Analysing User Activity in Online Collaboration Projects

Ronan McHugh and Birger Larsen

mchugh.r@gmail.com, blar@iva.dk
Royal School of Library and Information Science
Birketinget 6, 2300 Copenhagen S (Denmark)

Abstract
In this paper, we conduct a web usage analysis of user activity in two online collaborative projects, Open Street Map (OSM) and The Pirate Bay (TPB). User logs are downloaded and analysed to provide a picture of user productivity and user activity over time. We find that users of Open Street Map are both more productive and are active for longer than their counterparts in The Pirate Bay. We discuss the methods and heuristics used as an example of how to carry out webometric web usage analysis of online collaboration sites.

- Introduction

The web offers a wealth of valuable content to most people, but by using the web, people also invariably leave behind a wealth of traces, and the web has also become a rich source of information *about* the web users. As outlined by Björneborn & Ingwersen the research field of webometrics exploits such data to carry out studies of "(1) Web page content analysis; (2) Web link structure analysis; (3) Web usage analysis (including log files of users' searching and browsing behaviour); (4) Web technology analysis (including search engine performance)" (2004, p. 1217). The analysis reported in this paper belongs to the web usage analysis category as it carries out quantitative analysis of user activity changes in two collaborative online projects, *The Pirate Bay* and *Open Street Map*. The overall purpose of the research project is to analyse how online collaboration web sites *persuade* user to remain active and keep contributing to these sites. The project is framed in the theoretical perspective of Persuasive Design (Fogg, 2009), and uses two complementary methods: 1) heuristic analysis to evaluate the interface design of the sites, and 2) web usage analysis based on data downloaded and mined from the sites to assess user activity patterns over time. This paper is an extension of the preliminary results presented in McHugh & Larsen (2010a; 2010b). In the present paper we report initial results from the web usage analysis, and analyse user activity in terms of both total activity and rate of activity over time, and discuss the methods used as an example of how to carry out webometric web usage analysis of online collaboration sites.
The paper is structured as follows: The next sections give background information on online collaboration and summarises the two case sites. This followed by a discussion of the methods applied in the study. Next an analysis of the results is presented, followed by a discussion and reflections on future work.

Online Collaboration

In this paper, online collaboration refers to the phenomenon of voluntary participation in cooperative projects co-ordinated over the Internet. The phenomenon is embodied in projects such as *Wikipedia*[65], a collaboratively written encyclopaedia and *Peer-to-Peer University*[66], an online education resource. In general, such projects have several common characteristics, namely, that contributions are made on a voluntary basis, coordination of work tends to be organic and without formal hierarchy, and that projects are typically based around specific values.

[65] http://www.wikipedia.org
[66] http://p2pu.org

Online collaboration projects are a fascinating field of study, representing a unique opportunity for researchers. The digital nature of participation means that an unprecedented level of data are available for analysis, as user actions are recorded by the system in question and are often freely accessible. This data thus allows us to apply quantitative analysis to the study of human creativity and innovation.

- **Summary of Cases**

The analysis is based on complete user histories downloaded from two online collaborative projects; Open Street Map and The Pirate Bay. These sites were chosen because of their collaborative nature, but also because they stored detailed user histories in an easily accessible format. *Open Street Map*[67] is a collaboratively produced map of the world. Participants contribute by adding points to the map which they may have derived from exploring an area with a GPS transmitter or simply from local knowledge. *The Pirate Bay*[68] is a site which indexes torrent files which are used to download files collaboratively, from multiple computers at a time. While sites like Open Street Map or Wikipedia receive plaudits and public approbation for the public good, The Pirate Bay has been target for international controversy, criticism and litigation. Participants contribute by uploading torrent files and allowing other users to download files from their computer. A vast number of these files are in breach of copyright laws, leading to outrage from media providers and repeated, so far unsuccessful, attempts to have the site shut down (Dagens Nyheter, 2009; The Pirate Bay, 2010).

- **Methodology**

The methodology for this study is a quantitative analysis of user histories, available online. The data are processed and saved in table format in order to facilitate analysis. Two types of analysis are conducted; analysis of user productivity and analysis of user activity rates over time.

- *Data Retrieval*

The data for this study was retrieved by downloading histories of user activities stored publicly on the websites in question. URLs for user profiles were obtained by entering the unique sub-directories for user profiles into *Yahoo! SiteExplorer*[69] and downloading the first 1,000 results, the maximum number which one can download in Yahoo! Site Explorer. Duplicates were removed and a script based on Python's *Beautiful Soup*[70] module was used to download the full histories associated with each user, converting pages from a HTML format into a tabbed text file. Since user histories were typically spread over several pages, it was necessary to design the script to download all pages of user history and not just the most recent one.

- *Bin division of participants*

In order to facilitate analysis, it was decided to divide each sample of users into three bins based on total activity levels. After analysis of the Lorenz distribution of participation rates for both projects, it was decided to divide the samples based on the formula of 60, 30, and 10. The first 60% of participants are the lowest level contributors, the next 30% are medium level contributors and the final 10% are the highest level contributors. This method was chosen

[67] http://www.openstreetmap.org
[68] http://thepiratebay.org
[69] http://siteexplorer.search.yahoo.com
[70] http://www.crummy.com/software/BeautifulSoup/

because of the high rate of participation inequality observed within the samples, whereby a small number of participants are responsible for a large percentage of contributions while a majority of participants only ever contribute a relatively small amount. This "power-law" distribution, well-known from bibliometrics, is also common to such environments, to the extent that commentators accept a certain degree of inevitability to it (Nielsen, 2006; Shirky, 2008).

- *Data Analysis*

In order to analyse user productivity, we constructed frequency distributions in which users were sorted according to the number of times they had participated to their respective project. A hierarchy was created showing the proportion of users from each sample who had contributed once, twice, three times etc. This hierarchy thus shows the proportion of users who only contribute at these small scales. We construct a similar hierarchy based on user lifetime, thus showing the proportion of users who are only active for a brief period of time.

In order to analyse user participation rates over time, a series of spreadsheet formulae were used to number all user participation events according to when in the user's lifetime they occurred. Thus, all user activity could be charted on a timeline starting with their first ever contribution to the project. Using these timelines, a series of frequency distributions were derived which plotted the percentage of total contributions for each user group that occurred within a specific time-frame (e.g. two weeks, three months, etc). This methodology allowed us to make broad observations about the average lifetime participation rates of particular groups of contributors and compare these with other users of the same project or with the respective user division of the other project.

- **Analysis**

The TPB dataset consisted of 268,141 torrents produced by 1,495 users[71]. The set had an average contribution of 179.36 torrents per user with a median of 10. The OSM dataset consisted of 1,884,104 edits contributed by 762 users. This gives an average of 2472.58 edits per user, with a median of 299.

Due to problems with the data retrieved we have only analysed low and mid-level contributors. This is due to obvious flaws in the data retrieved for the highest level contributors to OSM, whereby these contributors had improbably low lifetimes, for example, some users with many thousands of uploads had lifetimes of only eleven days. This suggests that the data retrieved was only a partial representation of their total lifetime activity and as such lifetime based analysis of their contributions was thought not to be representative.

Table 1. Proportion of users contributing between 1 and 5 times

No. of contributions	*Open Street Map*	*The Pirate Bay*
1	1.31% (10)	17.24% (258)
2	1.31% (10)	8.35% (125)
3	0.13% (1)	6.41% (96)
4	0.52% (4)	5.08% (76)
5	0.26% (2)	3.00% (45)
≤5	3.54% (27)	40.10% (600)

[71] It was possible to download more than the maximum of 1,000 user pages for The Pirate Bay because TPB stores its user pages in two separate sub-directories.

- *Productivity Analysis*

Table 1 compares the proportion of users from each sample contributing at different scales. What is clear from this is the large difference in participation patterns between OSM users and TPB users. An extremely large proportion of TPB users only ever contribute once to the project, while only a small proportion of OSM users do the same. This indicates that the Pirate Bay is not very good at encouraging repeat contributions from users. Of course, we must remember that contributing a torrent will frequently involve more work than adding a point to a map. For this reason it is worth looking at lifespans of users as another measure of user behaviour.

Table 2: User drop-out between 1-5 days

Lifetime (days)	*Open Street Map*	*The Pirate Bay*
1	9.05% (68)	21.67% (324)
2	1.86% (14)	2.47% (37)
3	0.26% (2)	0.86% (13)
4	0.26% (2)	0.66% (10)
5	0.26% (2)	0.93% (14)
≤5	11.71% (88)	26.62% (398)

The average lifetime of TPB users is 308.35 days and the median is 169 days compared to 514.88 days and 516 days for OSM users. Table 2 shows the proportion of users from each project whose lifetimes last 1-5 days. As is apparent, OSM editors tend to remain involved with the project longer than TPB users. This indicates that the OSM project is better at persuading users to remain active than TPB is.

- **Lifetime Based Analysis**

- *Comparison of low level users*

Time-based analysis of contribution rates of low-level users across systems show a considerable amount of difference between the two projects. Low-level TPB users contribute proportionally far more in the first days of their lifespans than corresponding OSM users (see Figure 1 below). This difference is particularly apparent in the first two weeks of lifetime and the first day especially, where TPB users contribute 32.04% of their total uploads, while OSM users contribute only 1.62% of their total edits.

It is only after about 14 days that OSM contribution rates start to be consistently higher than TPB rates, with OSM editors contributing 1.28% of total lifetime edits while TPB users contribute 0.68%. This difference becomes more pronounced as time goes on: in the period between 330 days and 360 days after first activity, OSM editors contributed 3.69% of total lifetime edits while TPB users contributed 0.88% (McHugh & Larsen, 2010b). This comparison points to a different dynamic of participation which can also be seen in the different lifespans of users; the median lifetime of low-level TPB users is 19 days, while the median lifetime of low-level OSM users is 432 days. 36% of low-level TPB users contribute for only one day, while only 13% of low-level OSM users do the same. These statistics suggest that OSM is far better at persuading users to maintain their involvement in the project. The fact that the median lifespan of low-level OSM editors is well over a year suggests a far more sustainable level of involvement among OSM editors.

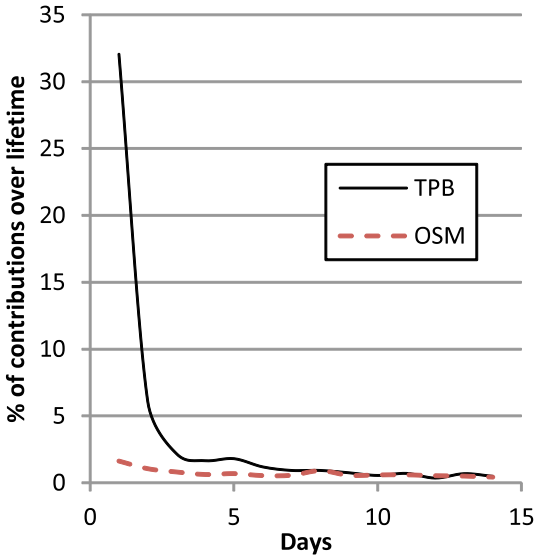

Figure 12. Relative contributions by low-level users over first two weeks of lifetime.

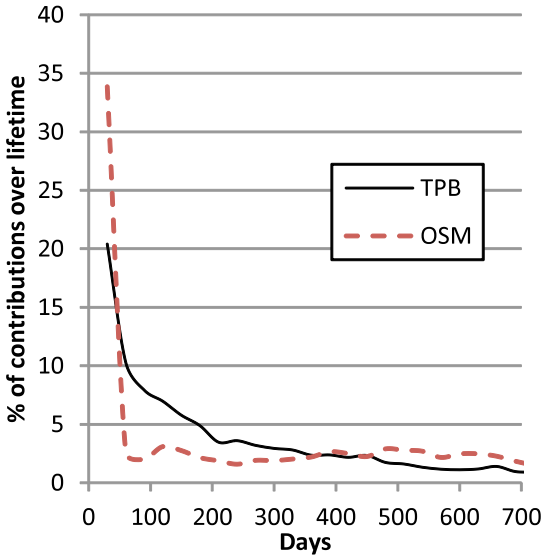

Figure 13. Relative contributions by mid-level users over first two years of lifetime.

- *Comparison of mid-level users across systems*

The lifespan analyses of mid-level users reveals some surprising results. As with the analysis of low-level contributors, mid-level TPB users start their activity periods by contributing more than their OSM counterparts, although the difference is not so great, 3.86% of total contributions in their first day vs 1.39% of OSM mid-level contributions. What is surprising is an extremely large rise in OSM contributions relative to those of TPB users after the sixth day. As can be seen in Figure 2, this increase in contributions is reflected in the two year timeline where the OSM contributions are more concentrated in the early days of lifespan than those of TPB users. This huge concentration of productivity in the second week of OSM user activity is followed by consistently lower productivity over the following months of activity, until 390 days where the OSM users again begin to outperform their TPB counterparts. The average lifespan of mid-level TPB users is 476.22 days, while the median is 406.5, OSM mid-level users on the other hand have an average lifespan of 784.25 days and a median of 785 days. This indicates that despite the flurry of activity in the first week, mid-level OSM users are both longer-lasting and more consistent than their TPB counterparts.

- **Discussion and conclusion**

The results of the quantitative analysis show several clear differences between the projects in terms of user behaviour. Users of OSM tend to contribute more consistently and for longer periods of time than their counterparts in TPB. On the other hand, TPB users tend to make the majority of their contributions at a very early stage of their activity and to remain active for shorter periods of time. This difference is likely related to the unique features of the projects in question and suggests that user participation rates are not inevitable in online collaborative endeavours. The different nature of the two sites may also have an impact here. As much of the activity on TPB is related to sharing of files that breach copyright laws the TBP users may tend to be more discrete and to contribute less and for a shorter time in order not to risk legal issues.

The present study serves as an example of how data from online collaboration sites can successfully be mined using simple techniques. The usage data of online collaboration sites supports a wide range of issues to be studied and can provide value information about such sites and their users. The present study thus adds another set of tools for webometrics analysis

of web 2.0, and follows other recent papers in this line of research, e.g. Angus, Stuart and Thelwall (2009) and Cheong & Lee (2010).

Future studies

Using this dataset, we can also analyse other aspects of contributor behaviour on such sites. For example, the user histories analysed in The Pirate Bay also provide details of how many contributors are seeding any given torrent, i.e. how many users are allowing other users to download the file from them. We can take this seeding activity as a mark of quality, as users would most likely not seed files of poor quality. Thus, using total seeders of a torrent, we can emulate the citation analyses that are typical for bibliometrics. As an example, we can construct a "p-index" for pirates, modelled on Hirsch's "h-index", whereby a pirate has index p if p of her/his Nt torrents have at least p seeders each, and the other (Nt - p) torrents have no more than p seeders each (Hirsch 2005).

Expanding the dataset, it would be of interest to use a script to download pages associated with a torrent for instance on a monthly basis over a year in order to analyse how seeding activity progresses over time and how comment activity plays a role in this. Comments could be analysed using sentiment analysis in order to test whether there is a correlation between good comments and high levels of seeders or bad comments and low levels of seeders.

Acknowledgments

We wish express our sincere thanks to dr. Kim Holmberg at University Åbo Akademi University, Finland for constructive comments on a draft of this paper. This work has been supported by the ACUMEN project (EU project grant No. 266632).

- References

Angus, E., Stuart, D., & Thelwall, M. (2009). Flickr: an academic image resource? In *Proceedings of the 12th International Conference of the International Society for Scientometrics and Informetrics: Vol. 2.* Brazil: BIREME/PAHO/WHO and Federal University of Rio de Janeiro, p. 904-905.

Cheong, M. & Lee, V. (2010). Twittering for earth: A study on the impact of microblogging activism on earth hour 2009 in Australia. In *Proceedings of ACIIDS 2010*, Part II, p. 114-123.

Björneborn, L. & Ingwersen, P. (2004). Towards a basic framework for webometrics. *Journal of the American Society for Information Science and Technology*, 55(14), 1216-1227.

Dagens Nyheter (2009): *The Pirate Bay sentenced to one year in prison*. Retrieved January 15, 2010 from http://www.dn.se/kultur-noje/musik/the-pirate-bay-sentenced-to-one-year-in-prison.

Fogg, B. J. (2009): A behaviour model for persuasive design. In *Persuasive '09: Proceedings of the 4th International Conference on Persuasive Technology*. New York: ACM, p. 1-7.

J.E. Hirsch. An index to quantify an individual's scientific research output. In *Proceedings of the National Academy of Sciences of the United States of America*, 102:16569-16572, 2005.

McHugh, R. & Larsen B. (2010a): Persuading Collaboration: Analysing Persuasion in Online Collaboration Projects. *International Journal on Social Media MMM: Monitoring, Measurement, and Mining*. 1(1), 102-105.

McHugh, R. & Larsen B. (2010b): Analysing User Lifetime in Voluntary Online Collaboration. In *Proceedings of the Tenth Danish Human-Computer Interaction Research Symposium (DHRS2010)*. Roskilde : Roskilde University p. 23-26. (Computer Science Research Report; 132).

Nielsen, J. (2006). *Participation Inequality: Encouraging more users to contribute*. Retrieved January 15, 2010 from http://www.useit.com/alertbox/participation_inequality.html .

Shirky, C. (2008). *Here Comes Everybody: How change happens when people come together*. London: Penguin Books.

The Pirate Bay (2010): *Legal threats against the Pirate Bay*. Retrieved January 15, 2010 from http://thepiratebay.org/legal.