Crowdsourcing, Attention and Productivity

Bernardo Huberman
Social Computing Lab, HP Laboratories
Palo Alto, CA 94304
bernardo.huberman@hp.com

Daniel M. Romero
Center for Applied Mathematics
Cornell University
Ithaca, NY 14853
dmr239@cornell.edu

Fang Wu
Social Computing Lab, HP Laboratories
Palo Alto, CA 94304
fang.wu@hp.com

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, uploaders compare themselves to others when having low productivity and to themselves when exceeding a threshold.

Categories and Subject Descriptors

H.1 [**Information Systems**]; J.4 [**Social and Behavioral Sciences**]

Keywords

Crowdsoucing, 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 encyclopedic knowledge, the centuries-old pattern has been one in which a 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, etc. and relatively few can attend to it all. This phenomenon, which goes under the name of *crowdsourcing*, is exemplified by websites such as `Digg`, `Flicker`, `YouTube`, and `Wikipedia`, where content creation without the traditional quality filters manages to produce sought out movies, news and even knowledge that rivals the best encyclopedias. 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 [2].

What makes crowdsourcing both interesting and puzzling is the under- lying dilemma facing every contributor, which is best exemplified by the well-known *tragedy of the commons*. 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 encyclopedic knowledge that can be freely accessed by anyone. Further-more, 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 free ride on the efforts of others, making everyone worse off [1, 3, 7, 5, 10].

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 analyzed from a public goods point of view. One possible explanation for this puzzling behavior, 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 downloads or news clicked on. As it has been shown, attention is such a valued resource that people are often willing to forsake financial gain to obtain it [6]. In the world of academia, for example, attention is often its main currency, for we publish to get the attention of others, we cite so that other researcher's work get attention, and we cherish the prominence of great work if only because of the attention it gathers [4]. Similarly, within online communities, status and recognition have been shown to be very important motivators for contributing [9].[1]

2. RESULTS

If attention is indeed the main driver of contributions to the digital commons, one should be able to observe a correlation between the rate at which content is generated and the

[1]Another important instance is open source software development. Several studies have shown however, that open source projects are characterized by a very small core of contributors [11] where the free-riding problem is not acute.

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 title, which measures how many times it has been watched. Our dataset contained 9,896,816 videos submitted by 579,471 users by April 30, 2008. For each video upload we obtained its datestamp, the uploader's id, and the final view count. Since older videos have been on the website for a longer time, they naturally will have more views. Thus we need to detrend the final view count data. We performed a linear regression of $v \sim at + b$ where v is each video's final count and t is the time when each video was uploaded. 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. $v_d = v - (at + b)$. 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 2-week periods, starting when they upload their first video and ending when they upload their last one. A common pattern we observed is that most periods between a contributor's first and last uploads contain no uploads at all (on average, 66% of these periods are empty), indicating an intermittent productivity. Because of the bursty nature of our data, we considered only the "active" periods for each contributor (i.e. periods containing at least one upload), and labeled them as $t = 1, 2, \ldots$.

We measured the productivity of each contributor by the number of videos n_t she uploads during the t'th active period, and the attention she receives by the average number of views v_t of the n_t videos. In other words we wanted to establish how v_t affects $n_{t+1}, n_{t+2}, \ldots$, which provides dynamical information on how each contributor responds to different amounts of attention.

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 that was active for $T > 10$ periods [12]. (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 that the α values come from a normal distribution with non-positive mean. The resulting p-value is less than 0.001, suggesting that the null hypothesis 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.

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 $\bar{v} = \text{median}\{v_t\}_{t=1}^{T-1}$ as her median received attention, where T is her 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.

Figure 1 shows the histogram of the different 20,061 Δ values for the group of contributors who were active for 2 to 9 periods. 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 becomes more productive after a good period than a bad period.

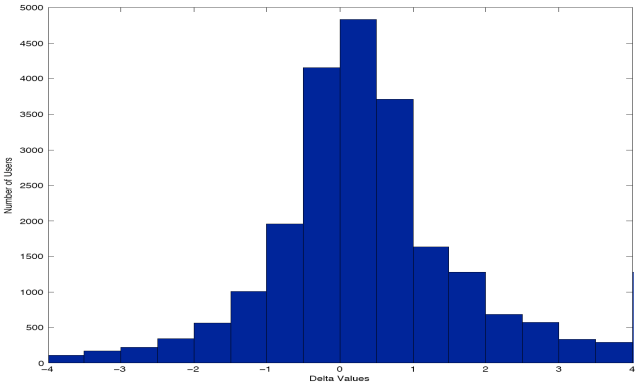

Figure 1: Histogram of contributor's Δ values for contributors that were active from 2 to 9 weeks. Notice that the maximum of the histogram is shifted to the right of the origin. The null hypothesis that data comes form a normal distribution with non-positive mean, can be rejected with p-value less than 0.001.

Figure 1 indicates that each contributor tends to become more productive after receiving a number of views that exceeds her own normal performance. One can also test whether his productivity increases as she outperforms 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 $\bar{v} = 10000$, and used it to measure the productivity difference between good periods (more than 10000 views on average) and bad periods (less than 10000 views on average) through the quantity $\Delta = n^G - n^B$. We divided the contributors into several different groups depending on their number of active periods, and tested the null hypothesis "$\Delta \leq 0$" for each group. Table 1 shows the results from these tests, including the number of contributors con-

sidered in each subgroup, the mean of the Δ values, and the p-values of the null hypothesis. 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 in Table 1 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 other contributors.

For comparison purposes we also tested the same null hypothesis for $\bar{v} = \text{median}\{v_t\}_1^{T-1}$ (i.e. the median view count of each contributor) which is not constant but varies from contributor to contributor. The results are listed in Table 2. 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.

Active weeks	Contributors	Δ-mean	p-value
2-9	20061	.65	$< .001$
10-19	24517	.53	$< .001$
20-29	7789	.38	$< .001$
30-39	2153	.20	.18
40-70	515	.09	.50

Table 1: 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 her contributions received more than 10000 views on average) and bad periods (less than 10000 views on average). As the number of active weeks increases, the mean of Δ decreases.

Active weeks	Contributors	$\Delta-$mean	p-value
2-9	85949	.05	.15
10-19	68317	.20	$< .001$
20-29	14757	.23	$< .001$
30-39	3303	.30	$< .001$
40-70	673	.43	$< .01$

Table 2: 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 her contributions received more than her median view count) and bad periods (less than her median view count). As the number of active weeks increases the mean of Δ increases.

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 [8]. 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 only 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.

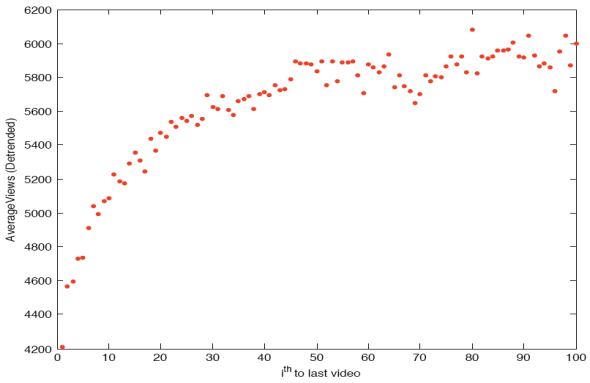

Figure 2: Average number of views vs. i'th to last video. The origin represents the last video. The average number of views decreases as contributers approach their last video. Only videos with at most 10000 views were used in this figure.

Finally, since it is a common observation that many contributors stop uploading videos, we decided to test if this behavior 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 four months previous to the date the data was collected.

Figure 2 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 exhibited 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.

3. CONCLUSIONS

By analyzing 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.

4. REFERENCES

[1] E. Adar and B.A. Huberman. Free riding on `Gnutella`. *First Monday*, 5(10), 2000.

[2] N. Anderson. The `YouTube` effect: HTTP traffic now eclipses P2P. *ars technica.*, `http://arstechnica.com/news.ars/post/20070619-the-youtube-effect-http-traffic-now-eclipses-p2p.html`, 2008.

[3] A. Asvanund, K. Clay, R. Krishnan and M. D. Smith. An empirical analysis of network externalities in peer-to-peer music-sharing networks. *Information Systems Research*, 15:2, pp. 155–174, 2004.

[4] G. Franck. Science communication, a vanity fair. *Science*, 286, pp. 53–55, 1999.

[5] N. Glance and B. A. Huberman. Dynamics of social dilemmas. *Scientific American*, pp. 76–8, March 1994.

[6] B. A. Huberman, C. Loch and A. Onculer. Status as a valued resource. *Social Psychology Quarterly*, vol. 67, no. 1, 103–114, 2004.

[7] D. Hughes, G Coulson, and J Walkerdine. Free riding on `Gnutella` revisited: The bell tolls? *IEEE Distributed Systems Online*, Vol. 6, No. 6, 2005.

[8] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37, 424–438, 1969.

[9] 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), article 5, 2007.

[10] 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.

[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), pp. 309–346, 2002.

[12] J. O. Street, R. J. Carroll and D. Ruppert. A note on computing robust regression estimates via iteratively reweighted least squares. *The American Statistician*, 42, 1, pp. 152–154, 1988.