How Prevalent is Content Bundling in BitTorrent?

Jinyoung Han
Seoul National University
jyhan@mmlab.snu.ac.kr

Taejoong Chung
Seoul National University
tjchung@mmlab.snu.ac.kr

Seungbae Kim
Seoul National University
sbkim@mmlab.snu.ac.kr

Ted "Taekyoung" Kwon
Seoul National University
tkkwon@snu.ac.kr

Hyun-chul Kim
Seoul National University
hkim@mmlab.snu.ac.kr

Yanghee Choi
Seoul National University
yhchoi@snu.ac.kr

ABSTRACT

Despite the increasing interest in content bundling in BitTorrent systems, there are still few empirical studies on the bundling practice in real BitTorrent communities. In this paper, we conduct comprehensive measurements on one of the largest BitTorrent portals: The Pirate Bay. From the torrents data set collected for 38 days from April to May, 2010, we study how prevalent bundling is and how many files are bundled in a torrent, across different types of contents shared: Movie, Porn, TV, Music, Application, E-book, and Game.

Categories and Subject Descriptors

H.4.3 [**Information Systems Applications**]: Communications Applications

General Terms

Measurement

Keywords

BitTorrent, Content Bundling, Measurement

1. INTRODUCTION

According to the Ipoque's report in 2009 [1], BitTorrent accounts for approximately 27-55% of Internet traffic. The huge success of BitTorrent is attributed to the attractive properties of its swarming operations. First, the swarming technique scales well even in the presence of flash crowds for popular files. Second, cooperations among peers in a swarm stimulated by the tit-for-tat incentive mechanism improve the overall system performance like throughput. Third, the tit-for-tat mechanism also addresses the free-riding problem.

Despite the success of BitTorrent, its swarming system suffers from a fundamental limitation: little or no availability of unpopular files [6]. That is, peers arriving after the initial flash crowd may end up with finding the file unavailable [6]. Recently, *bundling*[1] in peer-to-peer (P2P) swarming systems like BitTorrent has gained much attention, as it can mitigate the availability problem of unpopular files [5,6] as well as reduce download times [4–6].

However, to our knowledge, there have been few efforts to empirically investigate the practice of content bundling in P2P swarming systems. Most of the prior studies on bundling have been carried out by making simple assumptions on bundling strategies and users' accesses to bundles with no empirical basis, which motivates our measurement study on the bundling practices of the BitTorrent users.

As a first step towards providing empirical grounds for understanding and modeling content bundling in BitTorrent, we make the following contributions: (1) to our knowledge, this is the first measurement study on the bundling practice by observing one of the largest real BitTorrent portals – The Pirate Bay [2]. For 38 days, we have collected the trace data on 114 K torrents containing 1.6 M files. (2) We find that around 70% of BitTorrent torrents contain multiple files, i.e., bundling is widely used. (3) We find that the total volume of multi-file [2] torrents outweighs that of single-file ones; the volume of multi-file torrents accounts for over 80% of the total data volume contained in all the torrents in our data. (4) We show how many files are bundled in a multi-file torrent, across different types of contents shared: Movie, Porn, TV, Music, Application, E-book, and Game.

2. METHODOLOGY

We conducted a measurement study on a real BitTorrent portal, The Pirate Bay (TPB) [2], one of the most popular torrent hosting sites. For the purpose of data collection, we developed a crawling agent to periodically fetch newly released ".torrent" files[3]. For each torrent, the torrent metadata which consists of its .torrent information, category given by the torrent publisher, and published time is recorded.

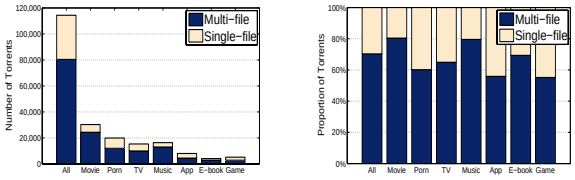
Our datasets have been collected from April 22 to May 29, 2010. Our crawling agent fetched the torrent data of 113,993 torrents from TPB, which contain 1,642,299 files. Throughout this paper, we investigate the bundling practice from the seven major (86% and 83% in terms of the total torrent counts and data volume, respectively) content categories given by the torrent publisher: Movie, Porn, TV,

[1]Bundling is a common practice in which a publisher packages multiple files (e.g. multiple episodes of the same sitcom) and disseminates them via a single swarm [6], instead of disseminating individual files via separate swarms.

[2]Throughout this paper, multi-file torrent indicates bundled torrent which contains multiple files, while single-file torrent indicates non-bundled torrent which contains a single file.

[3]Note that *a .torrent file* contains the metadata of content file(s) to download such as its name, size, and the tracker information, whereas *a torrent* refers to content file(s) to download in a swarm.

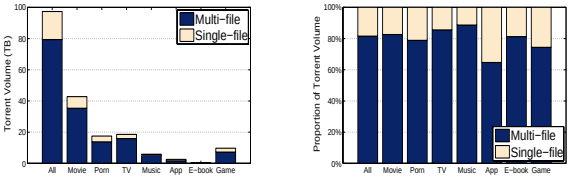
SIGMETRICS'11, June 7–11, 2011, San Jose, California, USA.
ACM 978-1-4503-0262-3/11/06.



(a) The numbers of torrents

(b) Percentage of torrent counts

Figure 1: Bundling is widely used in BitTorrent.



(a) Total volume of data

(b) Percentage of data volume

Figure 2: Total volume of multi-file torrents is substantially larger than that of single-file ones.

Music, Application, E-book and Game. The number of torrents (and total volume of data contained in the torrents) of the Movie, Porn, TV, Music, Application, E-book and Game categories constitute approximately 26% (36%), 17% (15%), 13% (16%), 14% (5%), 7% (2%), 4% (1%), and 5% (8%) of all the torrents at TPB, respectively.

3. BUNDLING PRACTICE

To analyze how prevalent content bundling is in BitTorrent, we first compare multi-file torrents with single-file ones in terms of the number and total volume of data contained in the torrents. Figure 1 shows that around 70% of the torrents contain multiple files, which means content bundling is widely used. In the Music category, over 80% of the torrents are multi-file ones, which indicates that BitTorrent users often share a collection of music files from the same genre, player, composer, or album. Likewise, over 80% of the torrents in the Movie category are multi-file ones, mostly because users often package: (i) multiple movie files of the same series (e.g. sequels), or (ii) a main video file and other supplementary files like subtitles. Meanwhile, around 50% of the torrents in the Application and Game categories are bundled; almost half of the torrents in these categories have a single compressed file, while the other half of the torrents consist of multiple files (e.g. installation files and subsidiary files such as how-to documents).

We next compare the total data volume contained in multi-file and single-file torrents in Figure 2. Figures 2(a) and 2(b) show that the total volume of multi-file torrents outweighs that of single-file ones; the volume of multi-file torrents accounts for over 80% of all the data that we investigated. The volume of all the torrents reaches around 100 TB.

Finally, we investigate how many files are bundled in the multi-file torrents. Total number of files of all the 80,152 multi-file torrents is 1,608,458; i.e, 20 files per multi-file torrent in average. Figure 3 shows the CDF of the number of

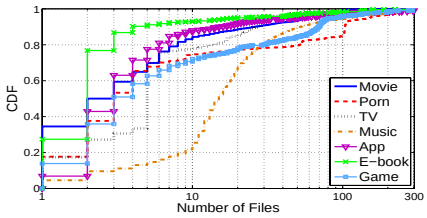


Figure 3: Number of files in a multi-file torrent

files in a multi-file torrent across different categories of torrents. As shown in Figure 3, multi-file torrents of the Music category contain significantly more number of files than those of the other categories; almost 80% of Music multi-file torrents contain more than 10 files. In contrast, 90% of E-book multi-file torrents are comprised of less than 4 files. Note that in the Porn multi-file torrents, around 70% contain less than 7 files, while over 16% contain a large number of files; over 100.

4. CONCLUDING REMARKS

We conducted measurements on the bundling practice from one of the largest BitTorrent portals: The Pirate Bay. From the BitTorrent data set collected for 38 days, we found that bundling is widespread for file sharing. Depending on the content categories such as movies and music, the bundling practice showed a wide variety of characteristics in terms of the number of files bundled in a torrent. Our ongoing work includes investigation of (i) how users access the bundled files in comparison with the non-bundled ones, and (ii) whether and how bundling patterns are similar or different, depending on the main incentives of content publishers: altruistic or profit-driven [3].

5. ACKNOWLEDGMENTS

This work was supported by NAP of Korea Research Council of Fundamental Science and Technology and the ITRC support program NIPA-2011-(C1090-1111-0004) of MKE/NIPA. The ICT at Seoul National University provided research facilities for this study.

6. REFERENCES

[1] The impact of p2p file sharing, voice over ip, instant messaging, one-click hosting and media streaming on the internet. `http://www.ipoque.com/resources/internet-studies/internet-study-2008_2009`.

[2] The pirate bay. `http://thepiratebay.org/`.

[3] R. Cuevas, M. Kryczka, A. Cuevas, S. Kaune, C. Guerrero, and R. Rejaie. Is content publishing in bittorrent altruistic or profit-driven? In *ACM CoNEXT*, 2010.

[4] N. Lev-tov, N. Carlsson, Z. Li, C. Williamson, and S. Zhang. Dynamic file-selection policies for bundling in bittorrent-like systems. In *IEEE IWQOS*, 2010.

[5] D. S. Menasche, G. Neglia, D. Towsley, and S. Zilberstein. Strategic reasoning about bundling in swarming systems. In *IEEE GameNets*, 2009.

[6] D. S. Menasche, A. A. Rocha, B. Li, D. Towsley, and A. Venkataramani. Content availability and bundling in swarming systems. In *ACM CoNEXT*, 2009.