ABSTRACT

Despite the tremendous success of BitTorrent, its swarming system suffers from a fundamental limitation: lower or no availability of unpopular contents. Recently, Menasche et al. has shown that bundling 1 is a promising solution to mitigate this availability problem; it improves the availability and reduces download times for unpopular contents by combining multiple files into a single swarm. There also have been studies on bundling strategies and performance issues in bundled swarms. Despite the recent surge of interest in the benefits of and strategies for bundling, there are still little empirical grounding for understanding, describing, and modeling it. This is the first empirical study that measure and analyze how prevalent contents bundling is in BitTorrent and how peers access the bundled contents, in comparison to the other non-bundled (i.e., single-filed) ones. To our surprise, we found that around 70% of BitTorrent swarms contain multiple files, which indicate that bundling has become widespread for contents sharing. We also show that the amount of bytes shared in bundled swarms is estimated to be around 85% out of all the BitTorrent contents logged in our datasets. Inspired from our findings, we raise and discuss three important research questions in the field of file sharing systems as well as future contents-oriented networking: i) bundling strategies, ii) bundling-aware sharing systems in BitTorrent, and iii) implications on content-oriented networking.

Keywords

Content Bundling, Swarming System, Peer-to-Peer System, BitTorrent, Content-Oriented Networking

1. INTRODUCTION

BitTorrent [5] has become the de facto standard in sharing contents on the Internet. According to the

1Bundling is a common strategy adopted by BitTorrent publishers by which a publisher packages a number of related files (e.g., episodes of a sitcom) and disseminates them via a single larger swarm [15], instead of disseminating individual files via separate swarms.

Ipoque’s report released in 2009 [1], BitTorrent accounts for approximately 27-55% of today’s Internet traffic. The ever increasing usage of BitTorrent is attributed to some attractive properties of its swarming systems: First, cooperation among peers in a swarm stimulated by the tit-for-tat based incentive mechanism to improve the overall system performance in terms of throughput. Second, the tit-for-tat strategy also prevents (or mitigates) the free-riding problem. Third, the swarming technique scales well even in the presence of massive flash crowds for popular contents [13,16].

Despite the success of BitTorrent, its swarming system suffers from a fundamental limitation: little or no availability of unpopular contents. That is, peers arriving after the initial flash crowd may often end up with finding the content unavailable, not to mention unpopular contents [15]. Recently, Menasche et al. has shown that bundling is a promising solution to mitigate this availability problem; it improves the availability and reduces download times for unpopular contents by combining multiple files into a single swarm [15]. There also have been studies on bundling strategies [9,14] and performance issues in bundled swarms [19].

Despite the recent surge of interest in the benefits of and strategies for bundling, there is still little empirical grounding for understanding, describing, and modeling it. This leaves researchers uncertain of (i) whether and how the assumptions (on both bundling pattern and user access pattern, with little if any empirical evidence) made in the aforementioned work are congruent with (or close to) reality, and (ii) the status quo and performance and architectural implications of bundling for the advancement of file sharing techniques or architectures (not to mention BitTorrent), as well as of the so-called future ‘content-oriented networking’. To begin to provide empirical grounds for understanding, describing, and modeling contents bundling in BitTorrent, we make the following contributions: (1) This is the first empirical study that measures and analyzes how prevalent contents bundling is in BitTorrent and
how peers access the bundled contents, in comparison to the other non-bundled (i.e., single-filed) ones. To this end, we first developed a BitTorrent monitoring software by modifying the Azureus, one of the most popular BitTorrent clients. Using the developed code, we have collected over a month of data comprised of 36K distinct swarms of 2.8 Million distinct peers. (2) To our surprise, we found that around 70% of BitTorrent swarms contain multiple files, which indicate that bundling has become widely used for contents sharing. (3) We also estimated that the amount of bytes shared in bundled swarms are much larger than that in single-filed swarms: the amount of data contained in bundled swarms are about 2 times larger than single-filed ones, the number of bundled swarms are about 2.3 times in average larger than that of single-filed ones, and the average popularity of bundled swarms are 1.2 times larger than single-filed ones, which we may approximately estimate that amount of bytes shared in bundled contents account for around 85% out of all the BitTorrent contents logged in our datasets. (4) According to our measurements, the bundling pattern and user access pattern are different depending on the seven BitTorrent content type categories (i.e., Movie, Porn, TV, Music, Applications, E-book, and Game). (5) Based on our findings and lessons learned, we raise and discuss three new important research questions in the area of file sharing systems as well as the future contents-oriented networking: i) bundling strategy, ii) bundling-aware sharing systems in BitTorrent, and iii) implications on content-oriented networking.

This paper is organized as follows. After reviewing related work in Section 2, we present measurement methodology, data, and results in Section 3. Section 4 discusses on implications of content bundling and our ongoing work. Finally, we conclude this paper in Section 5.

2. RELATED WORK

BitTorrent: BitTorrent [5] has been the de facto standard for content sharing. Due to its success in the real world, many studies have been conducted to investigate BitTorrent’s behavior in terms of throughput, fairness and incentive issues, revealing valuable insights into the performance aspect of the BitTorrent [13, 16–18]. To the best of our knowledge, this is the first empirical study that measures and analyzes on content bundling in BitTorrent swarming systems.

Sharing Multiple Files: Most studies on BitTorrent had focused only on sharing and transmission of single torrents until Guo et al. found that 85% of users concurrently access multiple torrents [8]. Yang et al. [20] proposed a novel incentive mechanism for nodes to remain as seeds to consider the situation when a user downloads multiple torrents. This mechanism calculates the aggregated downloading rate in a cross-torrent fashion in the peer selection phase, so that a node can get additional credits from providing pieces in another torrent which it participates in as a seed. These studies assume that only a single file is shared in a single swarm. In contrast, we focus on “bundling”, which allows peers to download multiple files from a single swarm.

Bundling: Our measurement results show that around 70% of BitTorrent swarms contain multiple files, which indicate that bundling has become widely used for contents sharing. Recently, Menasche et al. [14, 15] has shown that bundling can mitigate the availability problem by combining multiple files into a single swarm. Also Tien et al. studied the performance issues in sharing multiple files not only using multiple torrents but also using a bundle in BitTorrent [19]. Despite the recent surge of interest in the benefits of and strategies for bundling, there are still little empirical grounding for understanding, describing, and modeling it. Our work is the first work to begin to provide empirical grounds for understanding, describing, and modeling contents bundling in BitTorrent.

3. MEASUREMENT

In this section, we present an empirical analysis of BitTorrent content bundling through measurements and trace analysis to show that: (1) content bundling in BitTorrent has become widely used for content sharing, (2) the amount of bytes shared in bundled swarms are much larger than in single-filed swarms, and (3) the bundling pattern and user access pattern are different depending on the seven BitTorrent content type categories (i.e., Movie, Porn, TV, Music, Applications, E-book, and Game).

3.1 Methodology

We conducted a measurement study on real BitTorrent swarms starting from Apr 22, 2010 to May 29, 2010. We first developed a BitTorrent monitoring software by modifying the Azureus 2 [2] because the Azureus can discover new neighbors of most BitTorrent clients from other peers by the Peer Exchange Extension (PEx) protocol. We then developed a site monitoring agent to fetch the torrent file (i.e., a “.torrent” file) and its information from a torrent hosting site. Our agent periodically fetches the recently released torrent files and their information from The Pirate Bay [3] which is the most popular torrent hosting site, and then sends the torrent information to a subset of the monitoring clients. When a monitoring client receives a torrent information, it joins the swarm and begins to monitor each peer. Our client requests the peer list to the tracker several times so that it can get most of the peer infor-

2Azureus is also called Vuze, one of the most popular BitTorrent client.
mation in a swarm. In addition, it also leverages the PEX protocol to discover the rest of the peers to cover the entire swarm. We finally save all the swarm information from the hosting site, trackers, and each peer to the database. Each entry in the log data consists of torrent meta information (e.g., a torrent name, file names and a torrent creation time), information of peers (e.g., an IP address and its bitmap which they have), and information from the hosting site (e.g., categories of the torrent file and release time of the torrent file). We have collected data comprised of 36K distinct swarms and 2.8 Million peers.

Throughout this paper, we investigate the bundling pattern and user access pattern based on the seven BitTorrent content type categories (i.e., Movie, Porn, TV, Music, Application, E-book, Game) which accounts for 87% of the swarms logged in our datasets. The number of swarms of Movie, Porn, TV, Music, Application, E-book and Game accounts for 24%, 19%, 17%, 15%, 6%, 4%, and 4% of all the swarms, respectively.

### 3.2 Bundling is Widespread

To analyze how prevalent contents bundling is in BitTorrent and how peers access the bundled contents, we compare bundled swarms to the other non-bundled (i.e., single-filed) ones in terms of the proportion, swarm size (i.e., amount of data contained in a swarm), popularity, and availability.

Figure 1 shows that around 70% of BitTorrent swarms contain multiple files, which means content bundling is widely used in BitTorrent today for content sharing. For the music type, over than 80% of swarms are bundled ones, which we guess that users may prefer to share songs through bundling based on a famous singer or a recently released album. Meanwhile, around 50% of swarms are bundled for the types of game and application, which are lower than other content types, because we suspect that games and applications are often shared by a single installation file. Nevertheless, bundled swarms are more widely used for contents sharing regardless of content types.

We next compare the swarm size of bundled and single-filed swarms in Figure 2. Because there are a number of files in bundled swarms but there is only a file in single-filed swarms, the swarm size of bundled swarms is much larger (about 2 times in average) than that of single-filed ones. That is, the amount of data contained in bundled swarms are much larger than than that in single-filed ones.

To investigate the usage pattern of bundled swarms and single-filed ones in BitTorrent, we calculate the popularity of bundled swarms and the single-filed ones. As Dan et al. showed in [6], there are two methods to measure the content popularity; instantaneous popularity and the popularity over time. To measure not only the instantaneous popularity but also the popularity over time, we combine the concepts of two methods into a single novel content popularity metric, which we define “accumulated popularity”. The accumulated

---

3We used the content type categories information which are tagged in The Pirate Bay [3].

4The swarm size indicates the summation of file sizes in a swarm.
3.3 The Status Quo of Content Bundling: Bundling Pattern and User Access Pattern

In this subsection, we investigate the bundling pattern and user access pattern depending on the seven content type categories to understand the status quo of content bundling.

Figure 5 shows the comparisons between bundled swarms and single-filed ones in terms of the popularity, availability, and swarm size depending on the seven content type categories. Bundled swarms are more popular and available than single-filed ones in most content types in Figures 5a and 5b. More specifically, bundled swarms of the Movie, Porn, Music, and Game are much more popular and available than single-filed ones, while the popularity and availability of bundled swarms of the E-book, Application, and TV are almost similar or even lower than single-filed ones. We guess this is because many users often prefer to share contents of the E-book, Application, and TV with a single file (e.g., a single pdf file for the E-book, a single installation file for the Application, or the single episode of the TV shows for the TV).

Figure 5c shows that the swarm size of the bundled swarms are much larger than that of single-filed ones in all the content types. Especially, the swarm size of bundled swarms of the Game are around 2.8GB in average because we guess that the volume of games are mostly massive today.

Figure 6 represents the status quo of bundling in terms of bundling pattern (i.e., number of files in a bundle) and user access pattern (i.e., ratio of number of files selected in a bundle) depending on the categories. While bundled swarms of the Music contain much more files than other content types, bundled swarms of the E-book contain a small number of files compared to the other content types as shown in Figure 6a.

Figure 6b shows the CDF of ratio of files selected (or requested) in a bundled swarm. In current BitTorrent, there is a procedure that users can select a set of files among bundled files in a swarm. However, surprisingly, though users have the options to choose the files, most files in a bundled swarm are selected by users in Figure 6b because we believe many BitTorrent users regard a bundled swarm as a single content. While the ratio of number of files selected in Music and Porn are higher than other content types, the ratio of number of files selected in TV is comparatively lower than other content types.

We then examine the correlation of files in a bundled swarm. Here, we assume that if two files are correlated, users requested one of the files will also request other one. Based on this assumption, we calculate the correlation of two files by measuring user access history. Figure 6c shows the CDF of the correlation of files in a bundle. Because users usually make bundled swarms by combining multiple related files, files in a bundle are highly correlated as shown in Figure 6c. While the correlation of files in Music and Porn are higher than other content types, the correlation of files in TV is comparatively less correlated than other types, which are similar results with the ratio of number of files selected in a bundled swarm.

4. DISCUSSION
Inspired from our measurements and findings, we raise and discuss three important research questions in the field of file sharing techniques (or architectures) as well as future contents-oriented networking:

- Our measurement study shows that content bundling has become widely used for contents sharing, which reveals the tendency of users who are likely to share a set of related files (i.e., bundled contents), not just a single file (i.e., single-filed contents). In this sense, what lessons can we learn from our findings? What are the implications on content-oriented networking?

- So far, content bundling in BitTorrent has been done manually and in an ad hoc manner decided by publishers. If bundling is supported systematically in an automatic and efficient fashion, we believe that the system may enhance the availability and download speed. What information can be exploited to make a bundle? What should be considered to make a bundle efficiently?

- Is the current sharing mechanism for bundled contents in BitTorrent efficient? How can we maximize the performance of the bundled contents by taking the correlation of contents and cooperations of users into consideration?

**Implications on content-oriented networking:**
It is widely recognized that a single network address with an identity and a location information may not meet today’s content (or data) intensive application needs. There are several proposed remedies which aim to switch from host-oriented paradigm to content-oriented networking [4, 7, 10, 12]. Most studies have been conducted in naming, resolution and dissemination issues in sharing a single file on the content-oriented network, which assume that a content is regarded as a single file. That is, these content sharing systems presume the *file-oriented* paradigm, which means only a file is considered to search and share efficiently. However, it is questionable that the content in content-oriented networking should be just considered as a single file, because our measurement results show that content bundling in BitTorrent has become widely used for content sharing, which means users may prefer to share not only a file but also a set of related files (i.e., bundle). For example, users often share songs of a recently released album, or series of famous TV shows such as “The Oprah Winfrey Show”. In this case, a bundle of recently released album or a bundle of the TV show can be a content. Therefore, we believe that a bundle (or a set of files) also can be considered as a content in content-oriented networking. In this sense, we raise the question on the problem of *file-oriented* paradigm in content-oriented networking.

To answer the question, we conduct a simple study using 80 bundled contents in our measurement data. We calculate the similarities among files in a bundle in two ways (i.e., *title similarity* and *tag similarity*) to compare with the correlation of files which is described in Section 3. Among multiple candidate criteria to esti-
Correlation / Similarity
CDF
Correlation
Title Similarity
Tag Similarity
Figure 7: Sometimes, a single file has some limitations to identify/classify its content.

mate the file similarity such as file size, category (e.g., movie, TV show or game), and file hash, we first choose the title information to calculate the similarity based on the titles because the title is exactly what users use to search, download, and upload files. To calculate the similarity of file titles, we adopt a popular text classification (natural language processing) algorithm: Levenshtein distance. We also calculate the tag similarity by making 28 tags which contain various attributes (e.g., a director, a distributing agency, a main actor, and etc.) for each file using a search engine, and then we calculate the similarity among files based on the tag information. Figure 7 shows the CDF of two similarities and the correlation of our sample bundled contents. The correlations of 60% of bundled contents are close to 1, which means they are intrinsically the same content in our manual inspects. For example, a single big file can be divided into multiple files, or a video file, a sample file, and an information file of a same movie can be combined into a single content, which are intrinsically same contents. The tag similarity classifies the same content as well as the correlation, which means it can identify whether the files are intrinsically same content. However, the title similarity cannot classify the same content well because it is limited to identify whether the files are intrinsically same content from a single file information such as the title. For the system that need to classify (or identify) the same content such as a caching system in content-oriented networking, the file-oriented view may not be effective. We believe that our findings and lessons learned raise the important research question on the potential problem of file-oriented view in content-oriented networking.

How to make a bundle?: To answer the second question, we are studying bundling strategies to combine multiple files into a bundle considering correlations of files. In our strategies, we exploit two correlations: (1) in-content attributes such as a title, topic, tag, and context of the content, (2) out-of-content attributes such as user access pattern and history to the content. Based on the above correlation information, a bundle can be made. Bundling unpopular contents as discussed in [15] is one example of using an out-of-content attribute.

How to share a bundle?: To address the final question, we are studying a sharing mechanism to maximize the performance of the bundle. When a bundle is shared in BitTorrent, a user can decide whether they receive all the files or some of them which are chosen by users. To accelerate cooperations among users in sharing a bundle, a user can cache and share unwanted files or already downloaded files. In our mechanism, we use a cross-file fashion to give incentive for the cooperations and contributions of users because files in a bundle are mostly highly correlated. This incentive mechanism is similar to the cross-torrent manner in [20].

5. CONCLUSION
Content availability is a serious problem in today’s peer-to-peer swarming systems. Menasche et al. has shown that bundling is a promising solution to mitigate this availability problem, but there are still little empirical grounding for understanding, describing and modeling it. This is the first empirical study that measure and analyze how prevalent contents bundling is in BitTorrent and how peers access the bundled contents, in comparison to the single-filed ones. To our surprise, we found that bundling has become widespread for contents sharing. Inspired from our findings, we raise and discuss three important research questions in the field of file sharing techniques/architecture as well as future contents-oriented networking: i) bundling strategies, ii) bundling-aware sharing systems in BitTorrent, and iii) implications on content-oriented networking.

6. REFERENCES