to a decrease in the number of videos uploaded and the consequent drop in productivity, which in many cases asymptotes to no uploads whatsoever. Moreover, we were able to determine that uploaders compare themselves to others when having low productivity and to themselves when exceeding a personal threshold. More generally, these results show that the tragedy of the digital commons is partly overcome by making the uploading of digital content a private good paid for by attention.

References

[1] N. Anderson, The YouTube effect: HTTP traffic now eclipses P2P, *Argentina Mulls Open Source Move* (2007). Available at: http://arstechnica.com/news.ars/post/20070619-the-youtube-effect-http-traffic-now-eclipses-p2p.html (accessed 4 August 2009).
[2] G. Hardin, The tragedy of the Commons, *Science* 162 (1968) 1243–1248.
[3] E. Adar and B.A. Huberman, Free riding on Gnutella, *First Monday* 5(10) (2000).
[4] A. Asvanund, K. Clay, R. Krishnan and D. Smith, An empirical analysis of network externalities in peer-to-peer music-sharing networks, *Information Systems Research* 15(2) (2004) 155–174.
[5] D. Hughes, G. Coulson and J Walkerdine, Free riding on Gnutella revisited: The bell tolls, *IEEE Distributed Systems Online* 6(6) (2005).
[6] N. Glance and B.A. Huberman, Dynamics of social dilemmas, *Scientific American* 270(3) (1994) 76–81.
[7] S.S. Levine and S. Shah, *Cultivating the digital commons: A framework for collective open innovation*, paper presented at the annual meeting of the American Sociological Association 2004. Available at: http://www.allacademic.com/meta/p108966_index.html (accessed 4 August 2009).
[8] B.A. Huberman, C. Loch and A. Onculer, Status as a valued resource, *Social Psychology Quarterly* 67(1) (2004) 103–114.
[9] G. Franck, Science communication, a vanity fair, *Science* 286 (1999) 53–55.
[10] J. Lampel and A. Bhalla, The role of status seeking in online communities: Giving the gift of experience, *Journal of Computer-Mediated Communication* 12(2) (2007) 435–455.
[11] A. Mockus, R.T. Fielding and J.D. Herbsleb, Two case studies of open source software development: Apache and Mozilla, *ACM Transactions on Software Engineering and Methodology* 11(3) (2002) 309–346.
[12] E.S. Raymond, Homesteading the Noosphere, *The Cathedral & the Bazaar* (O'Reilly, Sebastopol, CA, 1999).
[13] G. Szabo and B.A. Huberman, Predicting the popularity of online content, to appear in *Communications of the ACM* 2009. Available at: http://arxiv.org/abs/0811.0405 (accessed 4 August 2009).
[14] J.O. Street, R.J. Carroll and D. Ruppert, A note on computing robust regression estimates via iteratively reweighted least squared, *The American Statistician* 42(1) (1988) 152–154.
[15] D. Wilkinson, Strong regularities in online peer production, *Proceedings of the 9th ACM conference on Electronic Commerce*, Chicago, IL, 2008 (ACM, New York, 2008) 302–309
[16] C.W.J. Granger, Investigating causal relations by econometric models and cross-spectral methods, *Econometrica* 37 (1969) 424–438.