

Persuading Collaboration:

Analysing Participation in Online Collaboration Projects

Name: Ronan McHugh

Date: September 2010

Supervisor: Birger Larsen

Pages: 70

(c) Ronan McHugh 2010

This work is licensed under the Creative Commons Attribution-ShareAlike 3.0 License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/3.0/> or send a letter to Creative Commons, 543 Howard Street, 5th Floor, San Francisco, California, 94105, USA, or see Appendix 4.

“To attribute, therefore, the industrial progress of our century to the war of each against all which it has proclaimed, is to reason like the man who, knowing not the causes of rain, attributes it to the victim he has immolated before his clay idol. For industrial progress, as for each other conquest over nature, mutual aid and close intercourse certainly are, as they have been, much more advantageous than mutual struggle... In its wide extension, even at the present time, I also see the best guarantee of a still loftier evolution of our race.”

Peter Kropotkin, *Mutual Aid: A Factor in Evolution*

“You never change things by fighting the existing reality. To change something, build a new model that makes the existing model obsolete.”

Buckminster Fuller

Abstract

The growth of worldwide internet usage has given rise to the phenomenon of “open movements”. These are communities that evolve based around the collaborative production of common resources, to which free access is typically provided to all who choose to use it. Such communities and the resources they create have developed rapidly over the past 10-20 years and are the cause for major scholarly interest.

This study takes the step of applying the lens of Persuasive Design to the study of open movements in order to show how site design can play a role in increasing participant activity and longevity. For this purpose it looks at two “open movement” resources, the collaboratively edited map OpenStreetMap and The Pirate Bay, a tracker for torrents uploaded by the file-sharing community.

The analysis is two-fold: first, it uses quantitative data of user participation in the systems, derived from downloads of system-generated user-histories to generate an overall picture of user-participation in each of the systems. Second, it applies a set of heuristics to evaluate the persuasive design of the systems in question. It then connects the results of the quantitative analysis with the heuristic analysis in order to see how the persuasive design of the systems impact on user participation.

This thesis will primarily be of value to researchers of online collaboration, although it may also be of interest to researchers in the field of Persuasive Design.

Acknowledgements

Writing this paper has, appropriately enough, been a collaborative endeavour. The original idea to work with the concept of Persuasive Design stems from a class taken with Jette Hyldegaard and Haakon Lund. Working with Birger Larsen made the study of necessity a quantitative affair. This approach was a challenge for me and I am grateful for Birger's help and patience as well as his wealth of experience in advising me. The technical aspects of data retrieval would not have been possible were it not for the help of Dr. Toine Bogers, who gave generously of his time to design retrieval scripts for the project. This spirit of generosity was also present in several online forums, whose users were kind enough to help me with certain technical problems of analysis. Where appropriate, these users are mentioned in footnotes to the text. Of course, all responsibility for errors is my own.

Theoretically, the impetus to work with the question of online collaboration stems from the work of Clay Shirky as well as Yochai Benkler whose analyses of online collaboration combine academic rigour with an appreciation of the moral force that lies behind such projects. José Ortega's analysis of Wikipedia has served as a "best case" example of what good quantitative analysis looks like.

This paper was written and researched using free software. \LaTeX was used for word processing and typesetting, OpenOffice Calc and R Commander were used for data analysis and graphs. My thanks to Grethe, Danielle and Dara for feedback and comments.

Contents

1	Introduction and Problem Statement	7
1.1	Introduction	7
1.2	Online collaboration and “open movements”	8
1.3	Persuasive Design	8
1.4	Summary of case: The Pirate Bay	9
1.4.1	Pirate and proud	9
1.4.2	Technical details of contribution	10
1.4.3	Other forms of contribution	10
1.5	Summary of case: Open Street Map	10
1.5.1	Geographers unite!	11
1.5.2	Technical details of contribution	11
1.5.3	Other forms of contribution	11
1.6	Can The Pirate Bay and OSM be compared?	11
1.7	Problem statement	12
2	Review of related literature	14
2.1	Defining “open movements”	14
2.2	Studies related to our cases	15
2.2.1	Previous studies of Open Street Map	15
2.2.2	Previous studies of The Pirate Bay	16
2.3	Related research about open movements	17
2.3.1	Quality in open movements	17
2.3.2	Explaining motivation in open movement projects	18
2.3.3	Social psychology in open movements	18
2.3.4	Participation patterns in open movement projects	20
2.3.5	Participation patterns and bibliometrics	21
2.4	Related studies about analytic evaluation	21
2.4.1	Analytic evaluation in Interaction Design	21
2.4.2	Analytic evaluation in Persuasive Design	22
2.4.3	Analytic evaluation of our cases	23
3	Methodology	24
3.1	Glossary of terms used	24
3.2	Selection of data sources	25
3.3	Data sources	25
3.4	Retrieval and processing of data	26
3.4.1	Yahoo! Site Explorer	29
3.4.2	Creation of productivity bins	29

3.5	Analysing Participation Patterns between case studies	30
3.5.1	Q1: How do user contribution patterns resemble and differ from each other? In particular, what is the difference between drop-out rate of new participants and lifetime contribution patterns in the two systems.	30
3.5.1.1	Drop-out Rate	31
3.5.1.2	Lifetime based analysis	31
3.5.2	Q2: Does participation in these projects follow standard rules (e.g. power laws) or is it different from case to case?	31
3.6	Participation patterns internal to projects	32
3.6.1	Statistical model of correlation	33
3.6.2	Role of feedback mechanisms in increasing participation	34
3.6.3	Quality analysis of torrents	34
3.7	Analytic evaluation of persuasive features	35
4	Analysis	37
4.1	Summary of data retrieved	37
4.1.1	Summary of The Pirate Bay data	37
4.1.1.1	Contribution rates	37
4.1.2	Summary of Open Street Map data	38
4.1.2.1	Contribution rates	38
4.2	Drop out rate	38
4.3	Contributions over lifetime	40
4.3.1	Lifetime analysis of Open Street Map editors	40
4.3.1.1	Contributions within the first two weeks of user lifetimes	41
4.3.1.2	OSM contribution rates over two years	42
4.3.2	Lifetime analysis of participation in The Pirate Bay	42
4.3.2.1	Participation patterns over the first two weeks of user lifetime	43
4.3.2.2	Participation patterns over two years	43
4.3.3	Comparison of lifetime participation rates across systems	45
4.3.3.1	Comparison of low-level users	45
4.3.3.2	Comparison of mid-level users across systems	46
4.4	Power Laws and Contribution Inequality	47
4.5	Heuristic Analysis	50
4.5.1	Analytic Evaluation of Open Street Map	51
4.5.1.1	Step One: Surface level use	51
4.5.1.2	Step Two: New participant	51
4.5.1.3	Step Three: Participation and Co-ordination	52
4.5.2	Analytic Evaluation of The Pirate Bay	52
4.5.2.1	Step One: Surface level use	52
4.5.2.2	Step Two: New participant	52
4.5.2.3	Step Three: Participation and Co-ordination	53
4.5.3	Summary of heuristic analysis	53
4.6	Miscellaneous Analyses	54
4.6.1	Measuring quality of contributions to The Pirate Bay	54
4.6.2	Effect of feedback on participation rates in The Pirate Bay	57

5 Results and Conclusions	60
5.1 Summary of results in relation to research questions	60
5.2 The value of persuasion	63
5.3 Relevance of this study to future research	64
Bibliography	65

List of Figures

3.1	HTML extract from an OSM user page	27
3.2	Extract of script used to process OSM user pages	28
3.3	Lorenz curve of TPB contributors plotted against an equality curve	30
3.4	Sample correlation line	33
4.1	Percentage of users remaining active over one year	40
4.2	Contribution rates over first two years of lifespan	42
4.3	Contribution rates over first two years of lifespan, excluding the first 30 days of lifespan	45
4.4	Contribution rates by low-level users over first two weeks of lifespan	46
4.5	Contribution rates by low-level users over first two years of lifespan	47
4.6	Contribution rates by mid-level users over first two weeks of lifespan	48
4.7	Contribution rates by mid-level users over first two years of lifespan	48
4.8	Lorenz curve of participation in The Pirate Bay	49
4.9	Lorenz curve of participation in Open Street Map	49
4.10	Quality Analysis of overall TPB sample	55
4.11	Quality Analysis of low level TPB users	55
4.12	Quality analysis of mid level TPB users	56
4.13	Quality analysis of TPB high-level users	56
4.14	Quality analysis of TPB sample overall with several outliers re- moved	57
4.15	Correlation between number of uploads and average number of comments	58
4.16	Correlation between user lifetime and average number of comments	59

List of Tables

4.1	Proportion of users contributing between 1 and 5 times	39
4.2	Proportion of users contributing between 1 and 50 times (inter- vals of 10)	39
4.3	Percentage of users dropping out of activity during the first week of lifetime	39
4.4	% of contributions by low-level and mid-level users over the first two weeks of their lifetimes	41
4.5	Percentage of contributions by all TPB user groups over the first two weeks of lifetime	43
4.6	% of contributions by low-level, mid-level and high-level TPB users over the first two years of their lifetimes	44
4.7	Comparison of participation drop-off in Open Street Map with Lotka's Law	48
4.8	Comparison of participation drop-off in The Pirate Bay with Lotka's Law	49
4.9	Comparison of top 5 contributors in Open Street Map with Zipf's Law	50
4.10	Comparison of top 5 contributors in The Pirate Bay with Zipf's Law	50

Chapter 1

Introduction and Problem Statement

1.1 Introduction

Voluntary collaboration in web-based projects is one of the biggest success stories of our time. Based on the rapid growth of internet access since the 1990s, more and more people have come online and started collaborating to create common resources. These resources have expanded to influence an ever-increasing amount of human social life, both online and off.

The output of such activity is highly varied, ranging from the production of complex software packages, to the collective writing of encyclopedia articles, to web based filtering and aggregation of news content. Despite such huge differences in output, there are some striking parallels in the organisation of such projects. Contributions tend to be made on a voluntary basis, with participants contributing for reasons other than monetary reward. Often there is a more or less explicitly defined value system that motivates participants. Participant community structure tends to be quite non-hierarchical and where formal authority figures exist, often in the form of site owners and project initiators, they tend to govern in a hands-off way.

The scale and productivity of such collaborative endeavours are fascinating topics of study in themselves, but the digital nature of such projects makes them open to empirical study in a way that offline cooperative projects are not. This presents a wonderful opportunity for researchers across the academic spectrum to study the dynamics of these collaborative projects.

Two of the key questions facing researchers into open movement projects have been very basic: how do these projects function and why do they function? This thesis aims to contribute towards answering these two questions by undertaking a quantitative analysis and a heuristic walkthrough of participation in two such systems, The Pirate Bay and Open Street Map, and analysing the results within the paradigm of Persuasive Design. Before formulating the problem statement and research questions however, I will give a brief overview of the cases being studied as well as the general problematic of "open movements" and the field of Persuasive Design.

1.2 Online collaboration and “open movements”

The remarkable growth of online projects powered by voluntary collaboration has caused a great deal of interest within academic circles, and has correspondingly been researched and theorized from many different perspectives on the basis of specific projects or on the phenomenon as a whole. Among these perspectives I can find those of lawyers, interested in the legal ramifications of such projects [5], business theorists, interested in the creative potential of these communities [51] and social psychologists researching the social dynamics of online volunteer communities [8]. Moreover, there are also well regarded studies from participants in such communities that attempt to provide theoretical or empirical answers for problems salient to their projects [42, 36]. Ortega’s study of Wikipedia is particularly noteworthy as an attempt by a community member to develop a strong methodological framework for analysing Wikipedia’s growth, success and quality over time.

1.3 Persuasive Design

At the same time as “open movements” have become an important topic of research for academics, the field of persuasive design has begun to establish itself at a crossroads between academia and the business world. Incorporating elements from interaction design, psychology and social psychology, it seeks to examine the ways in which technology influences our behaviour, persuading us to perform certain activities at the expense of others and creating emotional associations that encourage us to repeat target behaviour.

The prime theorist of this field is undoubtedly B.J. Fogg, the author of *Persuasive Technology* and a number of other influential works in the field. Fogg defines persuasion as “an attempt to change attitudes or behaviours or both” [15], seeing persuasive technologies as those technologies which embody such attempts. Fogg is particularly interested in the ways in which the ubiquity of computing technologies allows designers to have a positive influence on user behaviour, for example, by decreasing water usage or helping people to give up smoking, although he acknowledges that the same technologies can also be used towards ethically questionable ends.

It is worth noting that Fogg argues that computer mediated communication should not be an issue for Persuasive Design, since the focus is on computers as a communicative tool rather than as a source of persuasion. This position is not convincing; clearly computer based programs can also persuade people to communicate and interact, in which case it is reasonable to look at these through a Persuasive Design perspective. Moreover, the processes of communication and interaction can in themselves be persuasive; if computers programs facilitate these processes, they can be labelled persuasive actors. It is telling that Fogg himself has devoted considerable time to analysis of Facebook as an example of Persuasive Technology; since Facebook’s main function is to persuade users to interact and communicate with each other this would suggest that Fogg has revised this position whether explicitly or not.[16]

R. Khaled et al. use the persuasive framework to analyse “Social Software” (SSW), a label which includes online collaborative projects but can also be much broader. They identify three principal themes for investigation: *affiliation*,

access and *participation*. *Affiliation* refers to the processes of identification that result in people joining SSW communities and motivate them to participate. *Access* refers to users' ability to observe each others' contributions which thus enables social psychological processes of learning, comparison and influence. *Participation* refers to the contributions that users make and the subsequent social processes that these entail. [25]

1.4 Summary of case: The Pirate Bay

The Pirate Bay sits somewhat uncomfortably within the array of open and collaboratively produced internet resources. While projects such as Wikipedia or Ubuntu receive plaudits and public approbation for their contributions to the public good, The Pirate Bay has been a target for international controversy, criticism and litigation. Although The Pirate Bay also relies on user contributions and collaboration to create content, this content is torrent files which enable peer-to-peer (p2p) downloading of files. A vast number of these files are in breach of copyright laws, leading to outrage from media providers and repeated attempts to have the site shut down.[33, 40]

The Pirate Bay consists primarily of an index for torrent files, a type of file which makes it possible for users to download large files from other users collaboratively. Torrents are small files which contain meta-data about another file. This file has typically been broken up into several pieces by the uploader. Those downloading the file (peers) download it a piece at a time, they can download pieces from several different users and provide pieces to other peers without themselves possessing a full copy of the file. This means that download times for even large files can be very rapid if there are enough people providing a full copy for download (seeding) or themselves downloading a copy (leeching). This has the opposite effect from traditional single source downloads whereby popular files are typically slower to download and cause a large drain on the source's server.[11]

1.4.1 Pirate and proud

These features in themselves do not make The Pirate Bay particularly interesting as a research subject; there are many different websites that act as trackers and search engines for user contributed torrent files, such as Mininova¹ and Demonoid². What is interesting about The Pirate Bay is that it directly courts controversy and attempts to rally a sense of community through appeals to a putative "pirate" identity. This identity is based on multiple different strands including a relatively traditional support for civil liberties and privacy rights, more novel arguments against copyright as an institution and antagonism towards government and media company efforts to enforce and extend copyright law.[38, 39, 40] It is expected that such appeals serve to reinforce contributors' willingness to participate by providing a moral justification for piracy and an opposition to industry group attempts to stigmatise media piracy. These issues have become popularised in recent years and pro-piracy political parties have been formed in several countries across the world.[24]

¹<http://www.mininova.org/>

²<http://www.demonoid.com/>

The Pirate Bay and related projects are also typically quick to respond to efforts to block access to the site, providing detailed technical information in order to circumvent such efforts (which they typically describe as censorship) and have even launched their own proxy service which makes it possible for users to download files through another IP address, thus hindering identification.[12] Such efforts help maintain participation in the project as well as overall use of the site, as The Pirate Bay traffic has contributed to rise despite numerous blocks by internet providers around the world[4, 14].

1.4.2 Technical details of contribution

Contributions to The Pirate Bay can take several shapes. The most significant form of contribution is the uploading of torrent files. For a torrent file to be uploaded, a user must first possess a copy of the file to be transferred, and must then create a torrent file using software such as Vuze or UTorrent before uploading this to The Pirate Bay. Some users go to great lengths in acquiring highly restricted proprietary software and “cracking it”, i.e., removing built-in security restrictions intended to prevent piracy. Users who have downloaded the full file can contribute by seeding, this implies that they keep their torrent program (such as Transmission, Vuze or BitTorrent) running and allow other users to download from them. Users who are downloading the file can also upload to others, this typically happens by default but users can also modify their settings to limit their upload speeds[11][13].

1.4.3 Other forms of contribution

Users can also contribute to The Pirate Bay in a number of other ways. On the main site, users can post comments on torrent threads indicating torrent quality, providing help with technical aspects of the process, requesting people to seed the file, etc. There is also a large bulletin board attached to the site in which users post requests for files, create tutorials for torrent creation and downloading, post links to their own torrent uploads, report problems with the site, suggest improvements and engage in general conversation.

1.5 Summary of case: Open Street Map

Open Street Map (OSM) is a online project to create a freely usable map of the world based on collaborative editing. The project is inspired by the success of Wikipedia in creating a large and high quality online encyclopedia through collaborative editing of articles. It was further inspired by the recognition that most publically available maps have legal or technical restrictions on their use, thus making it difficult for people to re-use map data for their own ends.[34] The Open Source license that Open Street Map is released under has made possible a number of derivative projects as well as several commercial applications.[35]

Unlike The Pirate Bay, OSM fits more comfortably into the “open movements” category. It is driven by a desire to make content that is available to everyone without charge, it promotes open standards for data and attempts to increase the amount of data in the public realm through contact with governments and private bodies. It has approximately 250,000 users, of whom

approximately 10% contribute regularly, i.e., at least once a month [52, 44]. Although OSM is driven by volunteer contributors, it is also supported by a not-for-profit foundation, the Open Street Map Foundation, which manages the OSM servers, organises events and makes contact with potential data donors. [18]

1.5.1 Geographers unite!

One of the striking features about OSM is the high level of community that appears to underpin the collaborative process. There are many local groups around the world, composed of editors living in a specific area. These groups coordinate efforts between editors, share skills and organise “mapping parties” which involve intensive mapping of specific areas. This high level of organisation around the OSM project was demonstrated after the earthquake that struck Haiti in January 2010; within two days Open Street Map editors had built the most detailed and up-to-date map of Port-au-Prince available, including roads, hospitals and refugee camps, primarily using aerial photographs as source material.[43] It is this community focus which makes Open Street Map of particular interest for research.

1.5.2 Technical details of contribution

Contribution to Open Street Map can take several forms. Ideally most activity should take the form of contributing GPS (Global Positioning System) traces; this involves participants travelling around an area with a GPS unit, taking notes about the area being travelled through. Once this has been done, the GPS data is uploaded to OSM to allow other editors to see it and compare it with the finished maps. Using this data, editors can mark roads, points of interest, and other features to maps. It is also possible to edit maps through comparison with aerial photographs using the Potlatch application that is integrated with OSM. This feature makes it possible for editors to create basic maps for areas that they have never been to, although OSM documentation emphasises that local knowledge is of crucial importance to the editing process.[34]

1.5.3 Other forms of contribution

As noted, there is a high level of coordination between editors as part of the map-making process; editors co-ordinate their activities through a variety of mechanisms including online chat channels, Wiki articles, mailing lists and forums.

1.6 Can The Pirate Bay and OSM be compared?

An obvious question that can be asked is about the extent to which contributions to The Pirate Bay and Open StreetMap be compared; can activities as different as uploading torrent files and adding points to a map be compared productively? In general, it needs to be recognised that single contributions cannot be equated between systems, i.e. a user with ten uploads to The Pirate Bay is not the equivalent of a user who has made ten edits to OSM. Despite this, I believe

that contribution rates on a system level can be compared. For example, I can compare the average rate of contributions in a sample with the median of the sample, thus examining the degree to which contributions are skewed towards a few high activity contributors.

Within the overall framework of Persuasive Design, I can use the analysis to identify points of persuasive success or failure; for example, the proportion of users with only one edit is likely to indicate a point of prime persuasive importance for both systems. Moreover, it is apparent that by incorporating two systems within the analysis I can identify common persuasive challenges and show how the systems deal with them differently or similarly. In this context, the fact that two very different systems are being analysed is likely to constitute a strength rather than a weakness, as it will allow the results of the research to be more generalisable than those that would emerge from a study of only one form of system.

1.7 Problem statement

The problem for this thesis is to assess the persuasive strength of collaborative online systems through a quantitative analysis of participation in two such systems combined with a heuristic analysis of system features. The subsidiary research questions that follow from this problem formulation are as follows:

- Q1: How do user contribution patterns resemble and differ from each other in Open Street Map and The Pirate Bay? In particular, what is the difference between drop-out rate of new participants and lifetime contribution patterns in the two systems.
- Q2: Does participation in these projects follow standard rules or is it different from case to case?
- Q3: How does user contribution correlate with other features of user participation within these systems? Does feedback have an effect on contribution rate? How can quality be assessed in relation to contribution rate?
- Q4: How can persuasive features of the above systems explain differences in user participation?

Question 1 looks at user contributions on a broad systemic level. It focuses on drop-out rates and lifetime contribution patterns as indicators of particular persuasive importance. Drop-out rate refers to the likelihood of participants to stop activity after a short period of time or small number of contributions. In analysing lifetime contribution patterns I try to assess how users' participation in the systems changes over time.

In Question 2, the patterns observed in Question 1 will be compared to several distribution rules that are typically used to describe such samples as well as some bibliometric rules that may be of relevance. This comparison will test (a) whether such rules are generally applicable to the data in question and (b) serve as a benchmark for comparing participation patterns between the systems.

Question 3 looks at the relation between participation patterns and other features of the system such as feedback and contribution quality. It will divide users for each system into three separate groups, corresponding with their level of contribution and investigate usage patterns for each group. It will also propose a measure of quality that can be used for analysing user contributions to file-sharing sites.

Question 4 combines the results of the preceding questions with a heuristic walk-through of the persuasive features involved in the systems. In order to do so, it is first necessary to develop heuristics suitable for application to collaborative online projects. Using these heuristics, I try to connect the persuasive design of the systems with the participation outcomes analysed in Questions 1-3.

Chapter 2

Review of related literature

The emergence of various forms of web-based voluntary collaboration has sparked a great deal of interest in academic circles. Subsequent research has come from many different angles and thus has highlighted different themes within the general discussion. In this section I will look at some of these themes, focusing on those most relevant to a persuasive understanding of voluntary collaboration. Before doing this, I will present a definition and brief history of “open movements” and present research that deals specifically with our cases.

2.1 Defining “open movements”

While many examples of projects that form a part of the “open movements” phenomenon have been studied individually and comparatively for a number of years. To the best of my knowledge the term itself is a relatively novel one, seemingly coined in Ortega 2009[37]. Ortega provides a definition of open movements which relies on three criteria:

1. Open movements are built primarily on voluntary work, which is coordinated by the participants themselves.
2. These movements create “knowledge outcomes”, such as software, media files, documents etc, in a digital format.
3. These movements make all products of their work available for free, typically under some form of license which stipulates conditions under which their work may be re-used and shared.[36]

The academic interest in “open movements” dates back to the Free Software or Open Source movement. This movement is rooted in the “hacker culture” that developed among computer science researchers in universities in the United States in the late 1970s and early 1980s. These researchers were used to collaborating with each other on a non-commercial basis, sharing tools, passwords and skills freely. However, these practices were threatened when commercial interests began to dominate the research, leading to increasing copyright restrictions upon much of their work. In response to these restrictions, one researcher, Richard Stallman, developed a way to “hack” copyright, by creating a copyright license (the General Public License or GPL) that made it illegal for users of a

program to prevent others from using it freely, for example by creating price barriers through selling of it. This “hack” kicked off a huge wave of creativity, as it allowed programmers to develop software that they could freely use and share with others. Through the internet, it became possible for many developers to collaborate in producing software together. Examples of contemporary free software products are Ubuntu, an operating system, Apache, a widely used software for web servers, and Firefox, an internet browser [48, 42].

As access to the Internet has moved from universities and research labs into people’s homes, the open movements phenomenon has expanded from tools created by and for specialists into projects aimed towards a much wider audience [42]. The potentials for harnessing voluntary contributions through the Internet has been adopted by the mainstream in the guise of “crowd-sourcing”, a term which refers to a practice by websites of turning the leading role in content creation over to users, while they themselves provide a framework which organises this content creation in a meaningful way.

Several such undertakings have been extremely successful; examples include, *Wikipedia.org*¹, a collaboratively edited encyclopaedia, *YouTube.com*², a site for sharing and viewing of homemade videos, and *Flickr*³, a site for sharing and viewing of photographs. Of the top ten most visited websites worldwide according to Alexa rankings, five were sites based primarily on user-created content at the time of writing (<http://www.alexa.com/topsites>).

Of course, this should not imply that “crowd-sourcing” is synonymous with the open movements phenomenon, many of the sites that rely on crowd-sourcing for content creation embody values that are substantially different too and often at odds with the values of open movements. Haklay for example, notes that “crowd-sourcing” is potentially highly exploitative as it relies on unpaid users to do the work of private entities for free[21]. Moglen has also discussed the ethical problems inherent in websites that provide hosting for user created content, noting that “spying comes free” in exchange for a certain amount of free web-hosting and a user interface [28].

2.2 Studies related to our cases

There exists little previous academic investigation into our two cases, particularly OpenStreetMap. In this section I will summarise some of the work most relevant to this project. Themes from this work will be developed further in Section 2.3 where I will look at studies of other open movement projects.

2.2.1 Previous studies of Open Street Map

Goodchild looks briefly at Open Street Map as part of a general overview of what he calls “Volunteered Geographic Information” (VGI), including other services such Wikimapia and Flickr geo-caching, i.e., embedding locational data into photos taken. Goodchild sees VGI as part of a general trend towards a “bottom-up” science, in which citizens are able to play a greater role in the construction of sophisticated products, based on web-based coordination and the proliferation

¹www.wikipedia.org

²www.youtube.com

³www.flickr.com

of powerful tools as part of consumer electronics. Although VGI has great potential, he concludes that it also faces great problems, not only of incomplete coverage but also of unequal access to the digital resources and skills that are a prerequisite to participation in the project[20].

Haklay and Weber provide a more detailed overview of Open Street Map, examining the user interface for editors, the technical features that power the system, motivations for editors as well as participation patterns in the project. On motivation, they cite one “core member” who notes that:

People have a range of reasons for getting involved in the project—from certain ideological views such as a belief in the provision of free information to improve the world, to anti-national mapping agency views, to those who enjoy going out and mapping or sitting at home and writing computer code, to those who enjoy feeling like part of a community.[22]

On the problem of participation inequality they note while this inequality is common to such projects, OSM may not be helped by requiring a high level of system knowledge to edit and classify data appropriately[22].

Haklay has also investigated map quality within the city of London in OSM based on three different elements using the British Ordnance Survey data as a benchmark of quality. Although Haklay is generally upbeat about the prospects for OSM, he raises several important problems. Firstly, there exists broad differences of quality within the OSM data set: based on a survey of road data, he estimates that on average OSM data approximates 70% accuracy with occasional drops to 20%. Similarly large discrepancies in quality exist in other aspects and Haklay argues convincingly that it is due to differences in ability and effort on the part of the responsible editors. Another significant problem occurs in coverage, whereby areas of middle and high-income are in general far more extensively mapped than areas of low-income. Again, this reflects the volunteer nature of the project and specifically a likely bias caused by a low participation of mappers from low-income areas[21].

One might also criticise Haklay’s choice of London as a test-case for OSM map quality. London was one of the first areas in the world to be added to OSM and has a high proportion of OSM editors resident. These factors would suggest that London is likely to be one of the areas in the world with the best coverage by OSM and as such, quality studies of London are likely to indicate a best case scenario rather than an example that can be generalised throughout the system.

2.2.2 Previous studies of The Pirate Bay

While The Pirate Bay has received a significant amount of coverage in the media, in academia this interest has tended to translate into work reflecting the legal controversy around the site[7] or even political or philosophical questions that arises from this[2].

However, there has been some level of research into peer-to-peer file sharing in general within the field of computer science. This research focuses on download speed in peer-to-peer downloads. Noting the importance of maintaining high “sharing ratios” among peers in order to maintain high download speeds, Mol et al. point out the relative effectiveness of private communities (i.e.

communities where membership is needed to download and membership typically requires some form of invitation) in enforcing high sharing ratios among peers[29]. Continuing this research, Meulpolder et al undertake a comparative analysis of different torrent trackers' download speeds, both private and public (including The Pirate Bay). They conclude that the private trackers they studied had higher download speeds and higher ratios of seeders to leechers than the public trackers, including The Pirate Bay[27]. Taken together, this work shows that distinct improvements in download speed derive from tighter community control in BitTorrent communities.

2.3 Related research about open movements

In this section I will look at studies of other open movement projects and those that deal with the phenomenon generally in order to develop some of the principal themes that are relevant for the current study.

2.3.1 Quality in open movements

It is telling that one of the most common points of departure for commentary on open movements is an assertion of the high quality of many open movement products; for many, the typical absence of formal barriers to participation in such projects inevitably puts a question mark over the quality of the finished product. However, there exists a wealth of evidence that such products are often of sufficiently high quality to compete with products produced by conventional means.

Some of the best known examples of open source software are the Ubuntu operating system and the Apache server software. The Ubuntu operating system is generally considered to be of comparable quality to the other main operating systems, Microsoft Windows and Apple OSX[50], while Apache server software is far more widely used than any proprietary based rival[49].

The collaboratively edited encyclopedia Wikipedia has been extensively researched. One well-known study compared article accuracy in Wikipedia to that in Encyclopædia Britannica using a sample of articles from each, finding that the two samples had very similar number of errors, such that article accuracy could be said to be approximately equivalent between the two[19]. This conclusion was disputed by Encyclopædia Britannica who demanded a retraction based on methodological errors and exaggeration of their findings; Nature however defended the article, arguing that its conclusions were justified and its methodology was sound[6][30].

Other research on Wikipedia has tried to uncover the mechanisms through which articles become of high quality. One study found a correlation between number of editors, number of edits and article quality (as measured by selection for "Featured Article" status by Wikipedia contributors)[54], although this conclusion was criticised by a later study which argued that articles with increasing numbers of editors only became significantly better when this increase was matched by an increase in coordination between editors[26]. The authors of this later study argued that similar patterns could be found in other open movement endeavours, particularly open source software where highly interdependent work tends to be done by a core group of users while work requiring less

coordination such as bug reports is done by a more diverse group of participants.

This conclusion leads to the necessity to take into account different forms of participation and contribution to open movement projects. In relation to Computer Supported Cooperative Work, Schmidt has argued that all collaborative undertakings involve a division of roles and a potentially infinite level of discussion and negotiation about the project[45]. Thus, online collaboration systems must not only persuade users to contribute, but to contribute in a number of different ways, defined by the specific features and role structure of the system in question. Where discussion and negotiation are a crucial part of the system, it must persuade users to engage in these.

2.3.2 Explaining motivation in open movement projects

Another key question that faces open movement researchers is motivation; how can the mass voluntary participation in commons based production be explained? In Clay Shirky's *Here Comes Everybody*, Shirky presents three reasons why people might participate in open movement projects: 1) inherent pleasure in intellectual work, 2) a desire to make a mark on the world and 3) the desire to do something good[47]. Similar themes are echoed by other writers on the topic. Eric Von Hippel, a writer on innovation, devoted his book *Democratizing Innovation* to the study of open movements and the potentials for innovation that emerge from them. In discussing motivation he cites an open source programmer as follows:

Creation is unbelievably addictive. And programming, at least for skilled programmers, is highly creative. So good programmers are compelled to program to feed the addiction (Von Hippel 2005, p. 124).

According to Von Hippel, this creative impulse is best expressed within group contexts, as programmers are obliged to demonstrate their skills to an audience of their peers in order to receive recognition. Thus the twin motivators of self-expression and social approval are key drivers of open source software development[51].

Rafaeli et al. conducted a survey of Wikipedia participants in the English and Hebrew editions of Wikipedia where they found that the strongest motivators for participants were "Learning new things", "Intellectual Challenge", "Pleasure", "Sharing my knowledge" and "Contributing to other people"[41]. While the first three motivations chime with both Shirky and Von Hippel's assessments, the last two seem to reflect the social values exposed by the Wikipedia project itself; appeals to community and collaboration feature heavily on the Wikipedia site. This illuminates the persuasive appeal that values can have in encouraging people to participate.

2.3.3 Social psychology in open movements

In his 2003 book *Persuasive Technology*, Fogg devotes considerable time to discussing how social psychology can play a role in persuasive design. As part of this discussion, he sets out a number of social psychological principles which can be of use to analysis or development of persuasive tools. Although his focus does not include collaborative production sites, it is clear that I can adapt

these principles to apply to our cases. Below I list those principles together with some discussion and a related hypothesis that applies them to our own problematic[15].

- *Principle of social comparison* - this states that people's motivation to perform a behaviour will increase if they are aware how their performance compares with that of others, particularly others that are similar to themselves. This leads to the hypothesis that a site based on persuading users to create content in a collaborative manner will be more effective if users can compare their contributions with those of other users. In our cases I can think of the "user history" for each user that shows the full details of user edits in the case of OpenStreetMap and uploads in the case of The Pirate Bay.
- *Principle of normative influence* - this states that persons can be influenced by normative influence (peer pressure) to perform or not perform a certain behaviour. Normative influence involves people altering their behaviours in accordance with the opinions of a social group. From this, I can hypothesise that a site based on persuading users to create content in a collaborative manner will be more effective if it enables users to state and re-state motivational norms. For example, I can look at the value based appeals present on the Wikipedia site and in the community, whereby users are encouraged to contribute in accordance with these values. From our own cases I can think of the "Legal Threats" section on The Pirate Bay in which the site publicly ridicules copyright holders, thus creating norms of disrespect for the entertainment industry that may help to persuade users to engage in illegal behaviour.
- *Principle of social learning* - this states that a person who can see others being rewarded for performing a behaviour will be more likely to perform that behaviour themselves. Monetary rewards are not typically applicable to open movements, where contribution is by definition voluntary; however, there are many other type of rewards possible, for example, social approval. I can hypothesise that a site based on persuading users to create content in a collaborative manner will be more effective if it enables users to see other users being rewarded for performing target behaviours. In our examples, I can think of the "VIP" status given to high performing users in The Pirate Bay.

This last principle can be seen as critical to collaborative behaviour on the internet. In a discussion on online co-operation, Cheshire and Antin note that the internet is different to many other social environments because "individuals can only observe cooperative behaviour; they cannot necessarily see evidence of non-contributions"[9]. By "non-contributions" they are referring to users who contribute nothing but still consume the service, also known as "free-riders". Of course, it is also possible for hostile users to vandalise collaborative sites, through entering false information or uploading fake torrents for example, but in general users can only learn from the good behaviour of users and are therefore more likely to contribute in a positive way themselves.

Cheshire and Antin conducted a study of social psychological processes in what they call "internet information pools", but are to all intents and purposes

identical with our online collaboration projects. Arguing that monetary rewards are usually unfeasible within such projects, those developing them should instead focus on providing users with “intrinsic rewards” facilitated through awareness of social processes. Using a large sample of users who interacted with a website banner-based game they created and distributed on several dozen sites and a dedicated website they found significant evidence that social psychological feedback effected rates of continued participation. The types of feedback that they tested were: gratitude (e.g. “Thank you for contributing!”), historical reminder (e.g. “You have contributed three answers so far.”) and relative ranking (e.g. “You’re in the top 30% of contributors!”).

They found that all three forms of feedback had a strong influence on persuading repeat contributions from users who interacted with the banner on an external site, however, they had little influence on persuading repeat contributions from users who contributed on the main site. This implies that the context of interaction is critically important to the effect of persuasion mechanisms. They suggest that the difference in reactions of users is related to the different relations which users had to the project; if users accessed the site directly they were more likely to identify with the values and goals of the project and were thus not interested in receiving computer-generated feedback. Users interacting via the website banner on the other hand were not so interested in the project as such and were more likely to treat contribution as a form of game[9].

However, since the contributors on the internal site on average contributed far more than users of the website banners, this conclusion is not particularly encouraging for our own problematic which concerns contributors to specific sites and not to banner games. I can argue that there is a crucial distinction between computer generated feedback and real social feedback. While computer generated feedback may have a weak influence on committed contributors, it seems very possible that human feedback will have an influence within collaborative production environments.

2.3.4 Participation patterns in open movement projects

Discussion of participation patterns in open movement projects has been a common theme in much literature that deals with the subject. The fact of participation inequality in online collaborative environments has been observed as far back as 1998, where Whitaker et al. found that on average 2.9% of posters in UseNet newsgroups accounted for 25% of total posts in that newsgroup[53]. Nielsen has also discussed this problem, calling it the 90-9-1 rule, which assumes that for any 100 users of an online project, 90 will be lurkers (i.e. people who use the content without contributing anything), 9 will be low-level contributors, and 1 will be a heavy contributor. He cites participation figures in blogging, Wikipedia and Facebook Causes as evidence for this[32]. Shirky has also written about the issue claiming that in a typical “open movements” project, 80% of contributors contribute only 20% of content while 20% contribute 80% of content. This is known as a “power-law” distribution such that if contributors are ranked according to contribution amount, the user in the n th position will have contributed $1/n$ th of the amount contributed by the user who has contributed most[47].

Shirky and Nielsen have different reactions to the issue. Shirky sees it as

inherent to the functioning of such projects, a pattern that results from the “spontaneous division of labour” that drives online collaboration and that cannot be managed by attempts to “iron-out” participation inequality. Although Nielsen also accepts a certain level of inevitability to this pattern, he sees it as problematic as it means that the overall system is unrepresentative of most users and the voices of a minority of users will drown out those of the majority. He suggests that websites should take some steps to improve the participation ratio by, among others: making contribution easier, making contribution a side effect of other actions, rewarding contributors and promoting quality contributors[32]. Interestingly, making contribution a side effect is already present in the case of The Pirate Bay, where users who download a file automatically upload some of this file as they do so.

2.3.5 Participation patterns and bibliometrics

It is interesting to note that such discussions of power laws and participation patterns, bear a resemblance to certain key bibliometric concepts, notably, Lotka’s Law and Zipf’s Law.

Lotka’s law refers to patterns of productivity among authors. He writes that for every 100 authors who write one article, there will be 25 who publish two, 11 who contribute three, 6 who contribute four etc, i.e. that there is a decrease in performance on the basis of an inverse square [10].

Zipf’s law is based on linguistic analyses of texts; it predicts that in any written text, words will be used in a decreasing rate of frequency, such that the *n*th word will be used $1/n$ times as often as the highest ranking word. What is interesting about this law, is that Zipf argues that since it is based on basic patterns of human effort, it will be found in all areas where human production takes place [1]. As shown previously, it is also one of the formulations that Shirky uses to describe power law distributions.

Lotka’s law also seems relevant for our study, based as it is on productivity of individuals. It is not assumed that Lotka’s law should be taken as a mathematical truth, but rather as a general principle which guides human productivity.

2.4 Related studies about analytic evaluation

This section looks at previous work on analytic evaluation using it as the basis for developing our own framework for evaluating persuasion in online cooperative environments. It looks at analytic evaluation from two different perspectives, Interaction Design and Persuasive Design.

2.4.1 Analytic evaluation in Interaction Design

The practice of analytic evaluations of systems is well established within the general field of Interaction Design[46]. Analytic evaluation typically involves experts who adopt the role of end-users by staging step-by-step walk-throughs of a certain process such as booking plane tickets online. They often use heuristics as a guideline for identifying and reporting usability problems.

The most well-known set of heuristics are those proposed by Nielsen, who suggests 10 “rules of thumb” for evaluating user interfaces[31]. These heuristics

have subsequently been adapted by Baker et al. to assess “groupware”, i.e., systems designed for cooperative work[3]. In this work, they propose 8 heuristics for analysing cooperative systems. Although one would immediately assume that these can be of use to the present study, they suffer from a number of assumptions that limit their applicability to voluntary collaborative environments.

Foremost among these is the unspoken assumption that such systems will necessarily resemble traditional work environments. This assumption leads Baker et al. to presume synchronous use of the systems and a drive to emulate real world patterns of interaction [3]. Obviously, synchronous use of the systems is not realistic when one considers massively distributed online cooperation efforts taking part in many different time zones. At the same time, attempting to model systems to allow them to replicate real world forms of interaction through features such as audio chat and avatars seems unimaginative and doomed to failure. This is because the technical and social possibilities that underpin computer supported cooperative work are very different to those that underpin real life cooperative work, whether voluntary or paid. An example is the extremely low uptake of video chat; despite numerous efforts by manufacturers to launch it, consumers have resisted, presumably because of the technical limitations of the software combined with the different social possibilities enabled by voice only conversation (e.g. talking on the phone while walking)[46].

Of the eight, heuristic 7: "Allow people to coordinate their actions", seems very appropriate to our study. As already shown above, the best Wikipedia articles come about when a large number of editors operate within a well structured frame. It seems very likely that allowing coordination is a necessary step for creating valuable common pool resources.

2.4.2 Analytic evaluation in Persuasive Design

In the field of Persuasive Design, Fogg has proposed the “Fogg Behaviour Model” (FBM) as an analytic tool for assessing persuasion. He writes that for persuasion to occur, there must be a combination of motivation, ability and triggers. Although these elements are interdependent, high levels of one may cancel out low levels of another, for example if motivation to complete a task is very high, then a user may still carry it out, even if the task is difficult. Designing persuasive technologies involves boosting motivation or ability or both, while also ensuring that the desired behaviours are triggered at the appropriate time.

Fogg writes that there are three main binaries that effect motivation: pain/pleasure, hope/fear, and social acceptance/rejection. Ability (which he also calls simplicity) on the other hand, is affected by time, money, physical effort, brain effort, social deviance and non-routine activity. Meanwhile, there are three types of triggers; spark triggers, which aim to increase motivation, facilitators, which make target behaviours easier to do, and signals, which indicate when a behaviour is appropriate[17].

The clarity of this model is very useful for helping one think about persuasion in a dynamic way; one can see how site designers can remove barriers to ability or attempt to increase motivation as part of their persuasive strategy. On the other hand, the model seems most applicable to conceptualising one-off persuasive goals, such as persuading users to click on a “sign-up” link, or to purchase a product. The model is less intuitively useful when applied to large-

scale collaborative projects, which involve repeated actions by users over an extended period of time. Since our cases do not deal with one-off actions such as buying a book or something similar, it seems appropriate to look at behaviour change in a somewhat different way.

2.4.3 Analytic evaluation of our cases

In order to apply the FBM to our cases, it is necessary to begin by setting out the target behaviour desired by the system. If they are to be successful, online collaboration projects should be oriented towards converting users into contributors, converting contributors into repeat contributors and converting repeat contributors into core contributors. The focus should be on constantly expanding the base of core contributors while increasing the number of overall editors by making it easier for users to contribute. In this sense, while both motivation and ability are important at the first stage, as users become contributors, motivation will become more crucial later on as already capable contributors are converted into repeat and core participants.

As I have seen from our literature review, open movement projects have a number of common characteristics that underpin them. These aspects are as follows:

- Participation in projects is primarily made up of volunteers.
- Successful projects rely on active communities of users who identify with the values of the projects in question.
- Coordination of work tends to occur organically and without formal hierarchy.
- Projects tend to be based around certain values which are embodied in the products created.

The design of the projects should be oriented towards strengthening these aspects. The heuristics I propose in 3.7 attempt to combine an understanding of these with the Fogg Behaviour Model and the social psychological principles I outline above. These heuristics are then used as the basis of analytic evaluation of the cases.

Chapter 3

Methodology

This section introduces the methodology employed in the analysis of persuasion in collaborative online environments. It discusses the practical requirements for selecting data sources, explains the methodology used to retrieve data and the tools used to extract this into a workable format. It presents the research questions from Chapter 1 and shows the methodology by which these are to be tested. First though, I will present a glossary of terms used in order to allow the non-technical reader to follow the discussion.

3.1 Glossary of terms used

HTML - HTML (HyperText Markup Language) is a mark-up language that structures content on a webpage. It identifies different elements such as links, tables, titles and paragraphs.

Script - refers to a small program that is created to perform various operations.

Seeder/Leecher - in BitTorrent downloads (see Torrent, below) a seeder is a user who possesses a full copy of the file in question and is allowing others to download this file from their computer, while a leecher is a user who is downloading the file.

Shell - the shell is an interface for entering commands on a computer. In this case, the Linux shell BASH was used. It allows several different programs to run in conjunction with each other.

Torrent - a torrent is a file that is used to download another file from multiple peers. Files downloaded via torrent are typically broken up into multiple pieces to enable downloading different pieces from different peers. The torrent file contains meta-data about the pieces that make up the file and urls for torrent trackers that enable different peers to connect with each other.

URL - a Universal Resource Locator specifies the location of a particular resource and a protocol for retrieving it. In this study it refers principally to webpages.

3.2 Selection of data sources

This analysis focuses on user contributions to collaborative online systems. It analyses these contributions in relation to those of other users and analyses them in terms of development over time. Data sources thus needed to a) provide detailed information on user contributions including precise date of each contribution and b) this information needed to be publically accessible. Initially, it was supposed that Wikipedia would be the other subject for analysis, but a lack of access to detailed user data, due to Wikipedia's article-based organisation of edit history, made this impossible¹. The requirement for detailed time data for each user contribution, made several other systems unusable as research subjects. It is a common feature among websites to give time data in an informal way, such as "About two months ago", or "A year ago", instead of a usable date format. These approximate datings made sites such as YouTube² and TVShack³ unsuitable for research. Presumably, such sites possess exact data on their own databases, but this is not given on the publically accessible webpages. Based on these criteria, The Pirate Bay and Open Street Map were selected as cases for research.

This selection process is noteworthy in itself as it points to a key problem with analysis that relies wholly on 3rd parties to make data usable and publically available. The selection of sources for this research was very much determined by the standards in which websites displayed their information; when websites display data in an unusable format or do not display it at all, much web-based behaviour becomes inscrutable without some special agreement between researchers and websites. This condition becomes particularly difficult to satisfy when the data being recovered concerns activity of dubious legality, as in the case of The Pirate Bay, where it can be presumed that websites would be uninterested in providing researchers with data for fear of compromising the privacy of their users.

3.3 Data sources

The data sources for The Pirate Bay and Open Street Map are very similar. Each user's contributions are tracked automatically by the site. Information about these contributions is stored under a specific user page. In the case of The Pirate Bay, this information is highly detailed and includes: category of torrent file, its name, date uploaded, number of comments on the torrent (if any), and an additional indicator if the uploader is a "Trusted" or "VIP" user⁴. Open Street Map sorts user contributions in a similar way, but provides less detailed information than The Pirate Bay. Each contribution is given a unique ID, a time and date, space is given to an explanation or comment by the user, and the code for the area of the edit is also given. In the case of both systems, all available information was retrieved.

¹Ortega's 2009 thesis involved a user based analysis among others, but this analysis was only possible after a high level of processing of data. The technical skills required to perform this processing put such an investigation out of the reach of this research.

²<http://www.youtube.com>

³<http://tvshack.cc/>

⁴These are special statuses assigned by Pirate Bay administrators in recognition of a user's contributions to the community. They also act as guarantees of file quality for other users.

I should again note that the fewer details available for OpenStreetMap data led to fewer possibilities for analysis. While The Pirate Bay data could be analysed in terms of comments on contributions by other users and number of seeders per torrent, corresponding possibilities do not exist in OpenStreetMap, and so this information is not generated on OpenStreetMap edit histories. While it is the case that comments on contributions are made in other parts of the OpenStreetMap system (such as user forums and chat channels), these were not studied as part of this research.

Such considerations underline the difficulties inherent in comparing complex systems where user participation and coordination takes place on a number of different levels. For these reasons, this study focuses predominantly on contribution rates, instead of trying to examine the larger picture of participation in discussion and coordination of work.

3.4 Retrieval and processing of data

The desired data was HTML files of user histories accessible online. Thus, in order to retrieve these pages, it was necessary to get a list of urls for user profiles. This was achieved by entering the directories <http://www.thepiratebay.org/users/> and www.openstreetmap.org/users/ into Yahoo! Site Explorer. Yahoo! Site Explorer displays all pages connected with a specific url that are indexed by Yahoo!, as well as links from other sites on the web to these pages. It is then possible to download a thousand of the pages associated with this url (in these cases, those connected with sub-directories of the category user, i.e. specific user histories). In the case of The Pirate Bay, it was possible to download urls for user profiles from two separate sub-directories, since The Pirate Bay stores user urls at both <http://www.thepiratebay.org/users/> and at <http://thepiratebay.org/users>. This meant that there were roughly twice as many user histories available for researching The Pirate Bay as there were for OSM. In both cases, there existed some duplicate users in the lists retrieved from Yahoo! Site Explorer, as urls were sometimes retrieved for several of the user pages. In order to prevent a single user history being downloaded multiple times, these duplicates were removed using some terminal based commands.

Once a list of urls for user histories had been generated and duplicates removed, a series of scripts for downloading the user pages and extracting the relevant information were developed using a combination of shell scripts, Python (a programming language) and BeautifulSoup, an HTML/XHTML parser for Python which is designed to optimise data retrieval from HTML pages⁵. In both cases, full user histories were often not available on one HTML page, but were instead spread over a number of pages once more than a certain number of contributions had been reached (30 in the case of The Pirate Bay, 20 in the case of OpenStreetMap). Thus it was necessary to design a script that was capable of recognising links to additional history pages and adding these to the download queue in such a way as to keep all a user's files together. A further concern was that the websites in question would interpret the page requests as a hostile act or a drain on server capacity and thus block the IP address of the

⁵Post-Doctoral researcher Toine Bogers was a great help in this crucial part of the methodology.

```

<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0
Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en"
lang="en" dir="ltr">
<title>OpenStreetMap | Changesets by alv</title> </head>
<body> <div id="content">
<h1>Changesets</h1> <p>Changesets by <a
href="/user/alv">alv</a> </p>
<p>
&laquo; Previous
| Showing page 1 |
<a href="/user/alv/edits?page=2">Next &raquo;</a>
</p>
<table id="changeset_list" cellpadding="3"> <tr>
<th>ID</th> <th>Saved at</th>
<th>Comment</th> <th>Area</th> </tr> <tr>
<td class="table1 date"> April 16, 2010 14:05 </td>
<td class="table1 comment"> traffic:hourly </td>
<td class="table1 area"> <a href="/?minlon=24.9448809
&minlat=60.2088358&maxlon=24.9456711&
maxlat=60.2089752&box=yes" title="show area
box">24.945,60.209,24.946,60.209</a> <!--11015388--> </td>
</tr> <tr>
</div> </body> </html>

```

Figure 3.1: HTML extract from an OSM user page

requesting computer. In order to minimise this danger, the downloading script was designed to put intervals of varying lengths between page requests.

In order to process the data, scripts were designed to recognise and extract the required data on each page. This data was then saved in a table format and all non-relevant data was deleted. Since HTML is a file format designed for optimising the structured display of information in internet browsers, it is not specifically oriented towards presenting data in a format usable for researchers. To design programs capable of recognising the relevant data, it was necessary to analyse each site's HTML coding of user history pages and to design scripts to match the specific mark-up for each piece of information. This was done using a combination of shell based programs such as grep, awk and sed in the case of Open Street Map and code written in Python module "Beautiful Soup" in the case of The Pirate Bay. These scripts then saved the desired data in table format where it could then be analysed using spreadsheet software and statistics programs. A HTML extract from a typical OSM user history page is shown in 3.1 for purposes of illustration.

A full summary of the scripts used to extract data is given in Appendix 2. For illustration purposes, an extract from the script used to process the OSM data is shown in Fig. 3.2 on the next page .

In the case of data from The Pirate Bay, data was analysed primarily using Open Office Calc, an open source spreadsheet program, capable of carrying out both sophisticated calculations and generating graphs. In the case of Open-

```

#!/bin/sh
if [ $# -ne 2 ]
then
echo "USAGE: ./ $0 <HTML PAGE> <USER NAME>"
echo ""
echo " <HTML PAGE> Locally stored HTML page"
echo " <USER NAME> Name of the user I're crawling for"
exit 1
fi
#Get change ID
cat $1 | grep "View changeset details" -A 2 | awk
'BEGIN{FS=">"}{print $2}' | sed 's/<\a//g' | grep "." >
temp1.txt
#Get Date
cat $1 | grep -A 1 "table[01] date" | sed 's/<td
class="table[01] date">//g' | sed 's/--//g' | sed
's/ (still editing)/May 1, 2010 12:00/g' | grep ":"
| sed 's/January/1/g' | sed 's/February/2/g' | sed
's/March/3/g' | sed 's/April/4/g' | sed 's/May/5/g' |
sed 's/June/6/g' | sed 's/July/7/g' | sed 's/August/8/g'
| sed 's/September/9/g' | sed 's/October/10/g' | sed
's/November/11/g' | sed 's/December/12/g' | sed 's/,//g'
| awk 'BEGIN{FS=" "}{print $2 "-" $1 "-" $3}' > temp2.txt

```

Figure 3.2: Extract of script used to process OSM user pages

StreetMap data, immediate analysis through Open Office Calc was impossible as the table generated was too long. Instead, a Geographical Information System (GIS) program called MapInfo was used to generate some results as this was capable of handling the data. Other analyses were performed by dividing up the data into smaller pieces and using Open Office Calc.

3.4.1 Yahoo! Site Explorer

As stated, the data retrieved for this research is based on a set of urls in the user directories of the sites in question retrieved from Yahoo! Site Explorer. Site Explorer is a tool designed for webmasters to allow them to analyse which pages of their site are indexed by Yahoo! and which sites link to their site. Although it is not a tool specifically for researchers, it was used in this study because of its ability to easily provide a list of urls for user pages of both sites, information that was a requirement for the study to take place. Unfortunately, Yahoo! provide no information on how these pages are indexed, so it is impossible to tell if the users used in the study are representative of the population as a whole. Importantly, the results do not seem to be based on inlinks from other sites, within the user sub-directory of OpenStreetMap, Site Explorer retrieves 33,179 pages but only 10 inlinks. If it were the case that Site Explorer retrieved urls on the basis of inlinks, it would be probable that the users studied would be uncharacteristically productive, as it is unlikely that links would be found to pages of unproductive users.

3.4.2 Creation of productivity bins

If we take for granted the assumption that different users will have different levels of productivity, it makes sense to divide users up into bins based on productivity in order to allow analysis of how participation rates differ between the different types of users. Based on the discussion of participation inequality and power laws of contribution in online projects I decided to make three divisions based on a 60-30-10 rule. This division was chosen on examination of Lorenz curves.

As discussed in 3.5.2, Lorenz curves are typically used to visualise inequalities within samples or populations. They are derived by ranking all sample members by number of contributions starting with the lowest. A cumulative percentage for these users' contributions is then derived which forms the basis of the curve. As shown in 3.3 on the following page, the first segment of the distribution is between 0 and 0.6, where the graph is almost flat. The next segment is between 0.6 and 0.9 where the graph starts to rise more rapidly. The final section is between 0.9 and 1.0 where the slope of the graph is extremely sharp, indicating a few, very high performing users. Although the shape of the Lorenz curve is not exactly the same for both distributions, it was decided to use the same divisions in order to allow the bins from the different systems to be compared to each other.

Applied to TPB contributors, this method of division yields one group of 150 who have contributed 228,132 torrents, one group of 448 users who had contributed 35,518 torrents and one group of 897 users who had contributed 4491 torrents.

Dividing OSM contributors into the three groups, I get one group of 75 editors, one group of 229 editors and a final group of 456 editors. Bin One, the

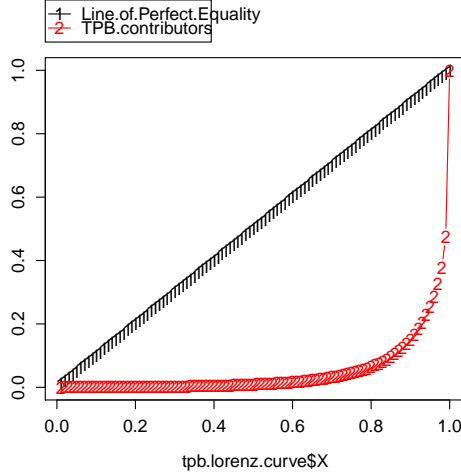


Figure 3.3: Lorenz curve of TPB contributors plotted against an equality curve

lowest level contributors had an average of 155.85 edits each and a median of 124 edits, making up 3.83% of total edits. Bin Two, the mid-level contributors had an average of 1,840.68 edits and a median of 1002 edits, making up 22.75% of total edits, Bin Three, the highest level users had an average of 18,130 edits and a median of 19,859 edits making up 73.40% of total edits.

3.5 Analysing Participation Patterns between case studies

Our first questions deal with the patterns of participation found in the cases studied. I am interested in testing the "power law" analysis of participation proposed earlier, and in seeing whether the same pattern applies to both samples. The results from these questions will then provide the backdrop for the analysis of persuasive features involved in both projects.

3.5.1 Q1: How do user contribution patterns resemble and differ from each other? In particular, what is the difference between drop-out rate of new participants and lifetime contribution patterns in the two systems.

To answer this question I perform two forms of analysis, an analysis based on user drop-out rate and an analysis based on contribution rates over user lifespans.

3.5.1.1 Drop-out Rate

Drop-out rate refers to the proportion of users that only contribute a relatively small amount or for a relatively short period before dropping out of activity. Drop-out is measured using both user contribution amounts and user lifetimes. This data is presented in a series of relative frequency distributions showing the the percentage of users that have contributed various amounts and the percentage of users whose lifetimes were between 1 and 7 days. The lifetime data is then graphed over a year, to show what percentage of the sample were still active within a year of their first participation event. By combining two different measures, it is hoped to minimise the problem of comparing contribution rates across systems that involve very different forms of contributions, as discussed in 1.6.

3.5.1.2 Lifetime based analysis

A lifetime based analysis is used to compare how participation rates alter throughout user lifecycles. In order to do this, a spreadsheet tool is used that assigns each user's contributions a value based on how many days have elapsed since that user's first ever contribution. In this way, a comprehensive picture of how each user's activity over time within the project is built up. The number of days between the first contribution and the last contribution of a user is used to represent the total lifespan of that user.

In order to analyse lifespan productivity rates, I divided each sample up into three groups based on the methodology discussed in 3.4.2. Frequency distributions were created based on days, these showed the number of contributions that happened in relation to user lifetime, i.e. all contributions made on the first day of a user's lifetime would be grouped together, regardless of when that user first became active. These numbers were then divided by the total number of contributions for that user group and multiplied by a hundred to give a proportionate amount for that user group. Thus each frequency distribution shows what percentage of that user group's total contributions were made at each stage of their lifespan.⁶

3.5.2 Q2: Does participation in these projects follow standard rules (e.g. power laws) or is it different from case to case?

In order to answer this question, I will compare the sample data with several of the power laws mentioned in 2.3.5.

Lorenz Curve First, I will use a Lorenz curve to compare inequality rates between the samples. A Lorenz curve is a graphical measure of inequality, typically used to visualise income inequality within countries, i.e., it shows the proportion of wealth in a country owned by different portions of society. In order to generate a Lorenz curve all users are listed by contribution starting from the lowest to the highest. Each contributor is divided by the total number

⁶For help with this part of the Analysis I am grateful to TessaES from the OpenOffice.org Community Forum <http://user.services.openoffice.org/en/forum/viewtopic.php?f=9&t=30690&start=0>

of contributors and each number is added to the last to give the cumulative percentage of the population. Each user's contributions are divided by the total number of contributions, and cumulated to give the cumulative percentage of contributions. In this way, at any point it can be shown what proportion of the population has created the corresponding proportion of contributions. A line of equality is created by plotting a sample based on the cumulative percentage of a perfectly equal rate of participation. This line of equality then serves as basis of comparison, the closer the Lorenz Curve is to the line of equality, the more equal the distribution of contributions is in that sample.

Pareto Principle Participation patterns will also be compared to the Pareto principle. This principle was originally developed by an Italian economist who used it to describe the distribution of wealth in a country, observing that 20% of the population typically owned 80% of the wealth. Recently, the principle has also been applied to explaining participation patterns in online activities. In our examples, I will use it to see how much of the contributions are created by 20% of the population in each case. Comparing the two will show how dependent each site is on a minority of "elite" contributors and will test Shirky's assertion that unequal participation is inevitable within online projects.

Lotka's Law Lotka's Law is a concept originally developed to describe productivity patterns within academia. It states that the number of researchers who publish n papers is $1/n^2$ of the number of researchers who publish only 1 paper. So if 100 authors write 1 paper, then $1/2^2(100) = 0.25(100) = 25$ will publish 2 papers, $1/3^2(100) = 0.11(100) = 11$ will publish 3 papers etc. I am interested in seeing whether this law of scholarly productivity can be applied to productivity of participants in online collaboration.

Zipf's Law Zipf's Law is based on content analyses of texts. Zipf found a numerical relationship between the frequency at which words are used, such that the n th most frequently used word will be used $1/n$ times the number of times the most frequently word is used. Interestingly, Zipf argued that his law could be applied to any field of human endeavour. For this reason, I will test whether it can be applied to contribution rates in online collaboration by using the number of contributions made by the highest level contributor as the basis for prediction.

3.6 Participation patterns internal to projects

- *Q3: How does user contribution correlate with other features of user participation within these systems? Does feedback have an effect on contribution rate? How can quality be assessed in relation to contribution rate?*

This part of the analysis focuses on how user participation correlates with other features of use. Since this section concerns information that is not available within both systems, these analyses will necessarily be internal, comparing different groups of users against each other. The variables it will focus on are: quality of uploads and the role of feedback mechanisms, both in The Pirate Bay.

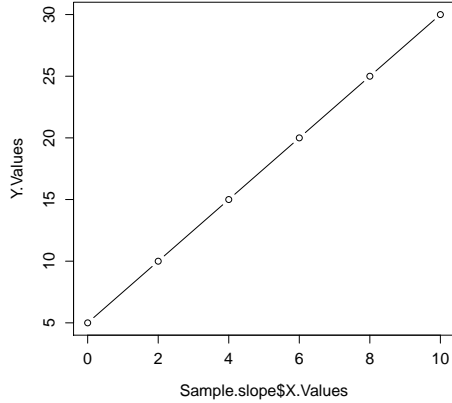


Figure 3.4: Sample correlation line

3.6.1 Statistical model of correlation

The analysis for this subsection is based on correlation between variables. It will investigate whether there are correlations between contribution rates and other variables such as quality of uploads, feedback from other users and user lifetime. Correlation between quantitative variables is typically represented as a linear relationship whereby the explanatory variable (i.e. the variable that is used to explain changes in the other variable) is represented on the x axis and the response variable (i.e. the variable that changes on the basis of the explanatory variable) is represented on the y axis. The slope of the graph created thus represents a relationship between the response variable and the explanatory variable of a sample.

This relationship is expressed by the formula $y = \alpha + \beta x$, where β represents slope and α corresponds to the y-intercept, i.e. the point on y where $x = 0$.

In Fig. 3.4, the upwards sloping line indicate a positive correlation between variables, whereby an increase in the value of x leads to an increase in the value of y . The angle of slope represents the strength of the correlation. Note though, that the slope of the graph is affected by the scale of the units of measurement, thus making comparison between different sample sizes based on slope impossible.

When applying this model to real data it is common to construct a scatterplot diagram in order to visualise the level of correlation. By observing the scatterplot, it is possible to judge whether or not the relationship between variables seems linear. A scatterplot with a U shaped distribution for example, will indicate that a linear relationship is not present. If it is decided that the relationship seems linear, the equation $\hat{y} = a + bx$ is used to predict the relationship. In that case, a represents an estimate of α and b represents an estimate of β .

This prediction equation will create a linear estimate of the relationship between x and y . Of course, it is unlikely that distributions will follow a strictly linear pattern, so it is also necessary to represent the degree of derivation from the prediction line. This is done through the sum of squares equation:

$$SSE = \sum (y - \hat{y})^2$$

In this equation, $y - \hat{y}$ represents the difference between any point y and the predicted value for that point according to the prediction equation.

None of the equations explained thus far can represent the strength of correlation between variables in a standard way. The slope of a prediction equation b can show whether a correlation is positive or negative, but its size is affected by the scale of the units of measurement. In order to compare correlations in a standard way, it is first necessary to derive the marginal sample standard deviations for both x and y . This is returned through the equations:

$$s_x = \sqrt{\frac{(x - \bar{x})^2}{n - 1}}$$

$$s_y = \sqrt{\frac{(y - \bar{y})^2}{n - 1}}$$

Once these values have been returned, the formula $r = \left(\frac{s_x}{s_y}\right) b$ denotes the degree of correlation between variables x and y . Since $-1 \leq r \leq 1$ this formula allows correlation to be tested in a standard way.

3.6.2 Role of feedback mechanisms in increasing participation

This sub-question looks at whether there is a correlation between feedback on user contributions and their rate of overall participation. Since it is not possible to comment directly on edits in OpenStreetMap, this question applies only to The Pirate Bay. It looks at whether comments on torrent uploads have any correlation with level of contribution.

To answer this question a script was used which calculated the average number of comments every user had received⁷. Then, the correlation between average number of comments and number of uploads was tested using R Commander. If feedback mechanisms such as comments play a role in encouraging repeat contributions, there should be a positive correlation between avg. number of comments and number of uploads.

3.6.3 Quality analysis of torrents

This sub-question proposes a methodology for determining quality of torrent uploads and applies it to investigate whether Pirate Bay users with more contributions tend to make uploads of higher quality. The methodology for determining quality of torrent contribution is a modification of that proposed by Hirsch for assessing researcher impact. That is:

“A scientist has index h if h of his/her N_p papers have at least h citations each, and the other $(N_p - h)$ papers have no more than h citations each.”[23]

⁷My thanks to the users of <http://www.programmingforums.org> and particularly user "Pat-sie" for their help with this script.

Hirsch argued that this method is far more robust than other measures of research quality because it takes into account both productivity (i.e. number of papers published) and consistency of paper quality (i.e. it will not be skewed by one highly cited paper).

To apply this formula to torrent uploads, I will reformulate this as follows:

“A pirate has index h if h of her/his N_t torrents have at least h seeders each, and the other $(N_t - h)$ torrents have no more than h citations each.”

I use the number of seeders as representative of the number of users interested enough in an upload to support it with their own bandwidth. While leechers could also be used to generate the "piracy h-index", leechers presumably do not possess a full copy of the file and are so unable to make any quality assessment of it. Seeders on the other hand, have a full copy of the file and give away some of their bandwidth to make it available to other users, thus implying that they believe the file is worth sharing. In this way, the index represents a "collective assessment" on the part of Pirate Bay users of a specific user's uploads, in much the same way as the h-index represents the assessment of a scientist's work by their peers. Applying this to torrents allows us to judge the impact of any contributor on the file-sharing site. It will then be possible to ask what variables correlate with high quality contributors, such as productivity, lifetime etc.

However, using this methodology to judge user quality is not without its flaws. Foremost of these is the fact that the number of seeders for a torrent changes over time. A new episode of a TV show might have 300 seeders in the first week after coming out, 100 the following week and as few as 10 six months later. Such patterns are presumably different from category to category; uploads of program files will most likely retain seeders longer than uploads of TV shows or movies. Thus a user's h-index will be affected both by the time the data is collected and the category in which the user tends to operate.

It can also be argued that a h-index will underestimate the value of one off contributors, who can at most have an index of 1, for example, a one off user with a single upload attracting 200 seeders will have the same h index as a similar user who only managed to attract a handful of seeders. It would also be possible to use average seeders per torrent as a method for measuring impact, but averages will tend to be affected too much by outliers. Moreover, the h-index is purposely weighted to value user productivity. For these reasons, I will use an unmodified h-index in this research as a best available tool for judging user impact.

In order to retrieve the H-Index, I used a series of formulas in OpenOffice Calc that counted the number of torrents each user had contributed and then found the number of torrents for which the number of seeders was equal to or greater to that number.

3.7 Analytic evaluation of persuasive features

- Q4: How can persuasive features of the above systems explain differences in user participation?

This question sets out to connect the results generated by the quantitative study with persuasive features of the systems themselves. Answering this question

will involve a walk-through of the persuasive features involved in contributing to these systems, based on a set of heuristics which I outline below. These heuristics are based on the Fogg Behaviour Model and the key concepts developed in the literature review.

With that in mind, I propose several heuristics for evaluation of online collaborative projects.

1. The values that underpin the site should be clearly visible to all users and should be reinforced regularly.
2. Triggers to participate in the project should be visible to users of the product.
3. Participation in the project should be as simple as possible and documentation of technical aspects should be easily available.
4. The project should encourage users to identify themselves with the project and feel as if their contributions are valued through rewards, feedback or other mechanisms.
5. The interface should facilitate interaction with other users and coordination of collaborative efforts.

These heuristics take into account the need to motivate users through values, the need to ensure simplicity of participation, the need to trigger participation, the need to enable self-organised coordination of efforts and the need to develop a community of users and leverage social influence. In this study I will walk-through participation in our cases analysing them on the basis of these heuristics.

I can see that the above heuristics involve different levels of engagement, with each becoming more or less relevant as one progresses through the system. For example, triggers to participate will be most relevant at the first level of engagement, as will values, whereas technical documentation only becomes relevant after one has first decided to participate. Social rewards and coordination will then become important on a third layer when one is already an engaged participant but lacks social reinforcement and encouragement.

I freely acknowledge the “ad hoc” nature of these heuristics. They are based on a limited amount of experience of such systems combined with a broad overview of related literature. Their relative lack of precision can be seen as an inevitable result of trying to apply new perspectives to a new field. Their application here should be taken as an opportunity for further refinement through experience.

It is also important to add that this analytic walk-through cannot be anything but a shallow account of project features. Self-organised and decentralised systems like the ones in question typically evolve highly complex structural and social dynamics such that it would take many months of engagement within the projects to fully appreciate all these processes. What I can do with this analysis is give an overview that can perhaps go some of the way in explaining the results generated in the other analyses.

Chapter 4

Analysis

In this chapter I present the results of the analysis under six main sections. These sections refer to the main findings derived from analysis. First, I present a summary of the data retrieved and discuss problems with data accuracy.

4.1 Summary of data retrieved

4.1.1 Summary of The Pirate Bay data

Once the scripts were written the data retrieval process was set in motion. The retrieval of data from The Pirate Bay was completed overnight on one computer. 2000 urls were retrieved from Yahoo! Site Explorer, of which 112 were duplicates (i.e. two different pages on the same account). After removing these, 1,888 unique users remained. Downloading user histories for these users returned information about 268,141 torrents produced by 1,495 users. The 393 users for which no torrents were returned likely indicates users who created accounts but never uploaded any torrents. This was confirmed by checking the user profiles for a small sample of the users for which no torrents were received. However, it was decided not to take these 393 as representative of the number of users who register an account but do not contribute.

4.1.1.1 Contribution rates

This gives an average contribution of 179.36 torrents per user. The median for the sample was 10 torrents and the mode was 1. The standard deviation was 1279.26, indicating a highly variable sample. The difference between the median and the average moreover indicates a sample in which the vast majority of contributors are below the average. This is further shown when I compare the standard deviation to the mean; 98% of users fall between $\bar{x} - s$ and $\bar{x} + s$, indicating the presence of a small percentage of outliers that heavily skew the sample. Note also that the range of the sample is 29,999, with 1 being the smallest number of observations and 30,000 being the highest. The figure of 30,000 is significant, because it is at this number of torrents that the retrieval script was programmed to stop retrieving more data. This suggests that this user had most likely contributed more than 30,000 torrents, but as this user is the only one in the sample with 30,000 uploads, it seems that the download

limit of 30,000 torrents per user for the script did not have a large effect on the sample.

4.1.2 Summary of Open Street Map data

The data for OpenStreetMap was considerably more difficult to retrieve due to the slowness and unreliability of the OpenStreetMap server. It was necessary to break up the retrieval process over two computers and replace the user list with a separate text file for each user. Using this methodology, the retrieval process took approximately 5 days. The original list of urls from Yahoo! Site Explorer contained 1000 urls, of which 999 were users, 994 of which were unique users. Downloading user histories for these 994 users retrieved information about 1,884,104 map edits, contributed by 762 users.

It is important to note that some of the users for whom no edits were retrieved actually have edit histories, indicating that the lack of results was at least in some cases the effect of problems with the OSM server rather than evidence of unproductive users. The problems with the data retrieval cast a shade of uncertainty over much of the OSM analysis, it is undeniable that there is a certain amount of flawed data, it is furthermore impossible to tell with certainty which data is flawed and which is not. However, one can reasonably suppose that the data retrieval process was considerably more likely to go awry for users with many contributions, as there was a proportionally higher chance of the server timing out while retrieving their history, rather than a user with only one or two pages of history.

4.1.2.1 Contribution rates

The OpenStreetMap 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 299.5. The range of the sample was 31,637, with the lowest number of edits being 1 and the highest number of edits being 31,638. The mode of the sample was also 1. As with The Pirate Bay sample, this sample indicates a high level of variability, with a standard deviation of 5552.09 and 87% of the sample falling between $\bar{x} - s$ and $\bar{x} + s$.

The top figure in the OSM distribution is quite notable for its relative smallness; the average number of contributions per user in the OSM sample is ca. 13.8 times greater than that of The Pirate Bay sample, while the top contribution in the OSM sample is only 1.05 times that in The Pirate Bay sample (this second figure is also likely to be an underestimate). This suggests immediately that the OSM project is less dependent on a few high performing individuals than The Pirate Bay, but this conclusion awaits further confirmation and is questionable when one bears in mind the problems with data retrieval.

4.2 Drop out rate

The rate at which users drop out of a project after becoming involved can tell us a lot about the persuasive success of the project in question. In this section, I look at the number of users who only participate briefly in both systems.

The most basic way to measure drop out rate is through a frequency distribution that compares the number of users in each system that have only a

Number of contributions	OpenStreetMap	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)

Table 4.1: Proportion of users contributing between 1 and 5 times

Number of contributions	OpenStreetMap	The Pirate Bay
1-10	4.85% (37)	50.76% (760)
11-20	3.01% (23)	10.75% (161)
21-30	2.62% (20)	5.67% (85)
31-40	3.28% (25)	3.40% (51)
41-50	1.57% (12)	2.60% (39)
≤ 50	15.35% (117)	73.21% (1,096)

Table 4.2: Proportion of users contributing between 1 and 50 times (intervals of 10)

small number of contributions. The results of two such frequency distributions are presented in Tables 4.1 and 4.2. As I can see from these results, the rate of drop out in The Pirate Bay is many times higher than that of Open Street Map, with 50% of Pirate Bay users contributing between 1-10 times and 17% contributing only once.

However, I must note that it is not unproblematic to compare gross contributions to the systems in this way. Editing Open Street Map will most likely involve adding many points at a single time, whereas there is no necessary connection between adding one torrent to The Pirate Bay and adding another.

For this reason it is necessary to look out drop out rate through another lens, that of user lifetime. Table 4.3 shows the proportion of users in both samples whose lifetime is up to one week while Figure 4.1 on page 40 shows the proportionate decrease in participation in both samples over a year, taking 100% participation as the starting point.

Days	OpenStreetMap	The Pirate Bay
1	9.06% (68)	21.67% (323)
2	1.86% (14)	2.48% (37)
3	0.26% (2)	0.87% (13)
4	0.26% (2)	0.67% (10)
5	0.26% (2)	0.93% (14)
6	0.26% (2)	0.6% (9)
7	0.93% (7)	0.33% (5)
≤ 7	12.93% (97)	27.58% (411)

Table 4.3: Percentage of users dropping out of activity during the first week of lifetime

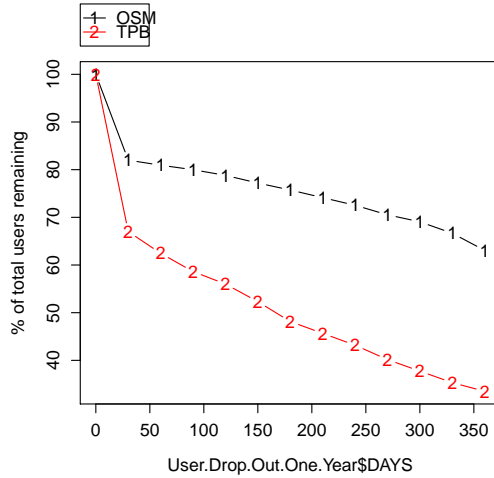


Figure 4.1: Percentage of users remaining active over one year

This data confirms the finding that Pirate Bay users tend to drop out of activity far sooner than their Open Street Map equivalents do. Furthermore, the average rate of user drop out per 30 day interval over a year in TPB (-5.54%) is significantly higher than that of OSM (-3.07%), indicating that TPB loses users at a faster rate than OSM.

4.3 Contributions over lifetime

In this analysis, I look at how user productivity develops over user lifetime. This analysis is based on the productivity divisions specified in 3.4.2. Participation patterns for groups are compared based on their relation to other groups from the same project and to the corresponding group from the other project, i.e. the low level contributors from one project will be compared to the medium and high level contributors from that project and also to the low level contributors from the other project. All groups contributions are compared on a daily basis over two weeks, on a weekly basis over six months and on a monthly basis over two years. Tables from those analyses not presented here are located in Appendix 3.

In the analysis, two general tendencies became clear. First, that the pattern for all distributions was one of gradual decline, whereby the majority of contributions were made during the early stages of lifetime and decreased over time. Second, several of the distributions displayed a step-like pattern of decline, in that contribution rates tended not to decrease smoothly, but to decrease to a lower level and stay on that level for some time before decreasing again.

4.3.1 Lifetime analysis of Open Street Map editors

For the lifetime analysis of OSM editors I have only analysed low and mid-level contributors. This is due to obvious flaws in the data retrieved for the highest

	Low-level contributors	Medium level contributors
Day 1	1.62 (1,154)	1.39 (2,729)
Day 2	1.05 (749)	0.15 (298)
Day 3	0.80 (575)	0.09 (176)
Day 4	0.61 (440)	0.07 (152)
Day 5	0.68 (487)	0.05 (115)
Day 6	0.53 (382)	1.43 (2,790)
Day 7	0.56 (398)	4.67 (9,113)
Day 8	0.90 (644)	2.51 (4,907)
Day 9	0.56 (397)	4.03 (7,878)
Day 10	0.59 (419)	13.94 (27,209)
Day 11	0.59 (422)	4.19 (8,180)
Day 12	0.53 (377)	0.11 (224)
Day 13	0.50 (357)	0.07 (151)
Day 14	0.42 (302)	0.06 (118)
\sum	10 (7,102)	32.81 (64,040)

Table 4.4: % of contributions by low-level and mid-level users over the first two weeks of their lifetimes

level contributors, 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 and as such lifetime based analysis of their contributions was not thought to be representative.

4.3.1.1 Contributions within the first two weeks of user lifetimes

Table 4.4 on page 41 summarises the contribution rates of low-level and medium level OSM contributors within the first two weeks of their lives.

What is interesting in these distributions is the striking differences between the two distributions. Both groups contribute approximately the same number of edits during the first day of their lifespans, the mid-level users contribute proportionally less thereafter, until the sixth day, where they suddenly begin to contribute far more, rising as high as nearly 14% of total edits on the tenth day of activity. After the eleventh day, the mid-level distribution sinks below the low-level distribution in both proportionate and absolute terms.

This finding is rather surprising, one would expect that the low-level contributors would have a higher percentage of contributions in the first few days of their lifespans and thereafter to have a consistently lower rate of contribution than the mid-level contributors. Instead of a steadily decreasing rate of participation in the project, the mid-level users contributions spike suddenly almost a week after their first contributions. This spike does not correspond with any similar increase on the part of the low level users. This might suggest that users interest in the project surges after about a week, this surge marking a user out as a highly committed participant in the project. On the other hand, this anomaly could be the result of flaws in the data retrieved.

It is also interesting to note that mid-level users in total contribute over 30% of their edits in the first two weeks of activity, while low-level users contribute

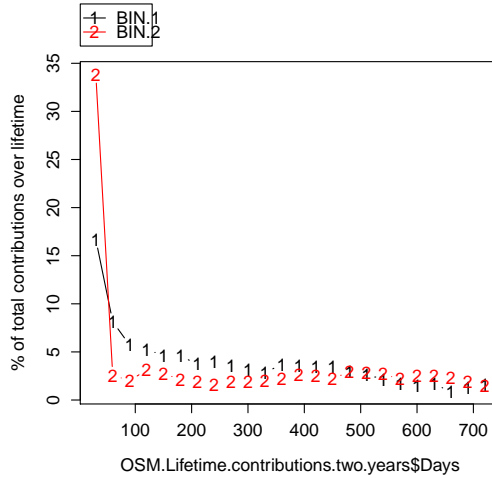


Figure 4.2: Contribution rates over first two years of lifespan

only 10% during the same period. One would expect that the result would be the reverse, but one can see that this high rate comes from the sudden spike that comes after the first week.

4.3.1.2 OSM contribution rates over two years

Figure 4.2 on page 42 shows the lifetime contribution rates of low-level and mid-level OSM users over the first two years of lifetime. This graph makes clear the unusual nature of the OSM mid-level user activity, as they contribute proportionally far more than low-level users in the early days of lifetime before sinking below low-level users. It is only at 510 days I can begin to see the proportion contributed by the mid-level users start to increase over that of the low-level users.

It is interesting to note that while these distributions appear to be highly one-sided, with the vast majority of user contributions occurring in the very early periods of activity, one can see that after this extremely high early period, the activity remains quite consistent, even increasing at some points, before eventually falling again. This seems to suggest that although users are most active at the start of their lifetimes, they fall into a rhythm of participation after this initial period is over.

4.3.2 Lifetime analysis of participation in The Pirate Bay

In this section I will look at how participation patterns alter throughout the lifetimes of the various samples of Pirate Bay users.

Days	Bin 1	Bin 2	Bin 3
1	32.05% (1,456)	3.86% (1,371)	1.04% (2,381)
2	6.09% (277)	1.93% (688)	0.63% (1,442)
3	2.17% (99)	1.00% (355)	0.66% (1,524)
4	1.65% (75)	0.86% (306)	0.16% (369)
5	1.80% (82)	0.72% (256)	0.59% (1,362)
6	1.18% (54)	0.68% (242)	0.16% (365)
7	0.92% (42)	0.73% (261)	0.18% (409)
8	0.92% (42)	0.67% (239)	0.19% (447)
9	0.74% (34)	0.61% (218)	0.57% (1,314)
10	0.55% (25)	0.54% (192)	0.18% (411)
11	0.70% (32)	0.48% (171)	4.52% (10,311)
12	0.35% (16)	0.55% (198)	0.17% (392)
13	0.68% (31)	0.51% (182)	0.60% (1,379)
14	0.48% (22)	0.44% (158)	0.57% (1,316)
Σ	50.34% (2,287)	13.62% (4,837)	10.26% (23,442)

Table 4.5: Percentage of contributions by all TPB user groups over the first two weeks of lifetime

4.3.2.1 Participation patterns over the first two weeks of user lifetime

Table 4.5 on page 43 compares participation rates among the different contributor groups in the first two weeks of their lifespans.

In these distributions one can see the huge proportion of contributions that low-level users (Bin 1) contribute in the first days of their lifespans. 50% of all their contributions are made within the first two weeks after their first contribution. This pattern is not so distinct for the mid-level (Bin 2) and high-level (Bin 3) contributors, although one can see that the mid-level users follow a similar pattern, although far less extreme. The high level pattern is rather strange for the sudden spike in contributions made 11 days into their lifespans. Upon closer inspection of the data, it became clear that this spike is caused by one extremely active user who has contributed a vast number of torrents on this one day. Without this user's contributions, contributions on the 11th day would not be statistically significant. It is worth bearing in mind the large effect single users can have on such a sample, where the sample size is relatively small and individual users are so productive.

4.3.2.2 Participation patterns over two years

Table 4.6 on page 44 shows contribution rates at intervals of 30 over a two year period. This table shows that all groups of users contribute the proportionally highest number of contributions during the very early stages of their lifespans. Although present in all distributions, this tendency is most pronounced for the low-level editors and least pronounced for editors with the highest number of contributions, with mid-level contributors falling in between.

It is also interesting to look at how the productivity rates of high-level and mid-level users develops on a smaller scale. One can see from Table 4.6 on

Days	Bin 1	Bin 2	Bin 3
30	56.79% (2,580)	20.40% (7,242)	14.63% (33,388)
60	7.19% (327)	10.27% (3,646)	5.38% (12,285)
90	5.48% (249)	7.91% (2,808)	7.15% (16,328)
120	3.10% (141)	6.96% (2,472)	3.95% (9,015)
150	3.43% (156)	5.77% (2,049)	4.08% (9,323)
180	2.62% (119)	4.88% (1,733)	3.70% (8,451)
210	2.15% (98)	3.47% (1,232)	3.89% (8,890)
240	2.02% (92)	3.60% (1,278)	3.05% (6,974)
270	1.80% (82)	3.19% (1,135)	3.08% (7,031)
300	1.49% (68)	2.93% (1,043)	2.42% (5,533)
330	1.14% (52)	2.80% (995)	2.13% (4,878)
360	0.88% (40)	2.36% (838)	2.11% (4,820)
390	1.18% (54)	2.38% (845)	2.67% (6,100)
420	1.03% (47)	2.17% (773)	2.50% (5,712)
450	0.59% (27)	2.33% (830)	2.44% (5,568)
480	0.81% (37)	1.74% (618)	2.13% (4,867)
510	0.59% (27)	1.60% (570)	1.99% (4,559)
540	0.57% (21)	1.32% (471)	1.78% (4,068)
570	0.46% (26)	1.16% (412)	1.65% (3,765)
600	0.48% (22)	1.11% (395)	1.54% (3,534)
630	0.48% (22)	1.18% (419)	1.56% (3,567)
660	0.57% (26)	1.39% (496)	1.39% (3,175)
690	0.48% (22)	0.96% (341)	1.37% (3,147)
720	0.22% (10)	0.91% (324)	1.26% (2,883)
Σ	95.64% (4,345)	92.87% (2,528)	77.96% (177,861)

Table 4.6: % of contributions by low-level, mid-level and high-level TPB users over the first two years of their lifetimes

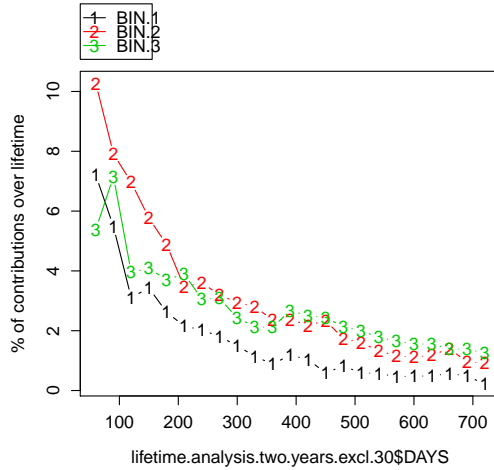


Figure 4.3: Contribution rates over first two years of lifespan, excluding the first 30 days of lifespan

page 44 that the percentage contributed by high-level users between 60 and 90 days increases nearly 2%, from 5.38% to 7.15%. What is interesting about this increase is that it is occurring in a period when the productivity rates of the other user groups are decreasing rapidly. If I look at Figure 4.3 on page 45, in which the first 30 days of activity have been removed to make the changes more visible, I can see that the contribution rate of the high-level users is subject to many temporary increases in activity, despite a general pattern of decreasing activity. This step like pattern can also be seen in the activity of the other two user groups, but it is not as pronounced or as frequent. While it is hard to draw conclusions from such a pattern, it seems to suggest periods of renewed interest in the project, after the peak of activity within the first few weeks of lifetime.

4.3.3 Comparison of lifetime participation rates across systems

In this section I will compare the lifetime participation rates between systems. Comparison will be based on the bin division specified, with each user group being compared to its counterpart from the other system. High-level users will not be compared due to aforementioned uncertainties regarding data accuracy.

4.3.3.1 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. 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.

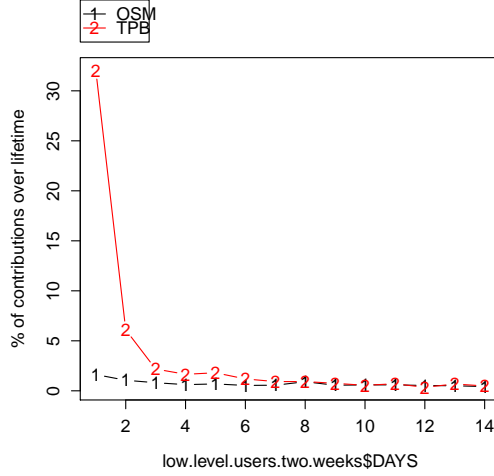


Figure 4.4: Contribution rates by low-level users over first two weeks of lifespan

It is only after about 14 weeks that OSM contribution rates start to be significantly 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, as can be seen in 4.5. 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%. 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 figures 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.

4.3.3.2 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 the extremely large rise in OSM contributions relative to those of TPB users after the sixth day. 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 leads to 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 av-

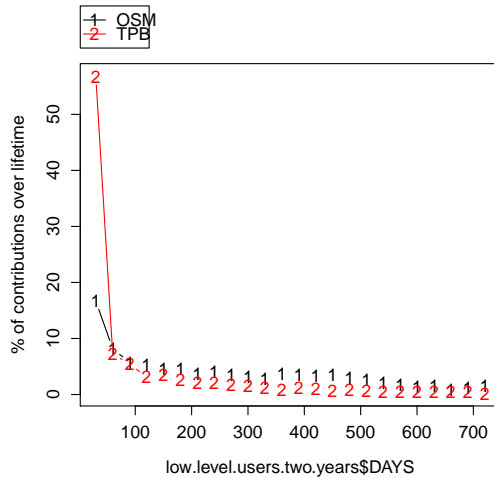


Figure 4.5: Contribution rates by low-level users over first two years of lifespan

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.

4.4 Power Laws and Contribution Inequality

In this section I test whether the data from the samples matches the power laws described in 2.3.4 and 2.3.5.

Figures 4.8 and 4.9 on page 49 show the Lorenz curve for each system, graphed against a Line of Perfect Equality. These curves show the scale of inequality within each system. As I can see, although both curves resemble power law distributions, it is clear that The Pirate Bay's distribution is considerably more unequal; the slope of the curve begins rising rapidly far later than it does in Open Street Map.

In order to test the Pareto Principle which states that 80% of the effects are given by 20% of the causes, I analysed how percentage of the total contributions were created by the top 20% of contributors. In The Pirate Bay, the top 20% contributed 93.61% of uploads. In Open Street Map, the top 20% contributed 89.63% of uploads. Both distributions obey the Pareto Principle although The Pirate Bay is slightly more dependent on its top contributors.

Lotka's Law states that every 100 researchers who publish 1 paper, 25 will publish 2, 11 will publish 3 etc. Testing it against the two samples produced the following comparisons where Actual shows the number of participants contributing each amount while Predicted shows the number predicted by the formula of Lotka's Law based on the number contributing once:

These tables show that participation in these cases does not resemble that predicted by Lotka's law, although participation in The Pirate Bay more closely resembles the pattern.

Zipf's Law tests participation rates using the highest level contributor as the

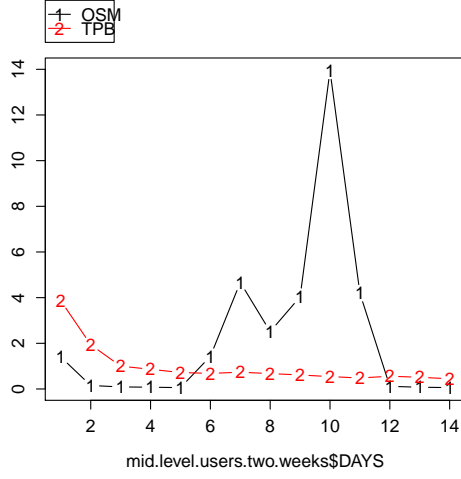


Figure 4.6: Contribution rates by mid-level users over first two weeks of lifespan

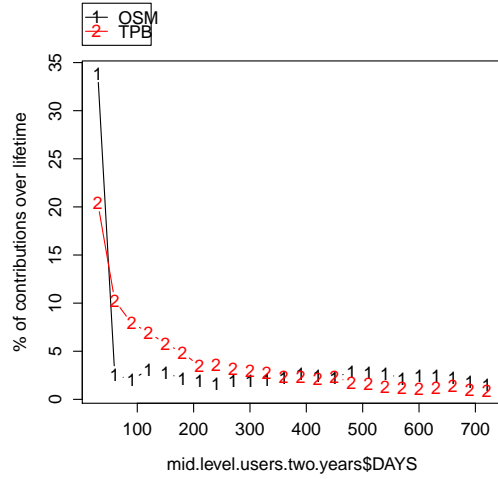


Figure 4.7: Contribution rates by mid-level users over first two years of lifespan

No. of contributions	No. of contributors	Predicted
1	10	10
2	10	3
3	1	1
4	4	0
5	2	0

Table 4.7: Comparison of participation drop-off in Open Street Map with Lotka's Law

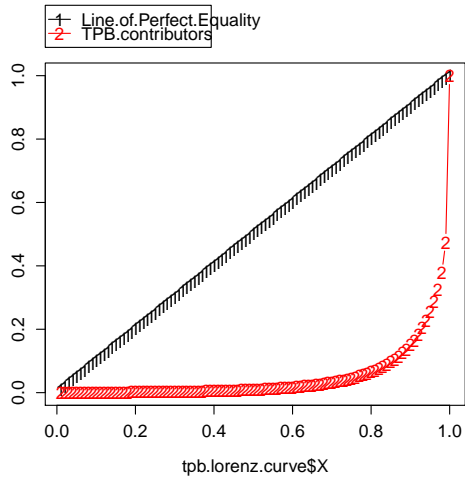


Figure 4.8: Lorenz curve of participation in The Pirate Bay

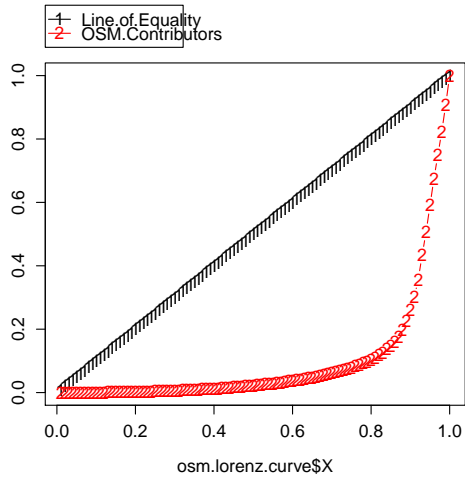


Figure 4.9: Lorenz curve of participation in Open Street Map

No. of contributions	No. of contributors	Predicted
1	258	258
2	125	65
3	96	29
4	76	16
5	45	10

Table 4.8: Comparison of participation drop-off in The Pirate Bay with Lotka's Law

Contributor rank	No. Contributions	Predicted
1	31,638	31,638
2	20,039	15,819
3	20,037	9,491
4	20,034	7,910
5	20,033	6,328

Table 4.9: Comparison of top 5 contributors in Open Street Map with Zipf's Law

Contributor rank	No. of contributions	Predicted
1	<i>58,372 (estimate)</i>	58,372
2	29,186	29,186
3	12,904	19,457
4	12,868	14,593
5	7,625	11,674

Table 4.10: Comparison of top 5 contributors in The Pirate Bay with Zipf's Law

basis for prediction, stating that the n th contributor will contribute $1/n$ as often as the most productive contributor. As stated, the history for the highest ranked Pirate Bay contributor was most likely cut off by the retrieval script. Therefore, for the purposes of this analysis I have predicted his number of contributions according to Zipf's Law, based on the number contributed by the second most prolific contributor. Thus, the match between the prediction and the reality in the first two ranks of The Pirate Bay table are artificial.

As can be seen from Tables 4.9 and 4.10, neither of the samples match Zipf's Law particularly well. However, the mismatch is for different reasons. The 2nd - 5th contributors in Open Street Map are all extremely close to each other and do not match the pattern of decline found in the Zipf prediction. This unusual pattern may well be the result of aforementioned problems with data retrieval which led to several of the user histories not being fully downloaded. In contrast, the decline in participation by Pirate Bay contributors is even more rapid than that proposed by Zipf's law.

4.5 Heuristic Analysis

In this section, I conduct a heuristic evaluation of OSM and TPB. The heuristics are as follows:

1. The values that underpin the site should be clearly visible to all users and should be reinforced regularly.
2. Triggers to participate in the project should be visible to users of the product.
3. Participation in the project should be as simple as possible and documentation of technical aspects should be easily available.

4. The project should encourage users to identify themselves with the project and feel as if their contributions are valued through rewards, feedback or other mechanisms.
5. The interface should facilitate interaction with other users and coordination of collaborative efforts.

As stated, these heuristics are applied differently at different levels of engagement. In our analysis I will go through three separate levels of involvement, at the first level is use of the site, there the challenge for the project is to motivate users to engage and trigger this engagement. The second level is for registered users who want to engage, how do they do this, is it easy or difficult? The third level is for users who have already participated, the challenge here for the system is to retain their engagement and encourage them to participate more and to engage with other users. Note that screenshots of the site walk-throughs together with more detailed analysis are provided in Appendix 1.

4.5.1 Analytic Evaluation of Open Street Map

4.5.1.1 Step One: Surface level use

The front page of OSM largely consists of a map and a sidebar. In the sidebar there are links to Help documentation, Copyright and License information, News blog, Shop and Map Key. There is a box for entering search queries, a short text in a small sized font explaining the project and a link for making donations. Above the map there are buttons for "View", "Edit", "History", "Export", "GPS Traces" and "User Diaries". In the right hand corner there are links for "log in" and "sign up". The "Edit" button brings one to a screen requesting login information or suggesting that the user can create an account. According to my schema, the most important factors at this stage are motivation and triggers. Triggers are highly visible on the screen; the "Edit" button and the "sign up" links are both triggers for participation. However, the values of the site are not so visible, the text explaining the values of the site is quite small and removed from the main attraction of the page, the map window. The meaning of the logo, which consists of a magnifying glass held over a map, is not at all obvious.

Registering for the site involves entering a small number of details (one screen) and accepting a licensing agreement (another screen). The licensing agreement states that all user contributions are in the public domain.

4.5.1.2 Step Two: New participant

Upon registering as a new user one receives a mail from Open Street Map thanking one for joining and providing links to various project resources. This email is a finely crafted example of persuasion, it is written in a friendly, informal manner and provides links to a Beginner's Guide, a videocast series, two blogs and encourages the new user to make contact with other users in their area. In sum, it makes the user feel welcome, increases their ability to participate and attempts to embed them in social networks. One of the blogs pointed to appears to be a good source of motivation for OSM participants: it is updated regularly with content with regular weekly projects for mappers to take part in,

selected images showing what can be achieved with OSM data and discussions about the OSM project. The blog thus helps users to feel their contributions are valuable, reinforces identification with the project and encourages them to participate more. It is also worth noting that many of the projects mentioned are humanitarian in nature, which presumably increases user loyalty to the project.

4.5.1.3 Step Three: Participation and Co-ordination

Once a user has registered and explored the project infrastructure the next challenge is to participate in the project. In the Beginner's Guide OSM provide detailed information about how users can participate, with the preferred method being through taking GPS traces. Since the cheapest GPS units available cost around €40, this represents a very high barrier to participation. It is also possible to add points through tracing over satellite imagery, or local knowledge, but GPS mapping is presented as the de facto standard. The wiki provides detailed information about GPS units to help people make a purchase. I would expect that this barrier would prevent a lot of people who register for OSM from participating actively in the project.

It is interesting that there is no integration between the automatic email sent upon registration and the location of the user (specified after registration) given the highly local nature of most OSM collaboration. There exists a high level of co-ordination within many countries including discussion forums and task lists, but the new user has to go through several levels in order to find these. It could be more effective for the system to ask users to specify their location upon registration (or derive this based on user IP) and then include in the introduction email links to discussion fora and project pages based on this.

4.5.2 Analytic Evaluation of The Pirate Bay¹

4.5.2.1 Step One: Surface level use

The front page of The Pirate Bay is primarily a search engine and the links to other parts of the site including the blog, the forum and the sign up form are not very prominent. The logo is a very effective piece of branding, expressing the site's values in a single image. There are statistics about the number of people using TPB which may help to give new users a feeling of community. There is also a prominent link entitled "How do I download?" which gives users instructions for downloading. There are no obvious triggers for users to contribute content.

4.5.2.2 Step Two: New participant

The Pirate Bay does not send out confirmation emails for newly registered users. When a user logs in they can see information about the number of torrents they have contributed and their IP address and they can alter some settings. The forum is not linked to more prominently from this page. The forum requires another login. There is plentiful documentation for new users as well as the ability to ask for more help. The site blog is not updated very frequently and

¹Note that at the time of evaluation (August 2010), it was impossible to register as a new user for The Pirate Bay. Therefore, I used a friend's account to investigate the internal system.

some of the content is very juvenile, much like the “Legal Threats” section ². There is little attempt to connect the site’s activities to a broader social context or movement.

4.5.2.3 Step Three: Participation and Co-ordination

Detailed instructions for uploading torrents are provided on the user forum. Uploading a torrent does not require anything other than a Pirate Bay account, a BitTorrent client and a copy of the file in question. The forum also allows users to request others to upload content and to promote their own content. The comment option on torrents also serves as a form of co-ordination, with users providing links to related content, requesting seeders and discussing torrent quality. Once a user has registered, there is no obvious sub-group for them to be active in, as is the case in OSM where users are expected to organise by locality and by project. It seems likely however, that users elect to be part of a sub-forum that matches their own interests, such as Film, TV or Music. A more developed mechanism for users to make connections and collaborate on uploading and seeding or on other related activities might be useful for such sites as it would add a more strongly social dimension to participation.

4.5.3 Summary of heuristic analysis

- 1. The values that underpin the site should be clearly visible to all users and should be reinforced regularly.
 - OSM - values are not very visible to surface level users but they are present in a regularly updated blog, and are often embodied in the various Projects of the Week.
 - TPB - prominent logo neatly captures many of values. However, there seems to be little readily accessible content discussing the broader context of file-sharing and copyright laws.
- 2. Triggers to participate in the project should be visible to users of the product.
 - OSM - “Sign Up” and “Edit” triggers are visible from front page. The Project of the Week acts as a recurring trigger for participants.
 - TPB - “Register” button visible from front page but not immediately obvious. No obvious encouragement to upload torrents present.
- 3. Participation in the project should be as simple as possible and documentation of technical aspects should be easily available.
 - OSM - plentiful documentation for new beginners linked to in registry email, including a beginner’s guide and screencast videos.
 - TPB - a large selection of tutorials are provided in the forum, as well as the ability to ask further questions on the forum.
- 4. The project should make users feel as if their contributions are valued.

²<http://thepiratebay.org/legal>

- OSM - Symbolic rewards are given to users based on the number of points they have uploaded, however these are on user pages and are not particularly obvious.
 - TPB - Users can achieve Trusted and VIP status based on their contributions to the site. In the forums users are given ratings based on how helpful their posts are.
- 5. The interface should facilitate interaction with other users and coordination of collaborative efforts.
 - OSM - Project encourages co-ordination via OSM Wiki. However, users must search for the appropriate forum or mailing list. There is a strong focus on making connections with other OSM users in one's area.
 - TPB - There is a single forum which is easy to find and has a large amount of material. The forum also makes it possible for users to request torrents. The comment feature on torrents enables users to request seeders, provide links to subtitles, and rate torrent quality, among other things.

4.6 Miscellaneous Analyses

This part of the analysis focuses on characteristics of participation that cannot be compared across the systems. These analyses tests for correlations between these variables and other features of participation in the systems.

4.6.1 Measuring quality of contributions to The Pirate Bay

As discussed, I will use a h-index based on seeders on torrents in order to judge the quality of a specific user's uploads. I tested for correlation on the basis of the three productivity bins outlined above and also on the basis of the sample as a whole. The results of these tests are summarised in figures 4.10, 4.11, 4.12 and 4.13.

In these figures, the dotted line represents the line predicted by $\hat{y} = a + bx$, that is the line for which SSE is at a minimum. The correlation values for the four different samples are as follows; for the sample overall, the correlation value is 0.35, for the core users the correlation value is 0.21, for the mid-level users the correlation is 0.38 and for the low-level users the correlation is 0.68. It is worth comparing these with the median H-indexes for the three samples, for the overall sample, the H-Index is 3, for the core users the median H-Index is 25, for the mid-level users the median H-Index is 8, while for the low-level users the median H-Index is 1.

It is not surprising that the correlation between uploads and H-Index is strongest for low level users; as already explained, users with low levels of productivity simply cannot earn a high H-Index. As these users begin submitting more torrents it is possible for them to earn a higher H-Index with a relatively low level of seeders per torrent. The next strongest correlation is for mid-level users, which suggests that these users also benefit from this effect, but to a much

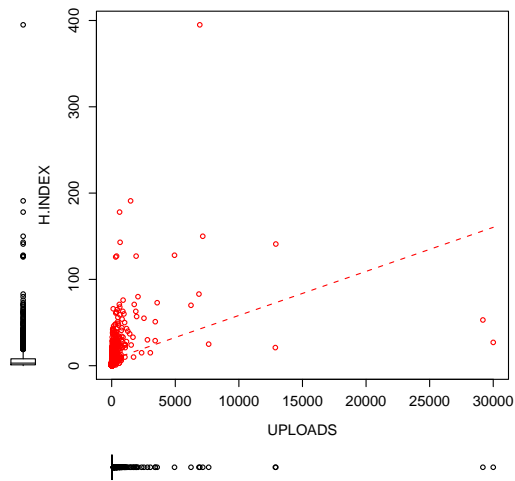


Figure 4.10: Quality Analysis of overall TPB sample

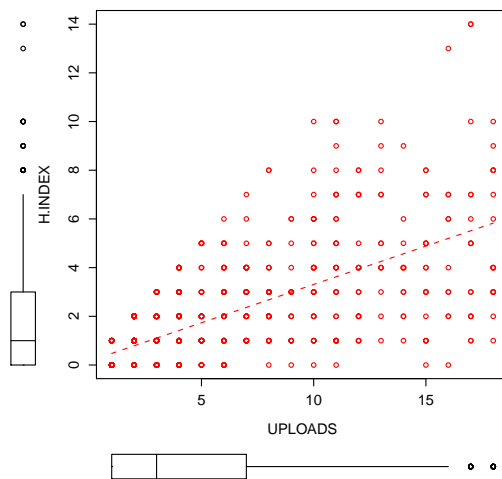


Figure 4.11: Quality Analysis of low level TPB users

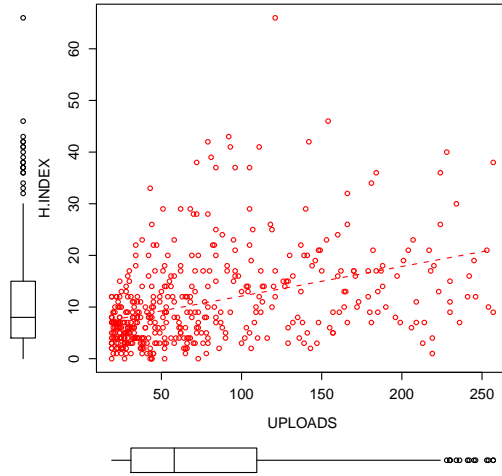


Figure 4.12: Quality analysis of mid level TPB users

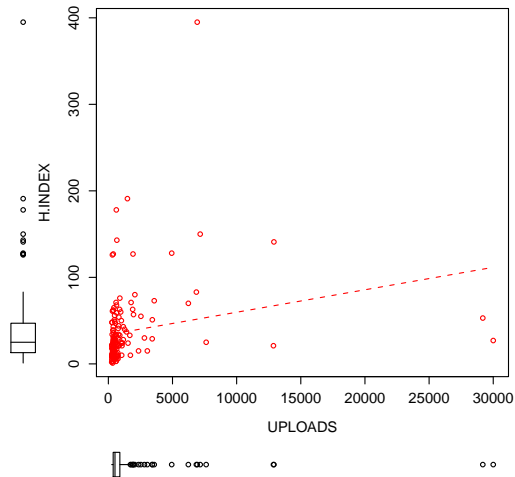


Figure 4.13: Quality analysis of TPB high-level users

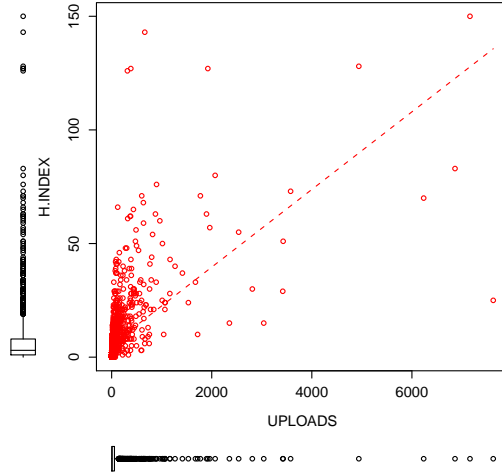


Figure 4.14: Quality analysis of TPB sample overall with several outliers removed

lesser extent. The correlation for the overall sample is similar to that of mid-level users, but with the difference that there are a far greater number of outliers present in this sample (indicated by the circles in the box plots on the graph margins). Outliers have a significant effect in skewing the correlation value and it is interesting to see the effect on the correlation if these are removed.

Figure 4.14 shows the scatterchart for the overall sample when six outliers (the top three outliers from both axes) are removed from the sample. This distribution appears to be more linear in nature and the correlation value is increased considerably, from 0.35 to 0.59.

The sample of core users has the lowest level of correlation and is also characterised by a relatively large number of outliers. It is interesting to note that the two highest uploaders in the sample have H-indexes below \hat{y} while the user with the highest H-index is the seventh largest uploader, submitting less than $\frac{1}{4}$ of the amount that the highest user has.

Overall, the evidence suggests that while there is reasonably strong correlation between uploads and user impact/quality, this correlation becomes less pronounced the more torrents one submits. While users with higher uploads tend to have higher H-indexes, the users with the highest H-indexes are not those with the most uploads.

4.6.2 Effect of feedback on participation rates in The Pirate Bay

In this analysis I test the correlation between the average number of comments a user receives on their torrents and both the total number of contributions they have made and their number of days they are active in the project. If there is a causal link between a user receiving comments on their torrents and that user participating more, I would expect there to be a positive correlation between

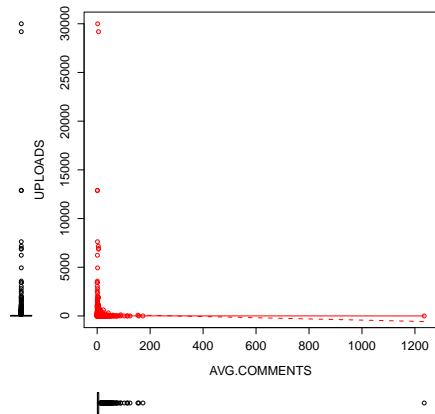


Figure 4.15: Correlation between number of uploads and average number of comments

the user's average comments per upload and their number of contributions and lifetime in the project. This connection was not visible in the evidence.

First I tested the correlation between average number of comments and the total number of contributions a user had made. There was found an overall correlation for the sample of -0.01. This pattern remained true when the correlation was tested for all groups of users; although the negative correlation was strongest for the core contributors and weakest for the low-level contributors.

Following this, I tested the correlation between a user's average number of comments and their lifetime in the project. Here too, I found a correlation of -0.01 for the entire sample.

Taken together, there appears to be little evidence in the sample for a connection between feedback on torrents (in the form of comments) and a user's number of contributions to the project or their longevity. Interestingly, in both samples there is one outlier who has a large impact on the sample. This user has contributed 1 torrent (and thus has a lifetime of 1 day) which received 1,204 comments.

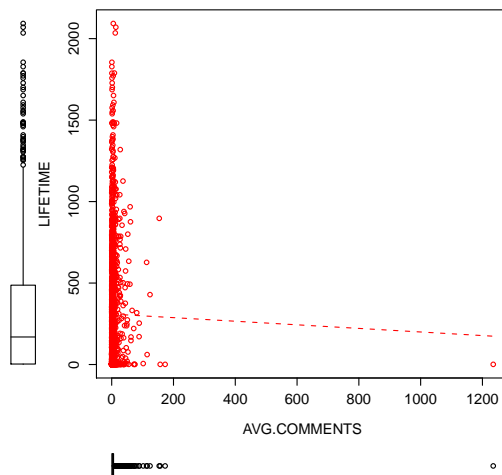


Figure 4.16: Correlation between user lifetime and average number of comments

Chapter 5

Results and Conclusions

This study set out to shed light on the role of Persuasive Design in maximising user participation in collaborative online projects. In this chapter I will discuss the results of the quantitative analysis presented in Chapter 4 in relation to the research questions set out and the literature reviewed presented in Chapters 1 & 2.

5.1 Summary of results in relation to research questions

- *Q1: How do user contribution patterns resemble and differ from each other in Open Street Map and The Pirate Bay? In particular, what is the difference between drop-out rate of new participants and lifetime contribution patterns in the two systems?*

The data analysis showed some important differences between contribution patterns in the two projects. On the most basic level, users in Open Street Map tended to contribute many more times than users of The Pirate Bay. This was not unexpected given the different forms of contribution in the systems.

User drop-out A common trend in both samples was a high number of contributors who dropped out of the project after only contributing a small number of times. As shown in the Analysis very large proportion of TPB users drop out after a very small number of uploads (between 1 and 5) while a very small proportion of OSM users do the same. This high rate of user drop-out cannot be fully explained by differences in contribution forms: Nearly 22% of TPB users drop out of activity after only one day of activity, while 9% of OSM users do the same. This suggests that OSM is better at persuading first-time users to remain with the system than TPB is. However, it is also possible that the relatively high barrier to entry for participating in OSM, created by the cost of GPS, also has the effect of deterring less committed users, leading to a higher level of commitment among users in general.

Contributions over lifespan The lifespan based analysis shows a common pattern of users contributing proportionally large amounts in the early days

of their lifespans and their contributions gradually dwindling away after this. The prime difference between OSM and TPB is in the rate of this decrease. In general, OSM users are active for longer than TPB users and their contributions decrease at a slower rate. This suggests that OSM is better at maintaining user motivation than TPB.

Of interest in the distributions was a step shaped rather than a smooth decrease in productivity. This pattern was evident in both the OSM and TPB samples at all levels. It is hard to know what it indicates, I would suggest that it shows phases of re-engagement with the project after a decreased level of involvement; users are not just gradually fading out, they are going through spurts of enthusiasm for the project as they gradually contribute less over time. It seems to show that despite decreasing activity, many users maintain loyalty to the projects.

One unusual result in the analysis was the tendency of mid-level OSM users to contribute a proportionally high amount during the second week of user lifetime. I can not offer any likely explanation for this anomalous result and would suggest that it results from flawed data.

- *Q2: Does participation in these projects follow standard rules (e.g. power laws) or is it different from case to case?*

The Analysis also tested the application of several laws of bibliometrics and social sciences to the samples. It found that the productivity of participants resembled a “power-law” distribution in that a majority of contributions were created by a minority of participants. Both samples were even more biased towards “the productive few” than suggested by the “80-20 Law” or the Pareto Principle. The Lorenz curves for both projects showed that while both samples followed the classic power law distribution, the Open Street Map distribution was somewhat less unequal. How this result is affected by faulty data however, is hard to predict.

However, the distribution of productivity in the samples did not match that which would be predicted by either Zipf’s Law or Lotka’s Law, although it seems plausible that participation rates within The Pirate Bay may follow some modification of Lotka’s Law. Looking at Lotka’s Law, the analysis makes clear the role of effort in contributing to the systems. Lotka’s Law is based on the rate at which academics submit research papers and its various permutations recognise the different productivity rates and submission cultures within different disciplines. Lotka’s Law seems somewhat appropriate to analysing The Pirate Bay, because uploading a torrent, like submitting a research paper, is a discrete event not necessarily connected to uploading another. Editing Open Street Map on the other hand, will most likely involve making many edits at the same time, thus Lotka’s Law is inapplicable.

- *Q4: How can persuasive features of the above systems explain differences in user participation?*

The clear differences between participation patterns in the projects seems to support Nielsen’s point of view regarding participation inequality; while some patterns are inevitable, site designers can have an effect on them nonetheless. Meanwhile, the analytic evaluation of the projects pointed to several key differences between the persuasive design of the systems.

There are two key differences discussed in the quantitative analysis, one, that TPB has a far higher drop out rate than OSM, and two, that OSM users stay active in the project for longer and contribute at a more even rate throughout their lifetimes. These differences require us to think on two different levels, on the level of one off users and on the level of repeat contributors. In both cases, the problem is likely to be one of either motivation or triggers, as first time contributors have already participated, so their ability is presumably high.

In terms of motivation, the analytic evaluation drew attention to OSM's sophisticated mechanisms for motivating users. OSM uses its first contact with newly registered users to welcome them to the community, point out help resources and encourage them to connect with other mappers in their area. The OpenGeoData Blog¹ connects users' contributions to broader themes of open information and humanitarian assistance. The "Project of the Week" tool constantly sets new goals and targets for mapping, thus acting as a recurring trigger for participation. All the while, the social nature of the OSM community embeds mappers in a network of their peers, where social approval derives from project contribution. The mapping parties and local gatherings of such local groups likely provides another form of recurrent triggering for participation.

The Pirate Bay, on the other hand does not send out a similar welcome message. Educational resources are provided and are easy to find, but the user must look for them. The Pirate Bay's blog fails to make meaningful connections with the larger movements around issues of copyright, censorship and open information instead allowing its tone to be dominated by antagonism with media corporations. This represents a significant lost opportunity to motivate users and mobilise supporters. Moreover, the social aspect of The Pirate Bay (represented primarily in its user forum) does not appear to be as developed as that of Open Street Map. The project does not promote autonomous user organisation (unsurprising given the semi-illegal nature of project participation), thus limiting the strength of the social ties that can be developed. Recurring triggers are present in the forum via requests for material, however these triggers are by nature limited to those users who possess or can get that material, thus diminishing their usefulness.

- *Q3: How does user contribution correlate with other features of user participation within these systems? Does feedback have an effect on contribution rate? How can quality be assessed in relation to contribution rate?*

Torrent Quality and User Participation The H-Index analysis found a correlation between number of uploads and increased torrent quality, this correlation was strongest at the lowest level and weakest at the highest. As discussed, using the H-Index is not unproblematic for analysing Pirate Bay contributions due to the fact that the number of seeders on a torrent changes over time. Despite this, I suggest that a modified form of the H-Index could still be appropriate for measuring pirate impact. Modifications could base the formula on the highest number of seeders a torrent has received (requires constant access to torrent databases) making it more reliable. Such a formula could be used by designers of peer-to-peer sites for assigning quality rankings to contributors, for example.

¹<http://www.opengeodata.org>

The effect of feedback on participation Based on the review of social psychological literature as well as previous studies on "common pool information resources", I proposed to analyse the effect of feedback on participation rates, where feedback was measured by number of comments on torrents. It was hypothesised that users whose torrents had on average a larger number of comments would participate in the project for a longer time and make more contributions. This result was not backed up by the evidence. Analysis found a very low negative correlation between average number of comments per torrent and user lifetime and between average number of comments per torrent and number of uploads by that user. This suggests that comments on a user's torrents do not have any impact on that user's level of participation in the project. This finding reinforces Cheshire and Antin's conclusion that feedback was not strongly related to repeat contributions in users who participated in a project directly, i.e. users who already had high levels of motivation [9].

I should note however, that comments on torrents do not represent the only way of providing feedback to users in The Pirate Bay. Feedback also occurs via participation in the user forum, where users request specific torrents and comment on each others' levels of participation. The connection between these interactions and a user's participation in the project were not included in this study.

5.2 The value of persuasion

This paper is based on the underlying assertion that the theoretical framework of Persuasive Design is of value to the study of online collaboration. At this point it is worth considering whether this assertion has been reinforced or discredited by the reality of the study. Persuasive Design has influenced this study in several ways: first, the assertion that participation rates are not inevitable but vary between different systems, second, the assumption that these variations are caused by persuasive features of the systems, and third, the methodological basis for investigating these persuasive features, i.e., heuristic analysis.

The first point has been convincingly proven by this study. Participation rates have been shown to be significantly different between the two systems. This has been true in respect to both gross contribution amount, user longevity and participation rates over user lifetime.

The second and third point are more difficult. In order to develop the heuristics used to evaluate site design, this study combined the Fogg Behaviour Method with insights from literature related to collaboration in general. In applying these heuristics it was found that unlike The Pirate Bay, Open Street Map has a strong focus on social connections between mappers and uses recurring weekly projects to motivate contributors. These factors could very well be responsible for the longer lifetimes of OSM users and their generally more consistent contribution rates. However, there is no clear way to prove this connection. This seems to be a problem more generally with combining heuristics and quantitative analyses - while interesting connections can be made, it is empirically very difficult to prove these.

In sum, although the use of Persuasive Design has thrown up some important methodological problems, there are substantial benefits to using this theoretical framework. One, its framework focuses on differences between sites,

thus allowing site designers and project leaders to consider interventions that will increase project participation. Two, it provides a useful starting point for heuristic evaluation and a set of concepts that site designers can incorporate into their projects.

5.3 Relevance of this study to future research

This study has both methodological and theoretical relevance to future studies. On a theoretical level, the study argues for the importance of considering persuasion in analysis of online collaboration and provides a prototypical example of such analysis. It is hoped that the arguments presented here can encourage further analysis of persuasion in online communities and thus assist the development and expansion of “open movement” projects. The use of the “H-Index”, while not a methodological novelty, represents a novel approach to considering on line file-sharing and in particular, user impact. This study has also contributed to the understanding of the role of feedback in online collaboration, reinforcing the rather surprising conclusion that feedback plays little role in increasing participant contributions.

Methodologically this study has shown how user histories can be used to carry-out system wide analyses of participation in online collaboration projects. It has presented “drop-out rates” and lifetime participation rates as useful concepts for analysing persuasive success in collaborative systems and shown how these can be measured. It has reinforced the assertion that participation rates in collaborative projects are not set in stone but vary to a certain degree from project to project. Perhaps most interestingly, it has developed and applied heuristics for analysing persuasion in online collaborative environments. These heuristics can be used by researchers as a starting for further analytic evaluation of persuasion in online communities.

Bibliography

- [1] Lada A. Adamic. Zipf, Power-laws, and Pareto - a ranking tutorial. <http://www.hpl.hp.com/research/idl/papers/ranking/ranking.html>, 2002.
- [2] Jonas Andersson. For the Good of the Net: The Pirate Bay as a Strategic Sovereign. *Culture Machine*, 10:64–108, 2009.
- [3] K. Baker, S. Greenberg, and C. Gutwin. Heuristic Evaluation of Groupware Based on the Mechanics of Collaboration. In *Proceedings of the 8th IFIP Working Conference on Engineering for Human-Computer Interaction (EHCI'01)*, 2001.
- [4] BBC-News. Piracy law cuts internet traffic. <http://news.bbc.co.uk/2/hi/7978853.stm>, April 2009.
- [5] Yochai Benkler. *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. Yale University Press, 2006.
- [6] Encyclopaedia Britannica. Fatally Flawed: Refuting the recent study on encyclopedic accuracy by the journal Nature. http://corporate.britannica.com/britannica_nature_response.pdf, March 2006.
- [7] Michael A. Carrier. The Pirate Bay, Grokster and Google. *Journal of Intellectual Property Rights*, 15:7–18, 2010.
- [8] C. Cheshire. Social Psychological Selective Incentives and the Emergence of Generalized Information Exchange. *Social Psychology Quarterly*, 70:82–100, 2007.
- [9] Coye Cheshire and Judd Antin. The social psychological effects of feedback on the production of internet information pools. *Journal of Computer Mediated Communication*, 1:705–727, 2008.
- [10] Russell C. Coile. Lotka’s Frequency Distribution of Scientific Productivity. <http://www.cna.org/documents/5500021600.pdf>, February 1978.
- [11] Brian Dessent. Brian’s BitTorrent FAQ and Guide. <http://www.dessent.net/btfaq/>, 10 2003.
- [12] enigmax. Pirate Bay Announces IPREDATOR Global Anonymity Service. <http://torrentfreak.com/pirate-bay-announces-ipredator-global-anonymity-service-090323/>, 03 2009.

- [13] Ernesto. How and Why BitTorrent Works, a Visualization. <http://torrentfreak.com/how-and-why-bittorrent-works-a-visualization-100217/>, 02 2010.
- [14] Ernesto. The Pirate Bay, A Year After The Verdict. <http://torrentfreak.com/the-pirate-bay-a-year-after-the-verdict-100417/>, April 2010.
- [15] BJ Fogg. *Persuasive Technology: Using Computers to Change What We Think and Do*. Morgan Kaufmann, 2002.
- [16] B.J. Fogg. Mass Interpersonal Persuasion: An Early View of a New Phenomenon. *Proc. Third International Conference on Persuasive Technology, Persuasive 2008*, 2008.
- [17] BJ Fogg. A behavior model for persuasive design. In *Persuasive '09: Proceedings of the 4th International Conference on Persuasive Technology*, pages 1–7, New York, NY, USA, 2009. ACM.
- [18] OpenStreetMap Foundation. About OpenStreetMap Foundation. <http://www.osmfoundation.org/wiki/OSMF:About>, May 2010.
- [19] Jim Giles. Internet Encyclopedias go head to head. *Nature*, 438:900–901, 2005.
- [20] Michael F. Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69:211–221, 2007.
- [21] Mordechai Haklay. How good is Volunteered Geographical Information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design*, 2008.
- [22] Mordechai Haklay and Patrick Weber. OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing*, 7:12–18, 2008.
- [23] J.E. Hirsch. An index to quantify an individuals scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102:16569–16572, 2005.
- [24] Pirate Party International. 22 Pirate Parties from all over the world officially founded the Pirate Parties International. <http://www.pp-international.net/node/471>, 04 2010.
- [25] Rilla Khaled, Pippin Barr, James Noble, and Robert Biddle. Investigating Social Software as Persuasive Technology. In *Persuasive Technology: First International Conference on Persuasive Technology for Human Well Being, PERSUASIVE 2006, Eindhoven, The Netherlands, May 2006, Proceedings*, 2006.
- [26] Aniket Kittur and Robert E. Kraut. Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *CSCW '08: Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 37–46, New York, NY, USA, 2008. ACM.

- [27] M. Meulpolder, L. D'Acunto, M. Capota, M. Wojciechowski, J.A. Pouwelse, D.H.J. Epema, and H.J. Sips. Public and private BitTorrent communities: A measurement study. In *IPTPS 10, 9th International Workshop on Peer-to-Peer Systems*, 2010.
- [28] Eben Moglen. Freedom in the Cloud: Software Freedom, Privacy and Security for Web 2.0 and Cloud Computing. <http://www.isoc-ny.org/?p=1338>, February 2010.
- [29] J.J.D. Mol, J.A. Pouwelse, D.H.J. Epema, and H.J. Sips. Free-riding, Fairness, and Firewalls in P2P File-Sharing. In *Eighth International Conference on Peer-to-Peer Computing*, 2008.
- [30] Nature. Encyclopaedia Britannica and Nature: a response. http://www.nature.com/press_releases/Britannica_response.pdf, March 2006.
- [31] J. Nielsen. Ten usability heuristics. <http://phillips.rmc.ca/courses/459-2007/lectures/03-heuristic-list.pdf>, 1994.
- [32] Jakob Nielsen. Participation Inequality: Encouraging more users to contribute. http://www.useit.com/alertbox/participation_inequality.html, April 2009.
- [33] Dagens Nyheter. The Pirate Bay sentenced to one year in prison. <http://www.dn.se/kultur-noje/musik/the-pirate-bay-sentenced-to-one-year-in-prison-1.846915>, 04 2009.
- [34] OpenStreetMap. Beginner's Guide. http://wiki.openstreetmap.org/wiki/Beginners%27_Guide, May 2010.
- [35] OpenStreetMap. OpenStreetMap - Fast Facts. pdf online, 1 2010. Published online.
- [36] Felipe Ortega. *Wikipedia: A quantitative analysis*. PhD thesis, Universidad Rey Juan Carlos, 2009.
- [37] Felipe Ortega and Kevin Crowston. Introduction to open movements: Floss, open content and open communities minitrack. In *Proceedings of the 42nd Hawaiian International Conference on System Sciences (HICSS 2009)*, 2009.
- [38] The Pirate Bay. OpenInternet. <http://thepiratebay.org/blog/174>, 09 2009.
- [39] The Pirate Bay. Riding out the storm. <http://thepiratebay.org/blog/172>, 08 2009.
- [40] The Pirate Bay. Legal threats against The Pirate Bay. <http://thepiratebay.org/legal>, 4 2010.
- [41] S. Rafaeli, T. Hayat, and Y. Ariel. *Cyberculture and New Media*, chapter Knowledge building and motivations in Wikipedia: Participation as Ba, pages 51–67. Rodopi Press, 2009.

- [42] Eric S. Raymond. The Cathedral and the Bazaar. <http://catb.org/esr/writings/homesteading/cathedral-bazaar/>, September 2002.
- [43] Riva Richmond. Digital Help for Haiti. <http://gadgetwise.blogs.nytimes.com/2010/01/27/digital-help-for-haiti/>, January 2010.
- [44] Andy Robinson. We're on the road to everywhere. <http://old.opengeodata.org/2008/07/index.html>, July 2008.
- [45] Kjeld Schmidt. *Cooperative Work and Coordinative Practices: Contributions to the Conceptual Foundations of Computer-Supported Cooperative Work*. PhD thesis, IT University of Copenhagen, 2007.
- [46] Helen Sharp, Yvonne Rogers, and Jenny Preece. *Interaction Design: Beyond Human-Computer Interaction*. John Wiley & Sons, 2007.
- [47] Clay Shirky. *Here Comes Everybody: How change happens when people come together*. Penguin Books, 2008.
- [48] Richard M. Stallman. *Free Software Free Society: Selected Essays of Richard M. Stallman*. Free Software Foundation, 2002.
- [49] Ilkka Tuomi. *Networks of Innovation.*. Oxford University Press, 2006.
- [50] TuxRadar. Benchmarked: Ubuntu vs Vista vs Windows 7. <http://www.tuxradar.com/node/33>, February 2009.
- [51] Eric Von Hippel. *Democratizing Innovation*. MIT Press, 2005.
- [52] Richard Weait. Today is the day: 250,000 contributors. <http://opengeodata.org/today-is-the-day-250000-contributors>, April 2010.
- [53] Steve Whittaker, Loren Terveen, Will Hill, and Lynn Cherny. The Dynamics of Mass Interaction. pages 257–264, 1998.
- [54] D.M. Wilkinson and B.A. Huberman. Assessing the Value of Cooperation in Wikipedia. *ArXiv Computer Science e-prints*, February 2007.

Appendix 1 - Screengrabs from Heuristic Walkthrough

OSM Walkthrough

Stage One - Signing Up

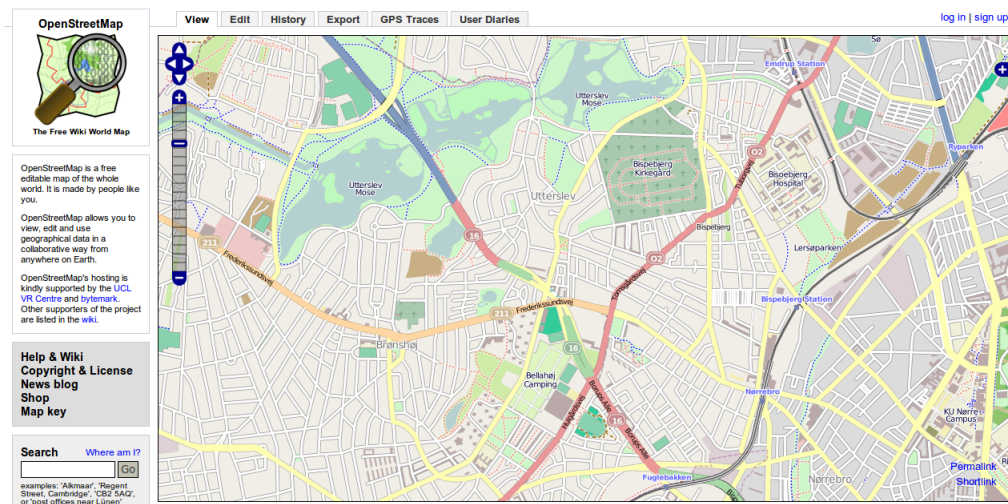


Figure 1: Prominent 'Edit' button and 'Sign Up' button take user to following page. Note though, small size of motivational material. Motivational material is also indirect, no direct encouragement to become involved.

Create a User Account

Fill in the form and we will send you a quick email to activate your account.

Email Address:

Confirm Email Address:

Not displayed publicly (see [privacy policy](#))


Display Name:

Your publicly displayed username. You can change this later in the preferences.

Password:

Confirm Password:

Figure 2: Create a user account screen



The Free Wiki World Map

[Help & Wiki](#)
[Copyright & License](#)
[News blog](#)
[Shop](#)

[Make a Donation](#)

[View](#) [Edit](#) [History](#) [Export](#) [GPS Traces](#) [User Diaries](#) [log in](#) | [sign up](#)

Contributor terms

Please read the agreement below and press the agree button to create your account.

Please select your country of residence: ☐ France ☐ Italy ☐ Rest of the world

Thank you for your interest in contributing data and/or any other content (collectively, 'Contents') to the geo-database of the OpenStreetMap project (the 'Project'). This contributor agreement (the 'Agreement') is made between you ('You') and The OpenStreetMap Foundation ('OSMF') and clarifies the intellectual property rights in any Contents that You choose to submit to the Project. Please read the following terms and conditions carefully and click either the 'Accept' or 'Decline' button at the bottom to continue.

1. You agree to only add Contents for which You are the copyright holder (to the extent the Contents include any copyrightable elements). You represent and warrant that You are legally entitled to grant the licence in Section 2 below and that such licence does not violate any law, breach any contract, or, to the best of Your knowledge, infringe any third party's rights. If You are not the copyright holder of the Contents, You represent and warrant that You have explicit permission from the rights holder to submit the Contents and grant the licence below.
2. Rights granted. Subject to Section 3 below, You hereby grant to OSMF a worldwide, royalty-free, non-exclusive, perpetual, irrevocable licence to do any act that is restricted by copyright over anything within the Contents, whether in the original medium or any other. These rights explicitly include commercial use, and do not exclude any field of endeavour. These rights include, without limitation, the right to sublicense the work through multiple tiers of sublicensees. To the extent allowable under applicable local laws and copyright conventions, You also waive and/or agree not to assert against OSMF or its licensees any moral rights that You may have in the

In addition to the above agreement, I consider my contributions to be in the Public Domain ☐ ([what's this?](#))

Figure 3: Terms and conditions

Stage Two - Getting involved

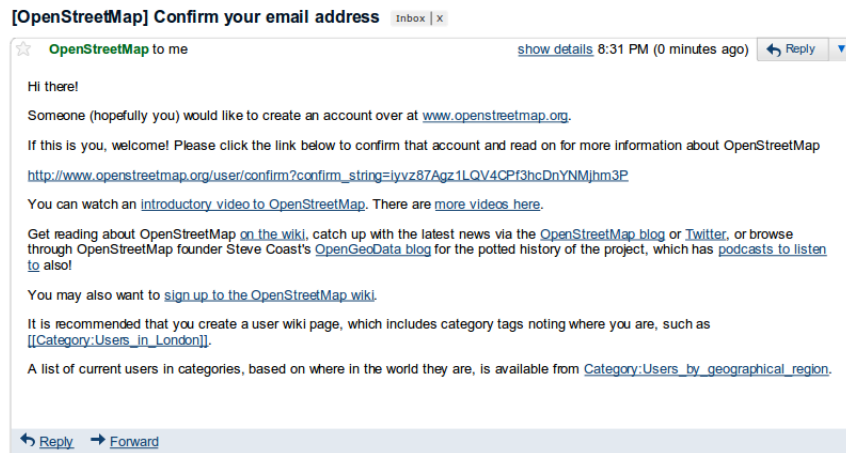



Figure 4: Once a prospective user has registered an account, they receive the above email. It is written in a light-hearted, conversational tone. It gives links to educational resources, documentation, blog, Twitter account, another blog and encourages users to sign up to the wiki based on where they live - thus making it easier for them to connect with other users in their area.



Figure 5: This is the first in a series of videos produced by an OSM community member which walks new users through the basics of editing Open Street Map.



[Page](#)
[Discussion](#)

[Read](#)
[View source](#)
[View history](#)

[Log in / create account](#)

Beginners' guide

(Redirected from [Beginners' Guide](#))

Available languages	Help
Български • Català • Česky • Dansk • Deutsch • Ελληνικά • English • Español • Eesti • Euskara • Suomi • Français • עברית • Hrvatski • Magyar • Interlingua • Italiano • 日本語 • 한국어 • Lietuvių • Latviešu • Nederlands • Norsk (bokmål) • Polski • Português do Brasil • Română (Moldova) • Русский • Slovenčina • Svenska • Українська • Tiếng Việt • 中文(简体)	
Missing languages	show

OpenStreetMap (OSM) follows a similar concept as [Wikipedia](#) does, but for maps and other geographic facts (despite its name, it's by no means only limited to streets and roads). People, like you and me, gather location data across the globe from a variety of sources such as recordings from [GPS](#) devices, from free satellite imagery or simply from knowing an area very well, for example because they live there. This information then gets uploaded to OpenStreetMap's central database from where it can be further modified, corrected and enriched by anyone who notices missing facts or errors about the area.

Anyone can freely [download](#) and use the full information for any purpose they like under an [open source license](#). Currently the most prominent use of OSM data is to render beautiful, rich maps such as the examples you can find on [www.openstreetmap.org](#), but there are plentiful other [applications](#) too.

The project was started because most maps you think of as free actually have legal or technical restrictions on their use, holding back people from using them in creative, productive, or unexpected ways. To foster these creative and unexpected uses, OpenStreetMap also does not limit the type of information people can add into the database, as long as it is factually correct, verifiable and does not infringe on anyone else's copyright. You never know for what interesting purposes it can be used for in future...

This Beginner's Guide covers the process of adding data to OpenStreetMap. To get an overview of the site's features and the

Contents:

- [Beginners' Guide](#)
- [1.1 - Gather data](#)
- [1.1.1 - GPS](#)
- [1.1.2 - YI/landsat/NPE Imagery](#)
- [1.2 - Upload data](#)
- [1.2.1 - Save your files to GPX](#)
- [1.2.2 - Uploading Data](#)
- [1.3 - Edit maps](#)
- [1.3.1 - General tips](#)
- [1.3.2 - Potlatch](#)
- [1.3.3 - JOSM](#)
- [1.3.3.1 - Downloading Into JOSM](#)
- [1.3.3.2 - First basic road](#)
- [1.3.4 - Merkaartor](#)
- [1.4 - Edit data](#)

Figure 6: The Wiki is the main documentation center for the OSM project. It includes an extensive User Guide and serves as a platform for co-ordination between users.

[Recent changes](#)

[Recent changes](#)

[Navigation](#)

[Main Page](#)
[The map](#)
[Mapping projects](#)
[Map Features](#)
[Help](#)
[Blog](#)
[Shop](#)
[Donations](#)

[Toolbox](#)
[What links here](#)
[Related changes](#)
[Special pages](#)
[Printable version](#)
[Permanent link](#)

Beginners Guide 1.1

Available languages	Help
Български • Català • Česky • Deutsch • Ελληνικά • English • Español • Eesti • Euskara • Suomi • Français • Hrvatski • Magyar • Italiano • 日本語 • 한국어 • Nederlands • Polski • Русский • Slovenčina • Українська • 中文(简体)	
Missing languages	show

Collecting Data

There are a variety of forms of gathering data for OSM:

- GPS** - This is currently the most common way of gathering data for OSM, and often preferred or even essential for collecting the initial geometry of roads, paths and other ways. If you want to add to an otherwise blank area and don't have a GPS, they are sometimes available for loan or you can create maps from other users' data.
- Local knowledge** - Perhaps the best source of data though is if you simply happen to know the area very well and thus for example the names of the roads, which shops or buildings there are, the local traffic rules and restrictions, or what ever else you might want to add to OpenStreetMap. If the basic road layout is already present, you often don't need any other technical devices and can start straight away.
- Yahoo! Imagery, Landsat and NPE maps** are available to OSM for extracting data from. Accuracy is important though, so only map places you've been. These sources add greatly to OSM, but data can't be built on these alone.
- Your own photography or maps** - Make sure these are completely free to copy and use with OSM. Most data is not as free as you would think.
- Data may already be collected** that requires people to convert it into a map. An example of this is the [collaborative mapping project for Korea](#).

What data to add

There is a lot of data that can be gathered and put into OSM: From common things such as 'street names' right down to fine details which includes things like parks, postboxes, hedgerows and cairns. Different people find different things more important, usually influenced by their main method of transport. If you feel it helps people find their way, then map it! A nice selection of some of the most commonly mapped features and the way to describe them in osm can be found on the [map features](#) page. But remember you aren't limited to what is listed there, so be creative and map what is important to you!

Commonly manners increase the detail as they go on, but start with the basics. So the main road network tends to be a good starting point, and lesser things such as

Contents:

- [Beginners' Guide](#)
- [1.1 - Gather data](#)
- [1.1.1 - GPS](#)
- [1.1.2 - YI/landsat/NPE Imagery](#)
- [1.2 - Upload data](#)
- [1.2.1 - Save your files to GPX](#)
- [1.2.2 - Uploading Data](#)
- [1.3 - Edit maps](#)
- [1.3.1 - General tips](#)
- [1.3.2 - Potlatch](#)
- [1.3.3 - JOSM](#)
- [1.3.3.1 - Downloading Into JOSM](#)
- [1.3.3.2 - First basic road](#)
- [1.3.4 - Merkaartor](#)
- [1.4 - Edit data](#)
- [1.4.1 - Adding Tags](#)
- [1.4.2 - Uploading changes](#)
- [1.5 - Render maps](#)
- [1.5.1 - Osmarender](#)

Figure 7: Collecting data by GPS is presented as the de facto standard. The cost of these devices likely acts as a large barrier to participation.

Stage Three - The Participant Community

August 8, 2010

Project of the Week: Monsoon Flooding in Pakistan



Monsoon flooding continues in Pakistan following what has been called the heaviest rains in 80 years. As many as 12 million people have already been affected and over 1600 are known dead to date. The availability of up to date aerial imagery has been hampered by the continuous cloud cover. New flood warnings are announced often, and the rains are continuing. Lists of dead and missing, additional flooding, displacements and injury seem overwhelming. It is expected that disease will become a very serious issue as access to clean water is reduced.

http://www.thenews.com.pk/top_story_detail.asp?id=30580

<http://www.csmonitor.com/World/Asia-South-Central/2010/0806/Pakistan-floods-d...>

<http://www.bbc.co.uk/news/world-south-asia-1088>

<http://www.earthtimes.org/articles/news/338523.relief-efforts-summary.html#9925>

OpenStreetMap has some data in some of the affected areas. You might help by donating money to an international relief agency. And you

Figure 8: The OpenGeoData Blog is a frequently updated blog connected to the OSM community. Its posts are aimed towards community members and consistently re-state value based norms for participation as well as encouraging users to take part in specific 'Projects of the Week'. The blog frequently connects OSM mapping activity to humanitarian causes (see the example above) which helps to encourage users to feel that their contributions are valuable.

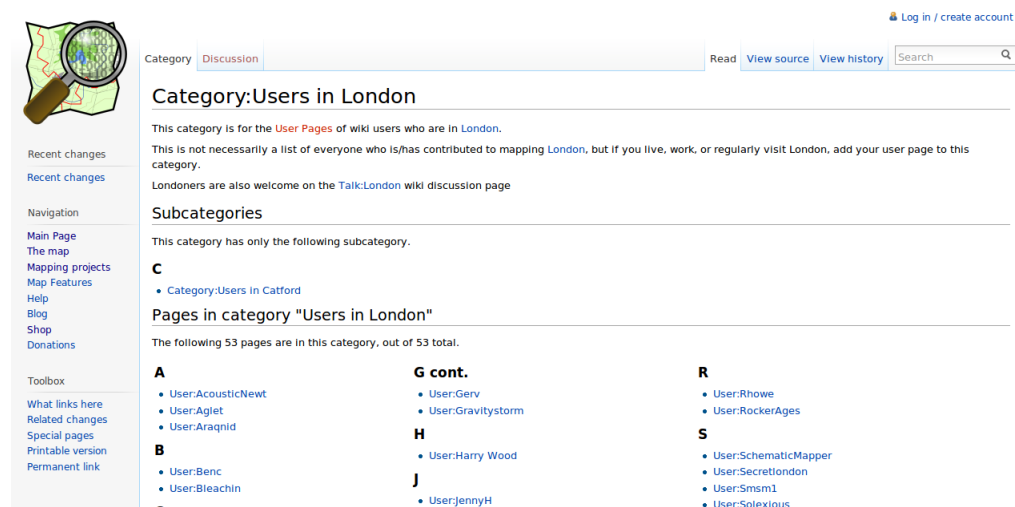


Figure 9: The 'Users In' category acts to connect OSM users with each other, see the link to the 'Talk:London' discussion page. It also gives an opportunity to visit the profiles of other OSM users. These profiles are distinct to the 'User Histories' on the main web page and are generated by the user themselves.

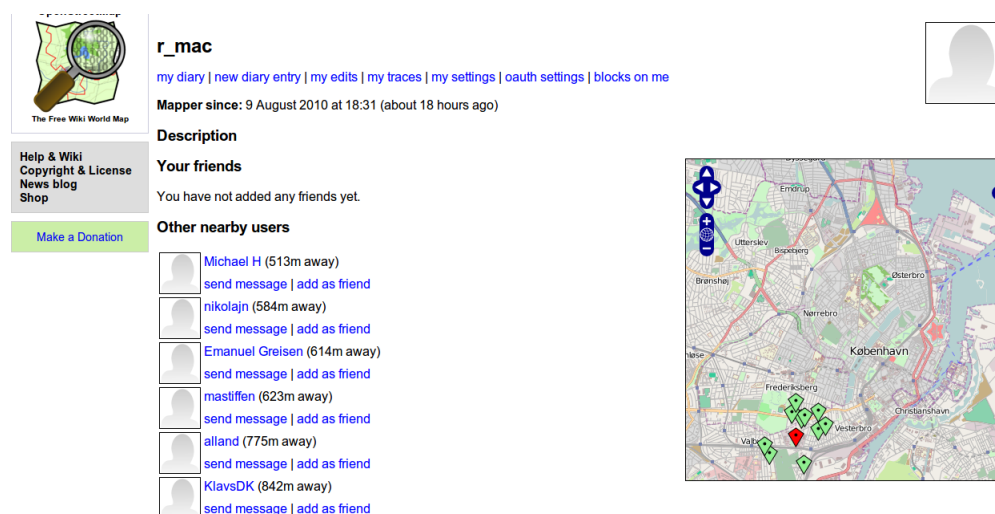


Figure 10: The user interface suggests that users contact other OSM users living nearby, based on the location given in the registration process. However, this list of users is populated based on those closest to the user and includes inactive users. There is no direct link at this stage to the discussion lists used by Danish users which would be of more use in connecting new users to active participants.

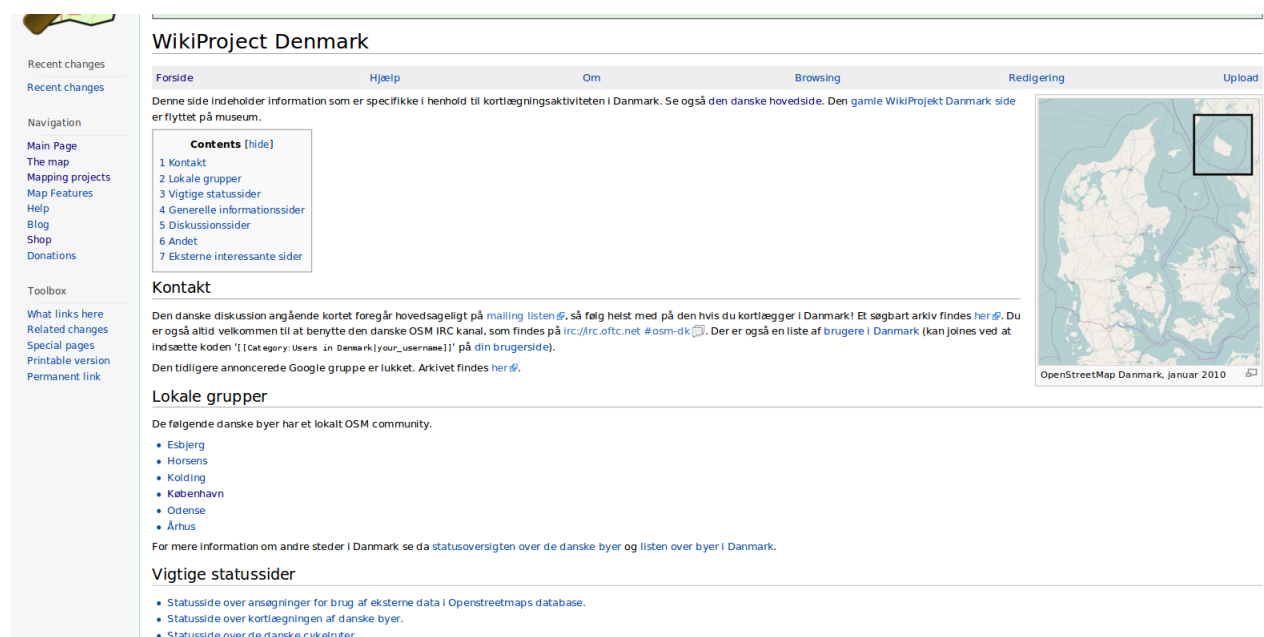


Figure 11: This pages gives information about local Open Street Map groups and links to the main co-ordination pages.

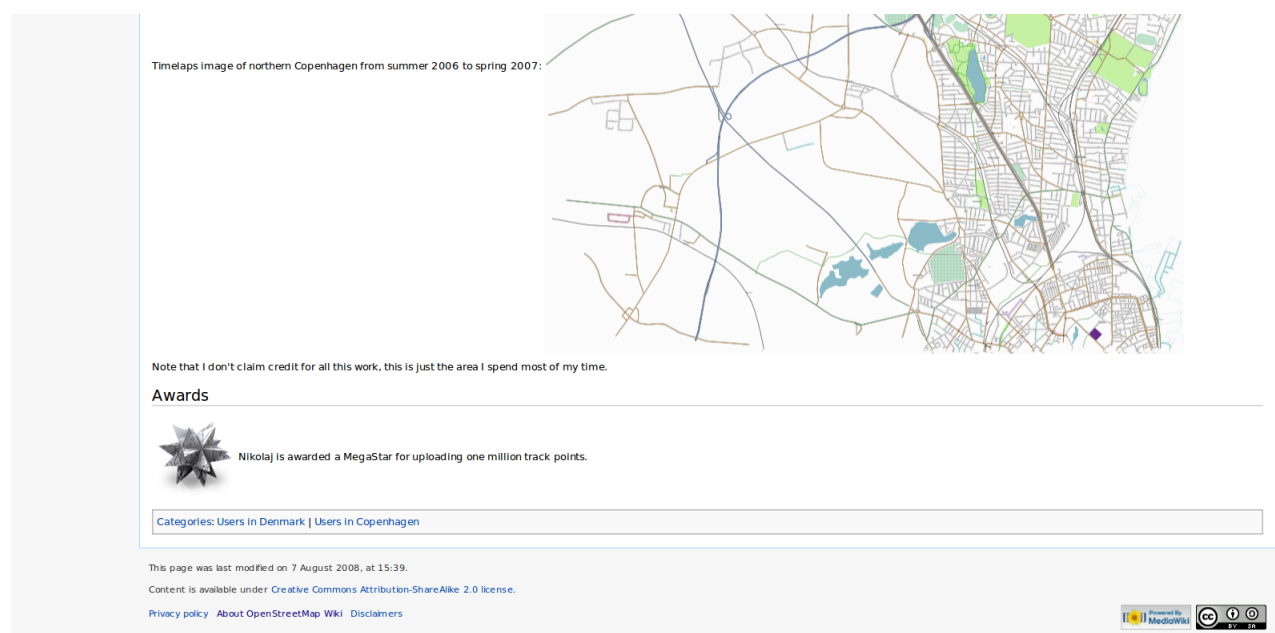


Figure 12: Users get different types of stars based on number of contributions. However, this is on the User page and is not otherwise visible to other users, limiting its ability to influence their behaviour.

TPB Walkthrough

Stage One - Signing Up

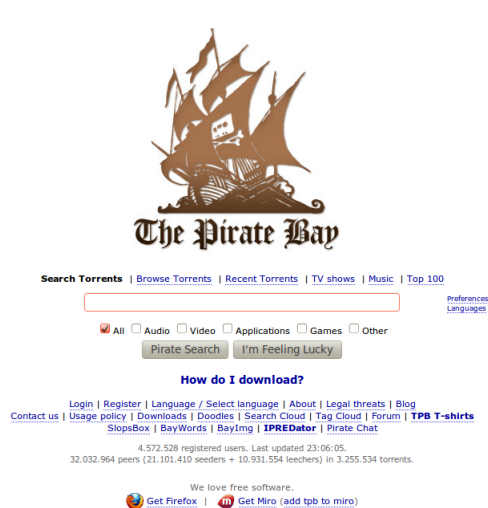


Figure 13: This is the main page. Its primary function is as a search engine. There is also a prominent link to instructions for downloading. The 'login' and 'register' links are all quite far down on the page, as are the Forum and the Blog. The Image can be seen as a tool to boost identity with the site, while the 'We love free software' text identifies TPB with free software initiatives. The 'How do I download?' link is meant to increase the ability of users to download, while the statistics on participants makes the user feel less isolated and encourages them to participate.

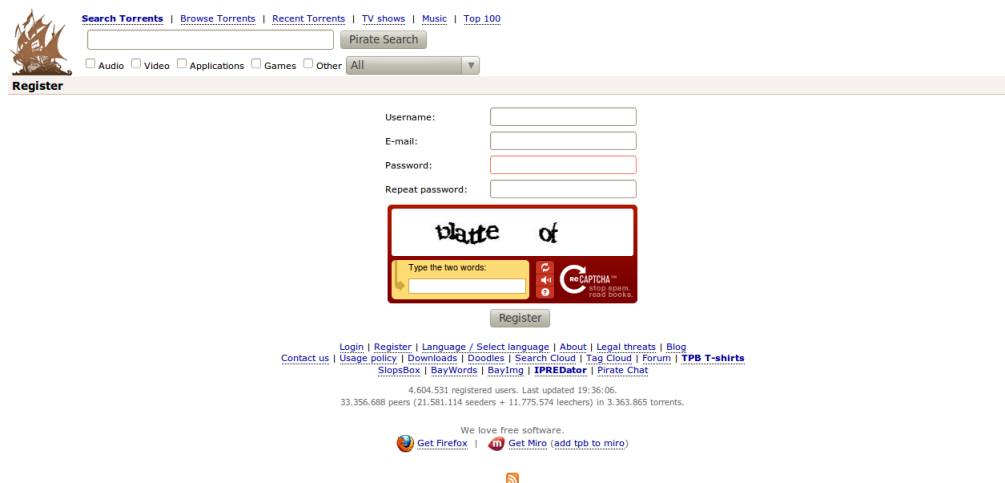


Figure 14: This is the registration page.

Stage Two - Getting Involved

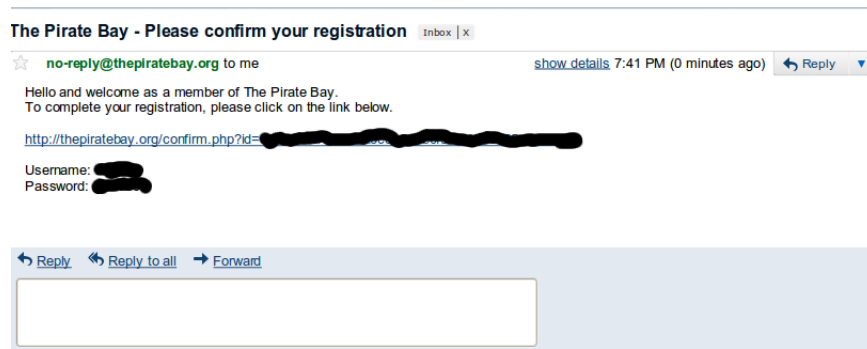


Figure 15: This confirmation email makes no attempt to encourage users to participate nor does it encourage them to join the forums or read the blog.

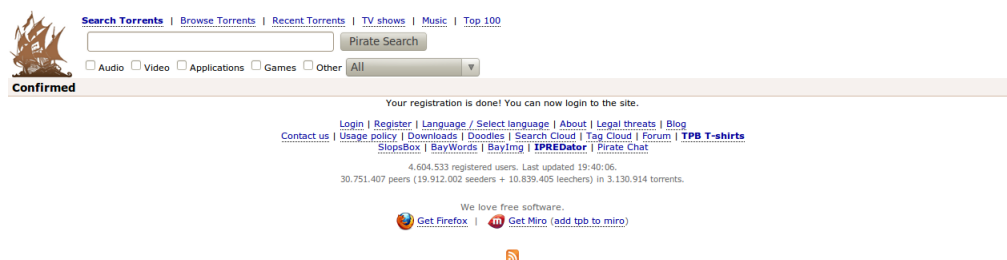


Figure 16: This is the screen that comes up when one clicks on the confirmation link in the window.

You have no active torrents.

Username: [REDACTED]
Current IP: [REDACTED]
Current language: English
Your public URL: [http://thepiratebay.org/user/\[REDACTED\]](http://thepiratebay.org/user/[REDACTED])
Your RSS feed: [http://rss.thepiratebay.org/user/\[REDACTED\]](http://rss.thepiratebay.org/user/[REDACTED])
Status: MEMBER
Show pornographic torrents: Yes ☐ No ☒
Default sort order: Relevance ▼

Old password: [REDACTED]
New password: [REDACTED]
Repeat password: [REDACTED]

[Account removal](#)

You have uploaded 0 torrent(s).
 You have written 0 comment(s).

Figure 17: This is the user preference panel.

Thread / Author	Replies	Views	Rating	Last Post (asc)
Forum Announcements				
Forum Rules (Last updated Aug 13 2009) Illuminated	-	-		13th August 2009 18:16
Important Threads				
Tutorial Thread Index Elin	1	65,730	★★★★★	24th August 2008 00:57 Last Post: Elin
FAQs *****Beginners, Start Here***** Adapa	1	74,924	★★★★★	11th August 2008 08:48 Last Post: Adapa
Normal Threads				
how to rip from youtube with nothing more than IE b.s.o.d	3	1,591	★★★★★	Yesterday 04:46 Last Post: Adapa
Tutorial: Block ads, malware, music, pop-ups and redirections on TPB Adapa	3	4,142	★★★★★	31st July 2010 02:42 Last Post: System Folder
[Tut] Show photo on your torrent page (aka Torrent Cover) Ekanh22	0	411	★★★★★	10th July 2010 07:14 Last Post: Ekanh22
[TUT] How to make a bittorrent use only a VPN DartheeroProductions	0	504	★★★★★	4th July 2010 00:37 Last Post: DartheeroProductions
-Timesaver- Bulk Movie thumbnail/screenshot creator (Drag & Drop) by Ekanh Ekanh22	0	463	★★★★★	2nd July 2010 13:24 Last Post: Ekanh22
Create and upload a torrent on TPB using BitComet room101belboy	0	2,624	★★★★★	18th March 2010 11:12 Last Post: room101belboy
Creating Torrents 202 - Minimize Seeding / Maximize Lifespan Adapa	2	8,436	★★★★★	18th March 2010 04:41 Last Post: Adapa
Create and upload a torrent on TPB using Azureus/Vuze room101belboy	0	1,839	★★★★★	13th March 2010 22:16 Last Post: room101belboy

Figure 18: This is the Tutorial Forum

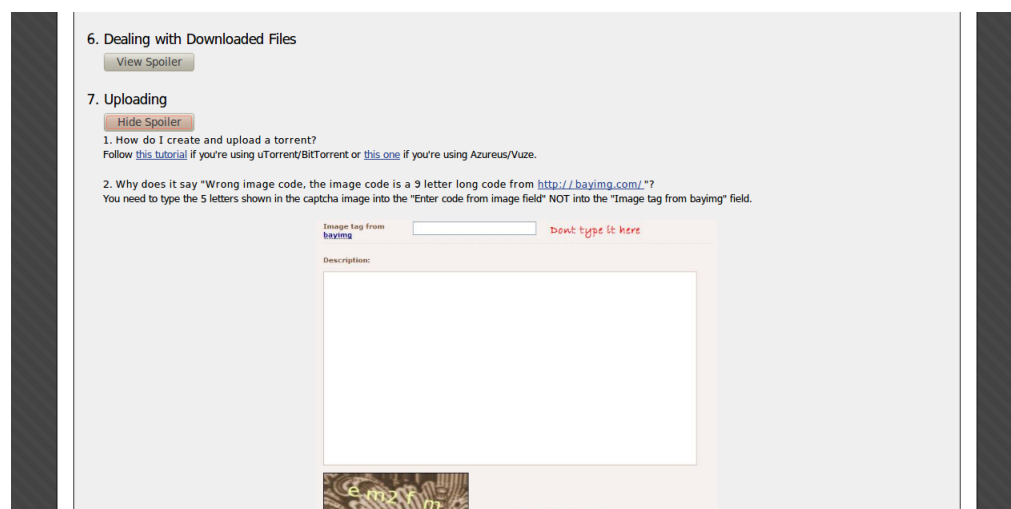


Figure 19: How to upload a torrent.

Stage Three - The Participant Community

PLZ LEARN: TPB CANT BE SHUT DOWN

LOL!

AS U MITE HAS READ OR NOTICD, PEEPS ONCE AGAIN R TRYIN 2 SHUT US DOWN. DIS WILL NOT SUCCED, LOL. OURS RLY NICE WEBHOST WUZ THREATEND WIF RLY HUGE FINE, SO WE DECIDD 2 MOOV TEH SIET SO DAT THEY DIDNT GOT INTO TROUBLE, LOL. TEH DECISHUN 2 MOOV WUZ TAKEN BY US, TEH PIRATE BAY, LOL.

TEH PIRATE BAY IZ AN UNSINKABLE SHIP. IT WILL SAIL TEH INTERWEBS 4 AS LONG AS WE WANTS IT 2. REMEMBR DAT, K THX.

TPB, ONLY IN IT 4 TEH LULZ SINCE 2003

 [851 comment\(s\)](#) | Posted 05-18 12:18 by Dr LOLCAT

More lies in the media

With great interested we read the press release at [Market Watch](#) that a company decided to buy TPB.

However, we have no deal with them. We have not even talked to them! Since this is quite heavy fraud and we don't want our users to buy shares or anything like that in a company that is claiming to work with us, we just wanted to point out that this is a lie.

What is not a lie is the problem with the digital bill in the United Kingdom. Please read more about it (you'll find a link on the frontpage!)

 [31 comment\(s\)](#) | Posted 04-28 05:49 by tpb

Merry X-mas 2009!

Figure 20: The TPB blog is not frequently updated. Its tone is also quite varied, ranging from well considered and reflective to juvenile. Its content is usually articles about the site but includes some references to digital liberty and anti-censorship activism.

```

> <http://thepiratebay.org/torrents-detail.php?r10=5630235&hl=1> &hl=1
> Enya - The Celts - 386 transfers
>
> As these albums are under Copyright and no EU or Swedish law allows
> unauthorized distribution of this ripped music, you are in violation of our
> Copyright.
>
> You are additionally in violation of Swedish and EU law as you are violating
> our Trade Mark by listing Trade Mark protected names (protected under
> Swedish law) on your web page without permission.
>
> You have 72 hours to completely remove the above links.
>
> After this you will receive 1 (ONE) legal note from GrayZone. You will
> receive no further notices.
>
> Also please note that making this email public or ridiculing it will result
> in immediate legal action and we are also contacting RIPE NCC for suspension.
>
> Regards
> Peter Pehrson
> aigle music / warner music international

Dear whatever-you-are,
thank you for providing us and our users with such great entertainment.
I'm not talking about Enya (hey, Enya fucking sucks), but instead of your
nonsensical email.

You have
- confused us with our ISP
- no knowledge whatsoever about BitTorrent
- no knowledge whatsoever of the applicable laws (trademark or copyright)
- made very entertaining threats (hey, go ahead and contact RIPE NCC,
please, I beg you)

You have scored 10 out of 10 points on our Legal Threats Entertainment
scale. You win the grand prize: A lifetime of ridicule on our legal
threats section (http://static.thepiratebay.org/legal/) !
Congratulations!

Please also note that I'm not currently out of toilet paper, so you may
wait a while before sending legal papers.

> =====

```

Figure 21: The infamous 'Legal Threats' page contains correspondence between Pirate Bay administrators and lawyers for content providers requesting that content be taken down. The Pirate Bay administrators refuse these requests and insult the sender. One can argue that this section serves a motivational purpose in developing the enmity between pirates and media corporations. It reassures would-be pirates that The Pirate Bay has no intention of complying with media corporations.

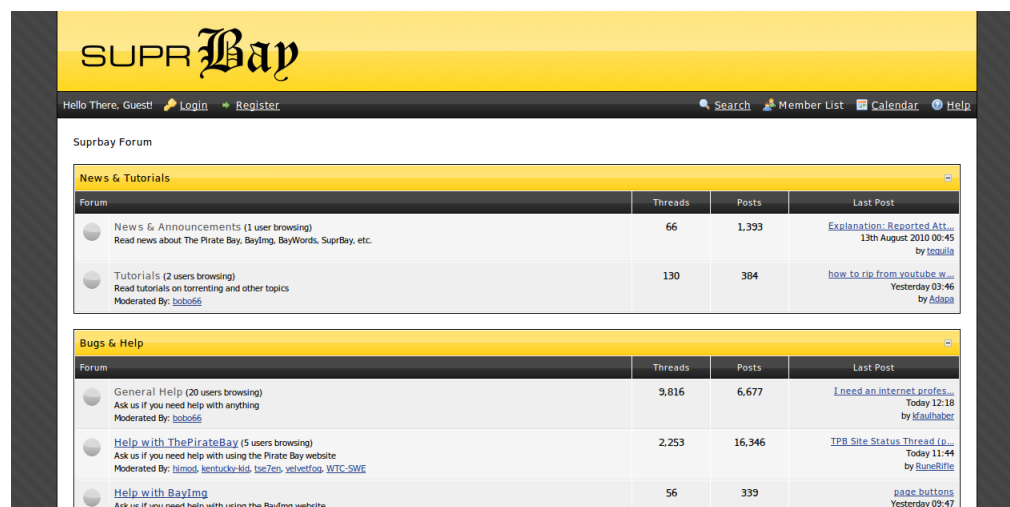


Figure 22: This is the main view of the user forum. The user forum is connected to the site. There is only one main forum rather than the many different ones in OSM. Tutorials and Help forums are very prominent, these act to increase user ability to upload torrents and pirate content.

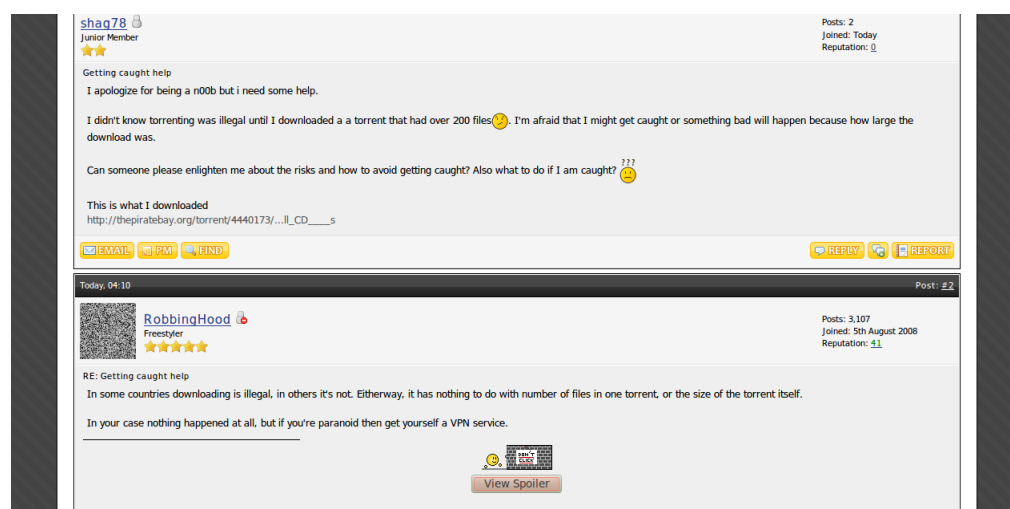


Figure 23: Tone on the forums is generally quite good; users help each other with technical problems related to torrenting, but also discuss computer games, movies, music and literature.

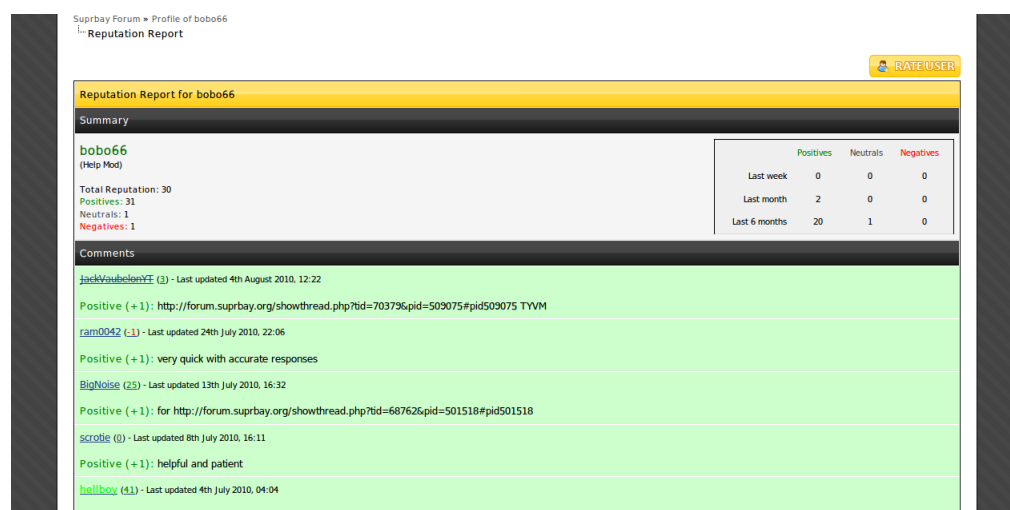


Figure 24: The forum has a rating system, whereby other users rate a user based on their posts on the forum. This is a form of social control which may help to make good forum behaviour more attractive.

Appendix 2 - Scripts used in data retrieval and analysis

This appendix contains all the scripts used for retrieving and processing data in this study. The script for retrieving and processing data from The Pirate Bay was written by Dr. Toine Bogers. I wrote the script that processed the Open Street Map HTML pages which was used in conjunction with Dr. Bogers' retrieval script. The script used to calculate average comments per user was designed by user "Patsie" after a request for help was posted on <http://programmingforums.org>.

The Pirate Bay

```
#!/usr/bin/python
# Import necessary libraries.
import sys, os, random, time, datetime, urllib2, unicodedata
from BeautifulSoup import BeautifulSoup
def getURL(user, page_no):
    return url_prefix + user + "/" + str(page_no) + url_suffix
def normalizeString(unicode_string):
    return unicodedata.normalize('NFKD', unicode_string).encode('ascii', 'ignore')
def formatDate(raw_date_string):
    # If there is an actual date and not "X minutes ago".
    if raw_date_string.find("ago") == -1:
        date_array = raw_date_string.split("&nbsp;")
        month_day = date_array[0]
        # If the second element is a timestamp, then we're talking about
the
        current year; add that instead.
        if date_array[1].find(":") == -1:
            year = date_array[1]
        else:
            year = time.strftime("%Y")
        # If the string contains "Today", just print today's date.
        if month_day == "Today":
            date_string = time.strftime("%d-%m-%Y")
        # If the string contains "Yesterday", just print yesterday's date.
        elif month_day == "Y-day":
```

```

today_object = datetime.datetime(int(time.strftime("%Y")),
int(time.strftime("%m")),
int(time.strftime("%d")),
0, 0, 0)
day_difference_object = datetime.timedelta(days = 1)
yesterday_object = today_object - day_difference_object
date_string = "%02d-%02d-%s" % (yesterday_object.day,
yesterday_object.month,
yesterday_object.year)
# Else convert the date to the same format.
else:
date_string = month_day.split("-")[1] + "-" + month_day.split("-")[0]
+ "-" + year
# If the string contains "X minutes ago", just print today's date.
else:
date_string = time.strftime("%d-%m-%Y")
return date_string
# Initialize variables.
url_prefix = "http://thepiratebay.org/user/"
url_suffix = "/3"
user_list, user_urls = {}, {}
max_pages = 1000
max_short_pause, max_long_pause = 2.0, 10.0
# max_short_pause, max_long_pause = 5.0, 150.0
micro_break_count, max_micro_breaks = 0, 30
# Quick and dirty parse of the command line options.
dataset, pair_list = None, None
no_of_arguments = len(sys.argv) - 1
if no_of_arguments == 1:
user_list_file = sys.argv[1]
else:
print "USAGE: ./traverse-user-pages.py <LIST OF USER NAMES>"
print ""
print "Note that this script only collects URLs associated with
one or more user names;"
print "it does not extract any other information or download the
pages."
print ""
print " <LIST OF USER NAMES> File containing a list of user names
for which we want"
print " to collect all associated URLs."
print ""
sys.exit()
# Open the user list file and store it in memory.
f = open(user_list_file, "r")
for line in f:
if line[0:1] != "#":
user = line.strip()
if user not in user_list:
user_list[user] = 0

```



```
f_crawl.close()
```

Processing average comments per user

```
#!/bin/sh
awk -F, '{
actions[$1]++;
comments[$1] += $3;
} END {
for (name in actions)
printf("%s,%d,%.02f\n", name, actions[name], comments[name]/actions[name]);
}'
```

Open Street Map

Retrieval Script

```
#!/usr/bin/python
# Import necessary libraries.
import sys, os, random, time, urllib2, unicodedata from BeautifulSoup
import BeautifulSoup
def getURL(user, page_no): return url_prefix + user + url_infix
+ str(page_no) def normalizeString(unicode_string): return unicodedata.normalize('NFKD',
unicode_string).encode('ascii','ignore')
# Initialize variables.
url_prefix = "http://www.openstreetmap.org/user/" url_infix = "/edits?page="
user_list, user_urls = {}, {} max_pages = 1000 max_short_pause, max_long_pause
= 2.0, 10.0 # max_short_pause, max_long_pause = 5.0, 150.0 micro_break_count,
max_micro_breaks = 0, 30
# Quick and dirty parse of the command line options.
dataset, pair_list = None, None no_of_arguments = len(sys.argv)
- 1 if no_of_arguments == 1: user_list_file = sys.argv[1] else: print
"USAGE: ./traverse-user-pages.py <LIST OF USER NAMES>" print "" print
"Note that this script only collects URLs associated with one or more
user names;" print "it does not extract any other information or download
the pages." print "" print " <LIST OF USER NAMES> File containing a
list of user names for which we want" print " to collect all associated
URLs." print "" sys.exit()
# Open the user list file and store it in memory.
f = open(user_list_file, "r") for line in f: if line[0:1] != "#":
user = line.strip() if user not in user_list: user_list[user] = 0
user_urls[user] = [] f.close()
# Loop through the users. For each user, traverse all of his pages
and subpages and store the URLs.
f_urls = open(user_list_file + ".urls", 'w') for user in user_list:
status = user_list[user] if status == 0:
# As long as the page contains search results, keep looking for
the next page
# as well.
```

```

        no_results_flag = False
        page_no = 1
        while no_results_flag == False and page_no < max_pages:
            # Take a random pause to prevent IP blocking by the server.
            if micro_break_count == max_micro_breaks:
                # Take a longer, random pause time.
                sleep(random.uniform(0.0, max_long_pause))
            micro_break_count = 0
        else:
            # Take a short, random pause time.
            sleep(random.uniform(0.0, max_short_pause))
            micro_break_count += 1

        # Download the current page.
        # print "Looking at page", page_no, "for user", user
        url = getURL(urllib2.quote(user), page_no)
        html_page = urllib2.urlopen(url)
        soup = BeautifulSoup(html_page)
        # Check whether the current page contains any search results.
        table_element = soup.find('table', {'id': 'changeset_list'})
        if len(table_element.findAll('tr')) == 0:
            no_results_flag = True
        # If the page does contain search results, process it.
        if no_results_flag == False:
            # Store the URL in our list.
            user_urls[user].append(url)
            print user, "\t", url
            f_urls.write(user + "\t" + url + "\n")
            # Update the page counter.
            page_no += 1
            # Close the URL file.
            f_urls.close()

```

HTML converter

```

#!/bin/sh
#Get change ID
cat $1 | grep "View changeset details" -A 2 | awk 'BEGIN{FS=">"}{print $2}' |
sed 's/<\/a//g' | grep "." > temp1.txt
#Get Date
cat $1 | grep -A 1 "table[01] date" | sed 's/<td class="table[01] date">//g'
| sed 's/--//g' | sed 's/ (still editing)/May 1, 2010 12:00/g' |
grep ":" |
sed 's/January/1/g' | sed 's/February/2/g' | sed 's/March/3/g' |
sed 's/April/4/g'
| sed 's/May/5/g' | sed 's/June/6/g' | sed 's/July/7/g' | sed 's/August/8/g'
|
sed 's/September/9/g' | sed 's/October/10/g' | sed 's/November/11/g'
|
sed 's/December/12/g' | sed 's/,//g' | awk 'BEGIN{FS=" "}{print $2 "-" $1 "-" $3}'
> temp2.txt
#Get Comment

```

```

    cat $1 | sed 's/#/hash1hash2/g' | grep -A 2 "table[01] comment"
| tr "\n" " " | sed 's/--/#/g' | tr "#" "\n" | sed 's/*.comment"> */g'
| sed 's/hash1hash2/#/g' > temp3.txt
    #Get Area Box
    cat $1 | grep -A 3 "show area box" | sed 's/*.box//g' | sed 's/<!--.*-->//g'
| sed 's/<\a>//g' | sed 's/--//g' | tr "\n " "&" | sed 's/#####/%/g'
| tr "'>" "\n" | tr "%&&" " " | grep "." > temp4.txt
    paste temp1.txt temp2.txt temp3.txt temp4.txt > osm-data-full.txt
    rm temp*.txt

```

Appendix 3 - Additional Tables and Graphs

Tables

	Low-level OSM	Mid-level OSM	High-level OSM
Avg. no of contributions	155.85	1840.68	18,130
Median contributions	124	1002	19,859
% of total contributions	3.83%	22.75%	73.40%

Table 1: Summary of OSM productivity bin details

	Low-level TPB	Mid-level TPB	High-level TPB
Avg. no of contributions	1.73	79.40	1,520.88
Median contributions	3	58	517
% of total contributions	1.67%	13.23%	85.08%

Table 2: Summary of TPB productivity bin details

Drop off rates

Number of contributions	OpenStreetMap	The Pirate Bay
1-10	4.84%	50.76%
11-20	3.01%	10.75%
21-30	2.62%	5.67%
31-40	3.27%	3.40%
41-50	1.57%	2.60%
≤ 50	15.33%	73.21%

Table 3: % of users contributing between 1 and 50 times (intervals of 10)

Lifetime participation rates - OSM

	Low-level contributors	Medium level contributors
Day 1	1.62	1.39
Day 2	1.05	0.15
Day 3	0.80	0.09
Day 4	0.61	0.07
Day 5	0.68	0.05
Day 6	0.53	1.43
Day 7	0.56	4.67
Day 8	0.90	2.51
Day 9	0.56	4.03
Day 10	0.59	13.94
Day 11	0.59	4.19
Day 12	0.53	0.11
Day 13	0.50	0.07
Day 14	0.42	0.06
\sum	10	32.81

Table 4: % of contributions by low-level and mid-level users over the first two weeks of their lifetimes

	Low-level contributors	Medium level contributors
Week 1	5.89	7.87
Week 2	4.10	24.93
Week 3	3.20	0.53
Week 4	2.81	0.39
Week 5	2.25	0.41
Week 6	2.08	0.61
Week 7	1.97	0.59
Week 8	1.72	0.67
Week 9	1.45	0.65
Week 10	1.48	0.58
Week 11	1.18	0.48
Week 12	1.36	0.33
\sum	29.53	38.08

Table 5: % of contributions by low-level and mid-level users over the first three months of their lifetimes

Days	Low-level contributors	Medium level contributors
30	16.69	33.85
60	8.14	2.56
90	5.71	2.02
120	5.25	3.09
150	4.63	2.75
180	4.57	2.14
210	3.74	1.87
240	4.00	1.58
270	3.53	1.91
300	3.15	1.89
330	2.84	2.02
360	3.69	2.20
390	3.51	2.65
420	3.42	2.49
450	3.44	2.24
480	2.93	2.89
510	2.65	2.78
540	2.12	2.69
570	1.67	2.16
600	1.42	2.47
630	1.64	2.47
660	0.90	2.27
690	1.24	1.84
720	1.53	1.49
Σ	92.53	86.46

Table 6: % of contributions by low-level and mid-level OSM users over the first two years of their lifetimes

Lifetime participation rates - TPB

Days	Low-level contributors	Medium level contributors	High level contributors
1	32.05	3.86	1.04
2	6.09	1.93	0.63
3	2.17	1.00	0.66
4	1.65	0.86	0.16
5	1.80	0.72	0.59
6	1.18	0.68	0.16
7	0.92	0.73	0.18
8	0.92	0.67	0.19
9	0.74	0.61	0.57
10	0.55	0.54	0.18
11	0.70	0.48	4.52
12	0.35	0.55	0.17
13	0.68	0.51	0.60
14	0.48	0.44	0.57
Σ	50.34	13.62	10.26

Table 7: % of contributions by low-level and mid-level users over the first two weeks of their lifetimes

Days	Low-level contributors	Medium level contributors	High level contributors
30	56.79	20.40	14.63
60	7.19	10.27	5.38
90	5.48	7.91	7.15
120	3.10	6.96	3.95
150	3.43	5.77	4.08
180	2.62	4.88	3.70
210	2.15	3.47	3.89
240	2.02	3.60	3.05
270	1.80	3.19	3.08
300	1.49	2.93	2.42
330	1.14	2.80	2.13
360	0.88	2.36	2.11
390	1.18	2.38	2.67
420	1.03	2.17	2.50
450	0.59	2.33	2.44
480	0.81	1.74	2.13
510	0.59	1.60	1.99
540	0.57	1.32	1.78
570	0.46	1.16	1.65
600	0.48	1.11	1.54
630	0.48	1.18	1.56
660	0.57	1.39	1.39
690	0.48	0.96	1.37
720	0.22	0.91	1.26
Σ	95.64	92.87	77.96

Table 8: % of contributions by low-level, mid-level and high-level TPB users over the first two years of their lifetimes

Graphs

Contribution levels - OSM vs TPB

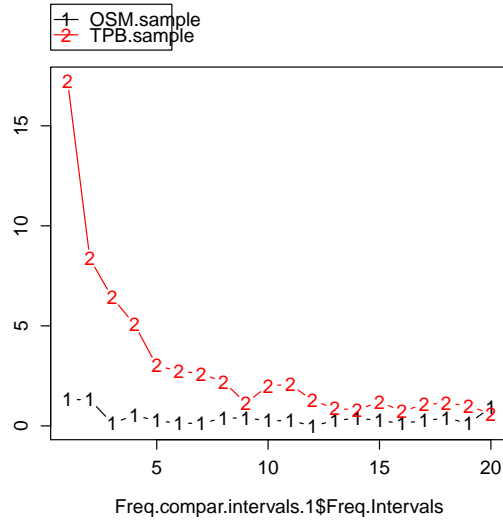


Figure 1: Contribution rates OSM vs TPB, Intervals of 1

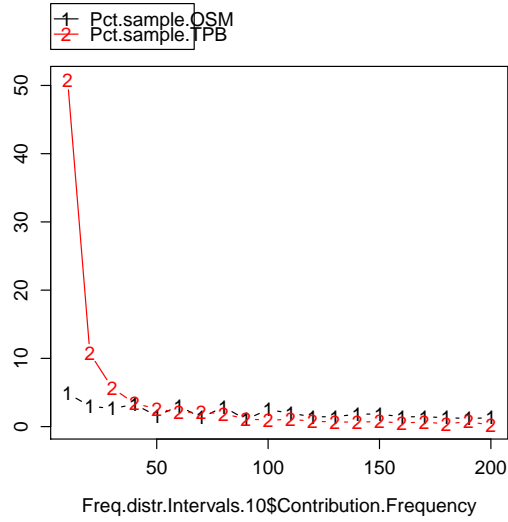


Figure 2: Contribution rates OSM vs TPB, Intervals of 10

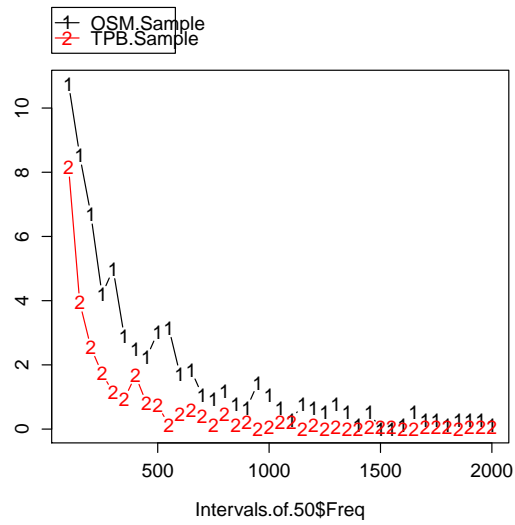


Figure 3: Contribution rates OSM vs TPB, Intervals of 50

Lifetime Contribution rates - OSM

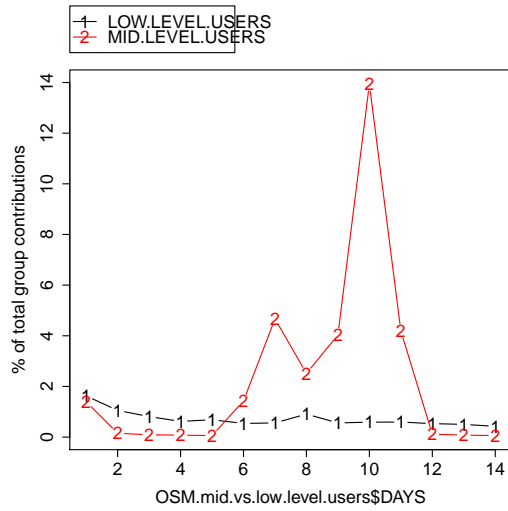


Figure 4: Contribution rates in first two weeks of lifespan, mid vs low-level contributors

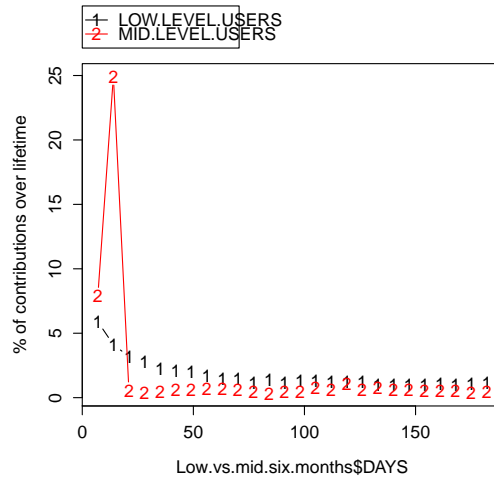


Figure 5: Contribution rates in first 6 months of lifespan, mid vs low-level contributors

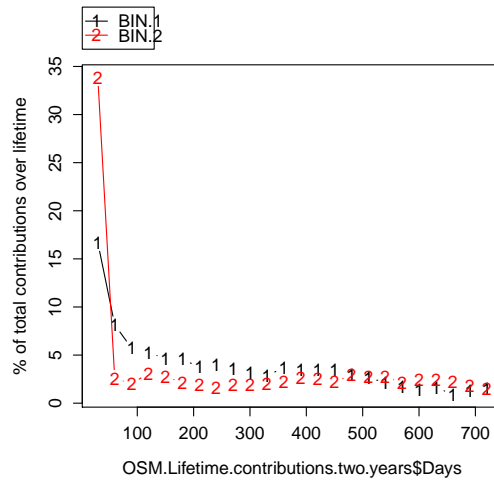


Figure 6: Contribution rates over first two years of lifespan

Lifetime contribution rates - TPB

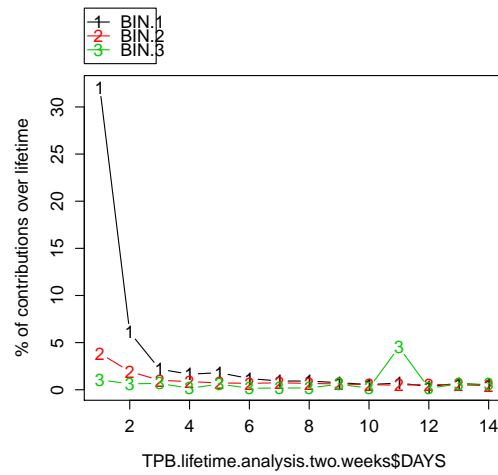


Figure 7: Contribution rates over first two weeks of lifespan

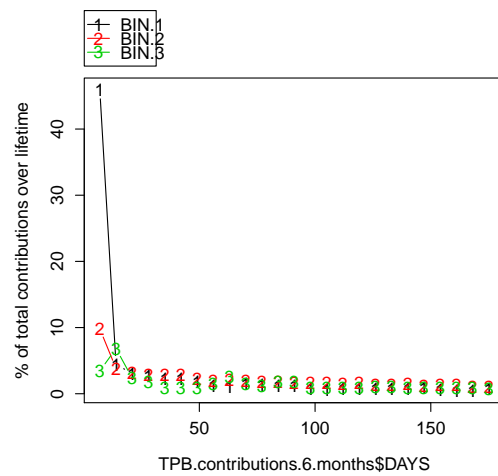


Figure 8: Contribution rates over six months of lifespan

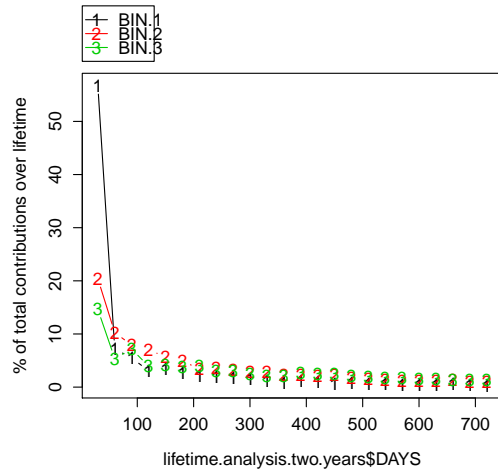


Figure 9: Contribution rates over first two years of lifespan

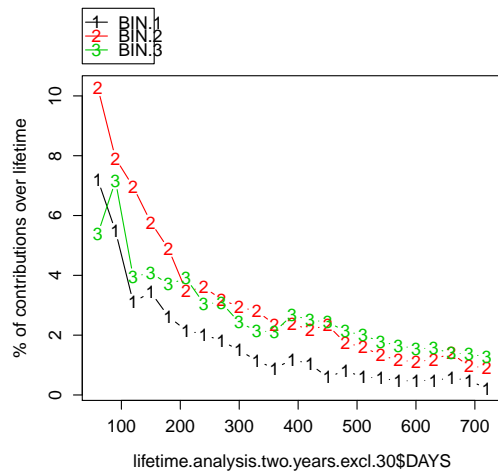


Figure 10: Contribution rates over first two years of lifespan, excluding the first 30 days of lifespan

Lifetime Contribution rates - OSM vs TPB

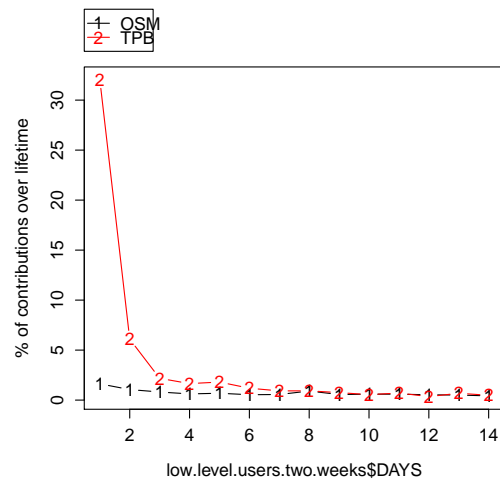


Figure 11: Contribution rates by low-level users over first two weeks of lifespan

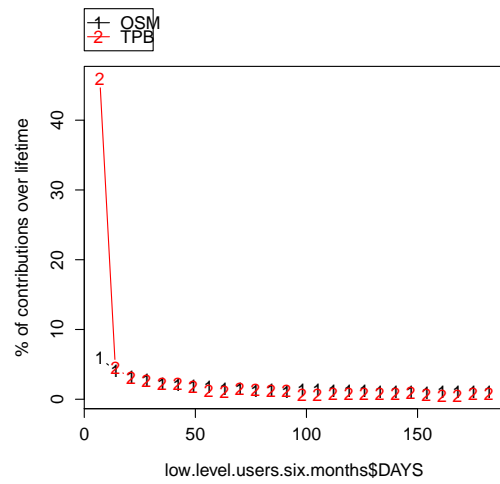


Figure 12: Contribution rates by low-level users over first six months of lifespan

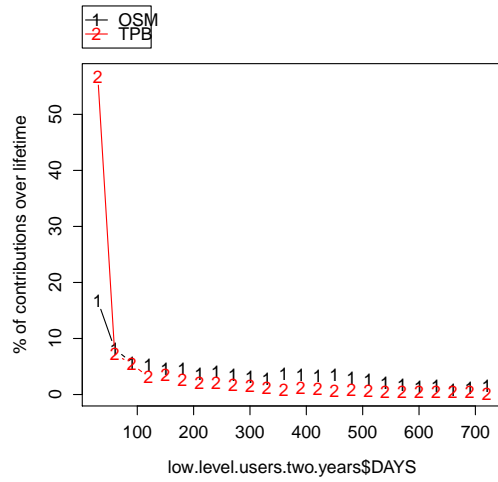


Figure 13: Contribution rates by low-level users over first six months of lifespan

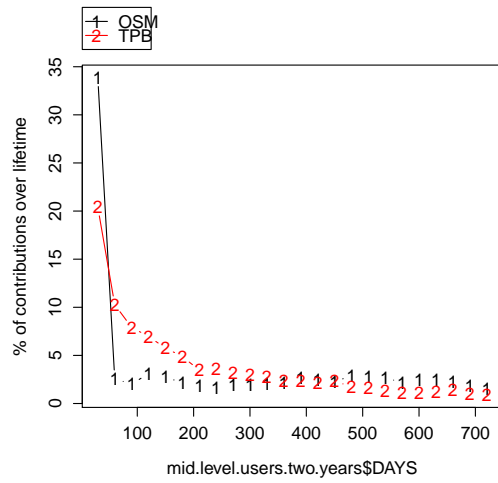


Figure 14: Contribution rates by mid-level users over first two years of lifespan

Appendix 4 - Creative Commons License Attribution-Share Alike

License

THE WORK (AS DEFINED BELOW) IS PROVIDED UNDER THE TERMS OF THIS CREATIVE COMMONS PUBLIC LICENSE ("CCPL" OR "LICENSE"). THE WORK IS PROTECTED BY COPYRIGHT AND/OR OTHER APPLICABLE LAW. ANY USE OF THE WORK OTHER THAN AS AUTHORIZED UNDER THIS LICENSE OR COPYRIGHT LAW IS PROHIBITED.

BY EXERCISING ANY RIGHTS TO THE WORK PROVIDED HERE, YOU ACCEPT AND AGREE TO BE BOUND BY THE TERMS OF THIS LICENSE. TO THE EXTENT THIS LICENSE MAY BE CONSIDERED TO BE A CONTRACT, THE LICENSOR GRANTS YOU THE RIGHTS CONTAINED HERE IN CONSIDERATION OF YOUR ACCEPTANCE OF SUCH TERMS AND CONDITIONS.

1. Definitions

- (a) "Adaptation" means a work based upon the Work, or upon the Work and other pre-existing works, such as a translation, adaptation, derivative work, arrangement of music or other alterations of a literary or artistic work, or phonogram or performance and includes cinematographic adaptations or any other form in which the Work may be recast, transformed, or adapted including in any form recognizably derived from the original, except that a work that constitutes a Collection will not be considered an Adaptation for the purpose of this License. For the avoidance of doubt, where the Work is a musical work, performance or phonogram, the synchronization of the Work in timed-relation with a moving image ("synching") will be considered an Adaptation for the purpose of this License.
- (b) "Collection" means a collection of literary or artistic works, such as encyclopedias and anthologies, or performances, phonograms or broadcasts, or other works or subject matter other than works listed in Section 1(f) below, which, by reason of the selection and arrangement of their contents, constitute intellectual creations, in which the Work is included in its entirety in unmodified form along with one or more other contributions, each constituting separate and independent

works in themselves, which together are assembled into a collective whole. A work that constitutes a Collection will not be considered an Adaptation (as defined below) for the purposes of this License.

- (c) "Creative Commons Compatible License" means a license that is listed at <http://creativecommons.org/compatiblelicenses> that has been approved by Creative Commons as being essentially equivalent to this License, including, at a minimum, because that license: (i) contains terms that have the same purpose, meaning and effect as the License Elements of this License; and, (ii) explicitly permits the relicensing of adaptations of works made available under that license under this License or a Creative Commons jurisdiction license with the same License Elements as this License.
- (d) "Distribute" means to make available to the public the original and copies of the Work or Adaptation, as appropriate, through sale or other transfer of ownership.
- (e) "License Elements" means the following high-level license attributes as selected by Licensor and indicated in the title of this License: Attribution, ShareAlike.
- (f) "Licensor" means the individual, individuals, entity or entities that offer(s) the Work under the terms of this License.
- (g) "Original Author" means, in the case of a literary or artistic work, the individual, individuals, entity or entities who created the Work or if no individual or entity can be identified, the publisher; and in addition (i) in the case of a performance the actors, singers, musicians, dancers, and other persons who act, sing, deliver, declaim, play in, interpret or otherwise perform literary or artistic works or expressions of folklore; (ii) in the case of a phonogram the producer being the person or legal entity who first fixes the sounds of a performance or other sounds; and, (iii) in the case of broadcasts, the organization that transmits the broadcast.
- (h) "Work" means the literary and/or artistic work offered under the terms of this License including without limitation any production in the literary, scientific and artistic domain, whatever may be the mode or form of its expression including digital form, such as a book, pamphlet and other writing; a lecture, address, sermon or other work of the same nature; a dramatic or dramatico-musical work; a choreographic work or entertainment in dumb show; a musical composition with or without words; a cinematographic work to which are assimilated works expressed by a process analogous to cinematography; a work of drawing, painting, architecture, sculpture, engraving or lithography; a photographic work to which are assimilated works expressed by a process analogous to photography; a work of applied art; an illustration, map, plan, sketch or three-dimensional work relative to geography, topography, architecture or science; a performance; a broadcast; a phonogram; a compilation of data to the extent it is protected as a copyrightable work; or a work performed by a variety or circus performer to the extent it is not otherwise considered a literary or artistic work.

- (i) "You" means an individual or entity exercising rights under this License who has not previously violated the terms of this License with respect to the Work, or who has received express permission from the Licensor to exercise rights under this License despite a previous violation.
 - (j) "Publicly Perform" means to perform public recitations of the Work and to communicate to the public those public recitations, by any means or process, including by wire or wireless means or public digital performances; to make available to the public Works in such a way that members of the public may access these Works from a place and at a place individually chosen by them; to perform the Work to the public by any means or process and the communication to the public of the performances of the Work, including by public digital performance; to broadcast and rebroadcast the Work by any means including signs, sounds or images.
 - (k) "Reproduce" means to make copies of the Work by any means including without limitation by sound or visual recordings and the right of fixation and reproducing fixations of the Work, including storage of a protected performance or phonogram in digital form or other electronic medium.
2. Fair Dealing Rights. Nothing in this License is intended to reduce, limit, or restrict any uses free from copyright or rights arising from limitations or exceptions that are provided for in connection with the copyright protection under copyright law or other applicable laws.
3. License Grant. Subject to the terms and conditions of this License, Licensor hereby grants You a worldwide, royalty-free, non-exclusive, perpetual (for the duration of the applicable copyright) license to exercise the rights in the Work as stated below:
- (a) to Reproduce the Work, to incorporate the Work into one or more Collections, and to Reproduce the Work as incorporated in the Collections;
 - (b) to create and Reproduce Adaptations provided that any such Adaptation, including any translation in any medium, takes reasonable steps to clearly label, demarcate or otherwise identify that changes were made to the original Work. For example, a translation could be marked "The original work was translated from English to Spanish," or a modification could indicate "The original work has been modified.";
 - (c) to Distribute and Publicly Perform the Work including as incorporated in Collections; and,
 - (d) to Distribute and Publicly Perform Adaptations.
 - (e) For the avoidance of doubt:
 - i. Non-waivable Compulsory License Schemes. In those jurisdictions in which the right to collect royalties through any statutory or compulsory licensing scheme cannot be waived, the Licensor

reserves the exclusive right to collect such royalties for any exercise by You of the rights granted under this License;

- ii. Waivable Compulsory License Schemes. In those jurisdictions in which the right to collect royalties through any statutory or compulsory licensing scheme can be waived, the Licensor waives the exclusive right to collect such royalties for any exercise by You of the rights granted under this License; and,
- iii. Voluntary License Schemes. The Licensor waives the right to collect royalties, whether individually or, in the event that the Licensor is a member of a collecting society that administers voluntary licensing schemes, via that society, from any exercise by You of the rights granted under this License.

The above rights may be exercised in all media and formats whether now known or hereafter devised. The above rights include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. Subject to Section 8(f), all rights not expressly granted by Licensor are hereby reserved.

4. Restrictions. The license granted in Section 3 above is expressly made subject to and limited by the following restrictions:

- (a) You may Distribute or Publicly Perform the Work only under the terms of this License. You must include a copy of, or the Uniform Resource Identifier (URI) for, this License with every copy of the Work You Distribute or Publicly Perform. You may not offer or impose any terms on the Work that restrict the terms of this License or the ability of the recipient of the Work to exercise the rights granted to that recipient under the terms of the License. You may not sublicense the Work. You must keep intact all notices that refer to this License and to the disclaimer of warranties with every copy of the Work You Distribute or Publicly Perform. When You Distribute or Publicly Perform the Work, You may not impose any effective technological measures on the Work that restrict the ability of a recipient of the Work from You to exercise the rights granted to that recipient under the terms of the License. This Section 4(a) applies to the Work as incorporated in a Collection, but this does not require the Collection apart from the Work itself to be made subject to the terms of this License. If You create a Collection, upon notice from any Licensor You must, to the extent practicable, remove from the Collection any credit as required by Section 4(c), as requested. If You create an Adaptation, upon notice from any Licensor You must, to the extent practicable, remove from the Adaptation any credit as required by Section 4(c), as requested.
- (b) You may Distribute or Publicly Perform an Adaptation only under the terms of: (i) this License; (ii) a later version of this License with the same License Elements as this License; (iii) a Creative Commons jurisdiction license (either this or a later license version) that contains the same License Elements as this License (e.g., Attribution-ShareAlike 3.0 US); (iv) a Creative Commons Compatible License.

If you license the Adaptation under one of the licenses mentioned in (iv), you must comply with the terms of that license. If you license the Adaptation under the terms of any of the licenses mentioned in (i), (ii) or (iii) (the "Applicable License"), you must comply with the terms of the Applicable License generally and the following provisions: (I) You must include a copy of, or the URI for, the Applicable License with every copy of each Adaptation You Distribute or Publicly Perform; (II) You may not offer or impose any terms on the Adaptation that restrict the terms of the Applicable License or the ability of the recipient of the Adaptation to exercise the rights granted to that recipient under the terms of the Applicable License; (III) You must keep intact all notices that refer to the Applicable License and to the disclaimer of warranties with every copy of the Work as included in the Adaptation You Distribute or Publicly Perform; (IV) when You Distribute or Publicly Perform the Adaptation, You may not impose any effective technological measures on the Adaptation that restrict the ability of a recipient of the Adaptation from You to exercise the rights granted to that recipient under the terms of the Applicable License. This Section 4(b) applies to the Adaptation as incorporated in a Collection, but this does not require the Collection apart from the Adaptation itself to be made subject to the terms of the Applicable License.

- (c) If You Distribute, or Publicly Perform the Work or any Adaptations or Collections, You must, unless a request has been made pursuant to Section 4(a), keep intact all copyright notices for the Work and provide, reasonable to the medium or means You are utilizing: (i) the name of the Original Author (or pseudonym, if applicable) if supplied, and/or if the Original Author and/or Licensor designate another party or parties (e.g., a sponsor institute, publishing entity, journal) for attribution ("Attribution Parties") in Licensor's copyright notice, terms of service or by other reasonable means, the name of such party or parties; (ii) the title of the Work if supplied; (iii) to the extent reasonably practicable, the URI, if any, that Licensor specifies to be associated with the Work, unless such URI does not refer to the copyright notice or licensing information for the Work; and (iv) , consistent with Section 3(b), in the case of an Adaptation, a credit identifying the use of the Work in the Adaptation (e.g., "French translation of the Work by Original Author," or "Screenplay based on original Work by Original Author"). The credit required by this Section 4(c) may be implemented in any reasonable manner; provided, however, that in the case of a Adaptation or Collection, at a minimum such credit will appear, if a credit for all contributing authors of the Adaptation or Collection appears, then as part of these credits and in a manner at least as prominent as the credits for the other contributing authors. For the avoidance of doubt, You may only use the credit required by this Section for the purpose of attribution in the manner set out above and, by exercising Your rights under this License, You may not implicitly or explicitly assert or imply any connection with, sponsorship or endorsement by the Original

Author, Licensor and/or Attribution Parties, as appropriate, of You or Your use of the Work, without the separate, express prior written permission of the Original Author, Licensor and/or Attribution Parties.

- (d) Except as otherwise agreed in writing by the Licensor or as may be otherwise permitted by applicable law, if You Reproduce, Distribute or Publicly Perform the Work either by itself or as part of any Adaptations or Collections, You must not distort, mutilate, modify or take other derogatory action in relation to the Work which would be prejudicial to the Original Author's honor or reputation. Licensor agrees that in those jurisdictions (e.g. Japan), in which any exercise of the right granted in Section 3(b) of this License (the right to make Adaptations) would be deemed to be a distortion, mutilation, modification or other derogatory action prejudicial to the Original Author's honor and reputation, the Licensor will waive or not assert, as appropriate, this Section, to the fullest extent permitted by the applicable national law, to enable You to reasonably exercise Your right under Section 3(b) of this License (right to make Adaptations) but not otherwise.

5. Representations, Warranties and Disclaimer

UNLESS OTHERWISE MUTUALLY AGREED TO BY THE PARTIES IN WRITING, LICENSOR OFFERS THE WORK AS-IS AND MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND CONCERNING THE WORK, EXPRESS, IMPLIED, STATUTORY OR OTHERWISE, INCLUDING, WITHOUT LIMITATION, WARRANTIES OF TITLE, MERCHANTIBILITY, FITNESS FOR A PARTICULAR PURPOSE, NONINFRINGEMENT, OR THE ABSENCE OF LATENT OR OTHER DEFECTS, ACCURACY, OR THE PRESENCE OF ABSENCE OF ERRORS, WHETHER OR NOT DISCOVERABLE. SOME JURISDICTIONS DO NOT ALLOW THE EXCLUSION OF IMPLIED WARRANTIES, SO SUCH EXCLUSION MAY NOT APPLY TO YOU.

6. Limitation on Liability. EXCEPT TO THE EXTENT REQUIRED BY APPLICABLE LAW, IN NO EVENT WILL LICENSOR BE LIABLE TO YOU ON ANY LEGAL THEORY FOR ANY SPECIAL, INCIDENTAL, CONSEQUENTIAL, PUNITIVE OR EXEMPLARY DAMAGES ARISING OUT OF THIS LICENSE OR THE USE OF THE WORK, EVEN IF LICENSOR HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

7. Termination

- (a) This License and the rights granted hereunder will terminate automatically upon any breach by You of the terms of this License. Individuals or entities who have received Adaptations or Collections from You under this License, however, will not have their licenses terminated provided such individuals or entities remain in full compliance with those licenses. Sections 1, 2, 5, 6, 7, and 8 will survive any termination of this License.

- (b) Subject to the above terms and conditions, the license granted here is perpetual (for the duration of the applicable copyright in the Work). Notwithstanding the above, Licensor reserves the right to release the Work under different license terms or to stop distributing the Work at any time; provided, however that any such election will not serve to withdraw this License (or any other license that has been, or is required to be, granted under the terms of this License), and this License will continue in full force and effect unless terminated as stated above.

8. Miscellaneous

- (a) Each time You Distribute or Publicly Perform the Work or a Collection, the Licensor offers to the recipient a license to the Work on the same terms and conditions as the license granted to You under this License.
- (b) Each time You Distribute or Publicly Perform an Adaptation, Licensor offers to the recipient a license to the original Work on the same terms and conditions as the license granted to You under this License.
- (c) If any provision of this License is invalid or unenforceable under applicable law, it shall not affect the validity or enforceability of the remainder of the terms of this License, and without further action by the parties to this agreement, such provision shall be reformed to the minimum extent necessary to make such provision valid and enforceable.
- (d) No term or provision of this License shall be deemed waived and no breach consented to unless such waiver or consent shall be in writing and signed by the party to be charged with such waiver or consent.
- (e) This License constitutes the entire agreement between the parties with respect to the Work licensed here. There are no understandings, agreements or representations with respect to the Work not specified here. Licensor shall not be bound by any additional provisions that may appear in any communication from You. This License may not be modified without the mutual written agreement of the Licensor and You.
- (f) The rights granted under, and the subject matter referenced, in this License were drafted utilizing the terminology of the Berne Convention for the Protection of Literary and Artistic Works (as amended on September 28, 1979), the Rome Convention of 1961, the WIPO Copyright Treaty of 1996, the WIPO Performances and Phonograms Treaty of 1996 and the Universal Copyright Convention (as revised on July 24, 1971). These rights and subject matter take effect in the relevant jurisdiction in which the License terms are sought to be enforced according to the corresponding provisions of the implementation of those treaty provisions in the applicable national law. If the standard suite of rights granted under applicable copyright law includes additional rights not granted under this License, such additional rights are deemed to be included in the License; this License

is not intended to restrict the license of any rights under applicable law.