Piracy, Public Access, and Preservation

An Exploration of Sustainable Accessibility in a Public Torrent Index

John D. Martin III

School of Information and Library Science, University of North Carolina at Chapel Hill

me@johndmart.in

@jdmar3

Abstract

Members of the Pirate Parties International *have* ***claimed that torrent networks have the potential to function as repositories*** *for popular cultural materials and historical primary source data.* The Pirate Bay *is the world's most popular public index for tracking and downloading torrents. Using a snapshot of torrents on the site,* ***this study considers the potential for torrent networks to preserve and provide access*** *to cultural materials in the form of digital media content. Metadata from 2.1 million torrents were categorized by media type and the robustness of given torrents was assessed. Trends over time, such as number of uploads and volume, were also investigated. This study found that* ***relatively few torrents exhibit long-term survivability****, even though the overall volume in the index shows continuous increase.*

Research Questions

1. What is the shape of *The Pirate Bay* as a repository in terms of media types represented by percentage?
2. What trends in media uploading and sharing can be identified in the *The Pirate Bay*?
3. How robust is a torrent-based network in terms of preserving media and making it available as represented by *The Pirate Bay*?

Data

The dataset used in this study was gathered via web scrape from *The Pirate Bay* in 2013 by Karel Bílek. The archive contains metadata for **2,142,134 torrent files** indexed by *The Pirate Bay*. Bílek published the data on *The Pirate Bay* so that it could be used by others for research purposes (`http://karelbilek.com/pirateay`).

RQ1. Shape of *The Pirate Bay*

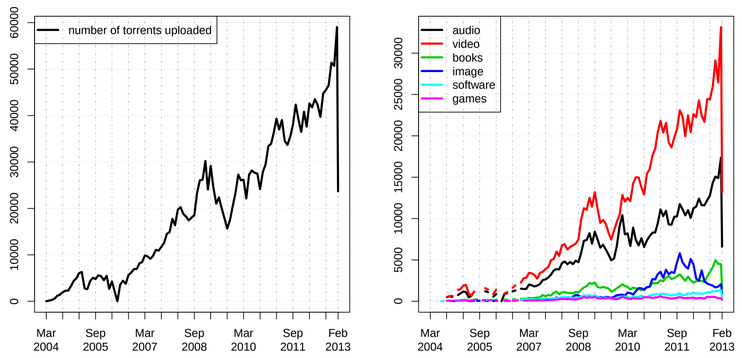
Media types were extracted by matching features in the title field of the torrent records. The records were then bootstrapped ($2,500 \times 10,000$ resamples) to assess the accuracy of the extraction.

Table 1: *Percentage of media types in* The Pirate Bay *compared with estimates from bootstrap using t-tests ($p < 0.001$, $df = 9999$). Bootstrapping was used to estimate the accuracy of the extraction technique.*

	All	Bootstrap estimates				t-tests
Type	**%**	**Mean**	**SD**	**Min**	**Max**	***t*-score**
Audio	27.407	27.421	0.901	24.280	30.960	-301091.90
Video	48.608	48.606	0.999	45.080	52.400	-481773.40
Books	6.644	6.641	0.496	4.680	8.880	-333924.00
Image	5.380	5.379	0.449	3.600	7.200	-118584.50
Software	2.277	2.278	0.299	1.320	3.480	-75423.96
Games	1.283	1.283	0.225	0.560	2.160	-56402.36

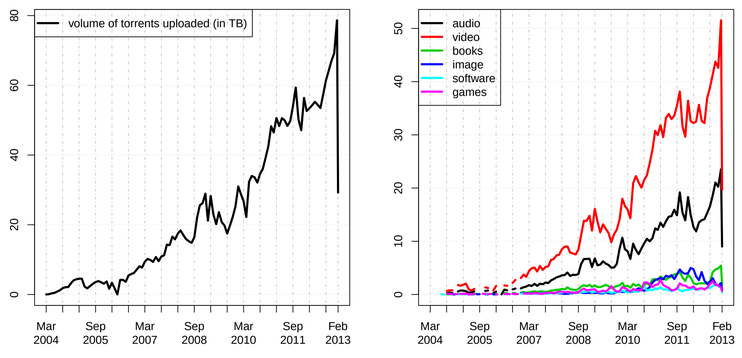
RQ2. Uploading Trends in *The Pirate Bay*

Trends in torrent uploads were assessed by considering **number of monthly uploads** (figure 1), **data volume** of uploaded per month (figure 2), and **total size** of all data uploaded over the life of *The Pirate Bay* (figure 3).



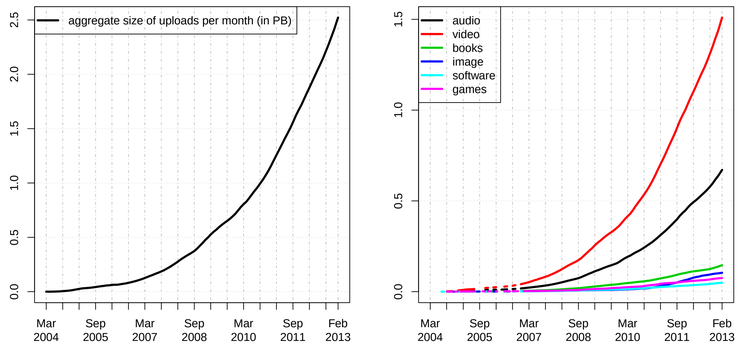
(a) (b)

Figure 1: *Number of uploads per month in aggregate (a) and split by category (b).*



(a) (b)

Figure 2: *Volume (in TB) of uploads per month in aggregate (a) and split by category (b).*



(a) (b)

Figure 3: *Total size (in PB) of uploads per month in aggregate (a) and split by category (b).*

Different media types have increased at different rates, in part due to demands and constraints on file type and size. For example, as the demand for HD video has increased, more and larger video files have been shared through torrent networks. **The number of torrents has continued to increase at a very high rate** to more than 5 million since this data was collected from *The Pirate Bay* in 2013.

Background

The Pirate Bay is a **public torrent index** used by millions of people all over the world to pirate various types of media content. **Swarms are resilient, self-healing networks of peers** running the BitTorrent protocol which allows users to exchange pieces rather than entire files. **Trackers aggregate information** about who and what is on the network.

Torrent: a metadata file for the BitTorrent protocol. It contains a **map** to verify the content being shared.

Seeders: peers **uploading** of pieces from full copies of all files indexed by a given torrent.

Leechers: peers **downloading** pieces from other peers in the swarm. They also upload pieces of partial files they hold.

RQ3. Robustness of *The Pirate Bay*

Based on a count of seeders and leechers, torrents were automatically sorted into 3 categories for survivability.

Table 2: *Robustness of torrents in dataset based on number of seeders (i.e., users uploading full copies of torrent-tracked content files available) in the swarm.*

Category	**Definition**	**# torrents**	**% torrents**
Extinct	0 seeders	778,655	36.35
Endangered	only 1 seeder	498,573	23.27
Robust	2 or more seeders	864,906	40.38
Total		2,142,134	

More than half of the content ever tracked by *The Pirate Bay* was either completely gone or in danger of disappearing.

Conclusions

The loss of torrent files from *The Pirate Bay* can be viewed in several ways:

1. while torrent networks may be resilient, the content is not;
2. **torrent swarms are not currently an effective method for preserving data** without dedicated seeders in place to keep torrents alive;
3. the drop off of torrents may represent a **naturalistic curation or culling process**, i.e., torrents fall away when users no longer view them as valuable.

The metaphor of torrent networks as analog or replacement for public libraries may have political currency, but the reality is quite different. If we accept the value of data in these networks, then the problem is no longer an increase in pirated content, but rather an increase in loss of content from lack of preservation.

Short Paper and References

Please follow the link or QR code to download this poster and read the accompanying short paper with references:
`https://goo.gl/ZEJJLF`

ASIS&T 79th Annual Meeting, 14-18 October 2016, Copenhagen, Denmark

The iSchool@UNC Chapel Hill