Exploring Cyberlockers Content

Nan Zhao
Télécom ParisTech
Paris, France

Loïc Baud
DREV, Hadopi
Paris, France

Patrick Bellot
Télécom ParisTech
Paris, France

Abstract

As the bandwidth of Internet rises by ISPs, the proportion of different Internet traffics and underlying used service has changed. There is about 20% decrease of the P2P traffic compared to an increase of more than 10% of the traffic of direct file sharing service through cyberlockers. In this paper we present a recent study over four cyberlockers: Rapidgator, Speedyshare, 1Fichier and Megashares. Compared to prior studies, we apply a bias-free sampling method to randomly gather hosted files on the four cyberlockers. We aim at giving a statistic study to find out the characteristics of the hosted files on cyberlockers. In our work, we analyse and estimate the total number of files and the total size of files on the four cyberlockers. We specifically discuss the size and number distributions of hosted files in file format and file content classifications. Our results show that different cyberlockers have different number and size distributions, but on most cyberlockers split- compressed files and uncompressed files take a relative large part of the volume. Additionally, the content classification analyse results show that users on different cyberlockers have different usage preferences. Rapidgator inclines to entertainment usage; Speedyshare and 1Ficher incline to personal and entertainment usage; Megashares inclines to entertainment and professional usage. And most of them like to upload series files.

1. Introduction

Cyberlockers are also referred as One-Click Hosting, which is a kind of Web services for file hosting and file sharing over the Internet. Cyberlockers allow Internet users to easily upload one or more files from users' local hardware devices as computers, tablets, smartphones, etc. to a remote hosting server with only one click. In return, cyberlockers generate a URL for the uploaded file. The cyberlockers users can keep this URL for them or share it with their friends, and even can publish it on some websites such as forums.

From 2005 as the popularity of cyberlockers have increased [1], the proportions of Internet traffic has drastically changed. In Ipoque Internet traffic study 2007 [2] and 2008/2009 [3], they pointed out that the cyberlockers service traffic has increased at least 10% in Eastern Europe and Germany, while in Southwestern Europe it has increased more than 20%. Meanwhile P2P that still takes the largest proportion in the Internet traffic had a decrease of at least 10% from 2007 to 2008/2009. This shows that Internet users' preference is changing from the P2P to the One-Click Hosting service, such as cyberlockers. Compared to P2P service, cyberlockers service does not depend on the number of available seeds to guarantee the transaction of files. Once a file is uploaded to the server, it is available all the time until it is removed. Cyberlockers service is based on the HTTP protocol, or in some case is FTP protocol, so it is easy to download files and there is no limit over downloading debits for premium users. What is more, the IP addresses of uploaders and downloaders are only known by cyberlockers service. In that case it is difficult to inspect the IP addresses of the devices connecting with cyberlockers sites via the third-part softwares [4]. These conveniences above explain why the Internet users' preference changes, which brings the rise of the number of cyberlockers.

As this change arises a lot researchers' attention, there exists several studies over the cyberlockers service. However, those prior studies only monitored and collected the user-end traffic. Their result is skewed to the downloading users behaviours, which cannot show the diversity of different files hosted on cyberlockers, and cannot give a comprehensive file size and file type distribution on cyberlockers neither. Therefore, our study aims to take a statistic study over the content stored on cyberlockers in order to figure out the general and representative characteristics for the

content on the cyberlockers. We perform a random file sampling from four cyberlockers: Rapidgator, Speedyshare, 1Fichier and Megashares. With retrieved information of each file, we analyse the properties of content on cyberlockers via two different classification systems, which are file form and file content classifications. We find that there are more small-size and medium-size files than the files bigger than 500 MB on cyberlockers. However, it is long-tailed for the files larger than 1 GB. Additionally, we also conclude the different usages for the four cyberlockers. Rapidgator is mostly for entertainment usage. Speedyshare and 1Fichier are mostly for entertainment and personal usage. And Megashares is mostly for entertainment and professional usage. The highlights of our work could be summarised as below:

- The analysed files in our study are sampled directly from the four one-click hosting servers. Compared to tracing the user-end HTTP traffic in the prior studies, the sampled files in our study are more random, more representative and bias-free.
- Different form [5], we use two different file classification systems, which are file form (split, single archived file and non-archived file) and file content (audio, document, picture, software and video) classifications to analyse the characteristics of files hosted on the cyberlockers separately. For the file content classification, we further divide the five types into detailed sub-types in order to have a deeper analysis over the file content properties.
- Our study reveals that there are about 18.7 millions accessible files on Rapidgator, about 2.37 millions accessible files on Speedyshare, about 2.32 millions accessible files on 1Fichier and about 697 millions accessible files on Megashares. The possible total amounts of files size on Rapidgator, Speedyshare, 1Fichier and Megashares are 6.93 PB, 0.18 PB, 1.0 PB and 177 PB separately.
- Both on Rapidgator and 1Fichier there are 60% of files no larger than 340 MB. On Speedyshare there are more than 60% of files no larger than 10 MB. Compared to the other three, on Speedyshare most of hosted files are small files. However, all the four cyberlockers are long-tailed for the files larger than 1 GB.
- More than 50% of files hosted on Rapidgator, 1Fichier and Megashares are archive files (both single and split). On Speedyshare, the archive files only take about 38%. The size of single-archived files mostly varies around 10 MB, 110 MB and 370 MB on the four cyberlockers. The size of split-archived files mostly varies around 10 MB, 110 MB, 210 MB, 270 MB, 530 MB and 1GB. For non-archived files, the file size varies around 10 MB, 370 MB, 730 MB. From those large files, it can be inferred that cyberlockers provide high bandwidth for file transfers.
- Video is not always the most hosted content type on cyberlockers, which is opposite to the result in [5-6]. On Speedyshare the number of video files only takes 25%. However, the volume of the video files does take the largest portion on cyberlockers. And the size of video files almost always varies around 110 MB, 190 MB, 210 MB and 370 MB.

The rest of the paper is organized as follows: the section II talks about the prior studies over the measurement and the traffic flows of the cyberlockers service. We describe our study method in details in the section III. Then we analyse the sampled files in order to reveal the characteristics of cyberlockers in the section IV. In the last section V, we give the statements for this study and indicate the direction of our future work about the cyberlocker service.

2. Related Work

Cyberlockers turned popular over the last ten years, so there are not many studies published yet. The most known is Antoniades, Markatos and Dovrolis' paper [4]. In their work, they traced "client-side" traffic from a research centre and a campus networks. They show that cyberlockers traffic volume surpasses that of popular video sharing service such as YouTube. They point out that the large objects as movies and softwares are often split into 100 MB or 200 MB archives for transfers. This is mostly caused by the maximum upload size limits on free users. They identify the traffic volume between free and premium users. There are more free users than premium users. In their work, they also point out that the cyberlockers service does not depend on the web cache, instead, it focus on large transfer and are less sensitive to delay.

Mahanti's work in [5-6] is based on monitoring the "client-side" traffic of a campus network. Besides this, [5-6] also crawled a cyberlockers search engine to get the published files. Mahanti and his colleagues give a detailed file hosting ecosystem study over RapidShare, Megaupload, zSHARE, MediaFile and Hotfile. By observing the downloaded traffic size, they point that most files are no larger than 100 MB. They also give the proportions of different content downloaded from the cyberlockers. They state that on several cyberlockers, the content of archives takes more than

80% of downloaded volume. And content as documents, audios, applications are just around 5%.

In Envisional' report [7], they crawled the Internet for collecting URLs pointing to cyberlockers in order to estimate the percentage of copyrighted material exchanged via cyberlockers. They list that most of files hosted on cyberlockers are movies, which is 35.8%. The remaining of the content hosted on cyberlockers are pornographic material (20.1%)., music (10.1%), software (10.0%), game (9.4%), TV shows (8.5%) and eBooks (2.6%). Over 90% of analysed files have copyright issues.

However, all the prior studies either monitored the "user-side" HTTP traffic, or crawled the forums with published cyberlockers links. Thence, their methods are biased to the users' preference, which only can show what kind of files are popular for the cyberlockers users, but cannot represent the general characteristics of cyberlockers. And in the prior studies, the classification of content type is not clear. As in [5-6], they take archives as a type of file content, but there is a correlation between other types such as video, audio etc., and archives. This correlation makes the content type distribution in their study ambiguous. Additionally, in [8] the conception of Deep Web and Surface Web is mentioned. Most methods used in prior studies are biased to Surface Web, which are the published static pages and links to other pages. This surface web can be crawled or spidered. However, the Deep Web is made of pages created dynamically as the result of a special search, which cannot be retrieved directly. In [8] it mentions that the size of Deep Web is about 500 times larger than that of Surface Web, which means with crawling and probing method in the prior studies, the result is focused on the Surface Web and not enough representative and comprehensive.

The study in this paper is designed to avoid the bias caused by sampling over the Surface Web. It therefore:

- Takes into account the Deep Web, in order to dig out the hidden information stored on the cyberlockers;
- Gives up using Internet monitors over a specific location to collect cyberlockers information, in order to avoid the bias of users' behaviours.

3. Methodology

In this section, we give the methodology that we used in our cyberlockers study.

3.1. File Sampling Methodology

Once a user uploads a file on a cyberlocker, the server generates a URL for this file. This URL allows anyone to access the location where the file is hosted on the cyberlocker. And this URL points to a webpage, which displays basic information of the file such as its name, size and some optional information designed by the cyberlocker as file description, file upload date etc. Table I shows the composition of URLs on Rapidgator, Speedyshare, 1Fichier and Megashares.

Table I. The Form of Cyberlockers URL

CLS	URL		
	URL Form	***Id L***	***Id element***
Rapidgator	http://rapidgator.net/file/<id>	8	0-9[1]
Speedyshare	http://www.speedyshare.com/<id>	5	0-9, lowercase and uppercase letters
1Fichier	http://<id>.1fichier.com/	6	0-9, lowercase letters
Megashares	http://d01.megashares.com/index.php?d01=<id>	7	0-9, lowercase and uppercase letters

From Table I, we can deduct that the URLs of each cyberlocker can be composed as "prefix+<id>", which "prefix" is the fixed information of cyberlocker server, and "<id>" is a sequence of randomly generated alphanumeric. Therefore, we would like to sample the files hosted on cyberlockers based on randomly generating cyberlocker URLs. In our experiment, we designed an ID generator on JAVA, which randomly combines ID elements to compose a possible URL for each cyberlocker. Once the URL is generated, we verify whether this URL points to an existing file on its cyberlocker or not. This method can be summarised as following:

a) Generate a random possible file URL for a cyberlocker.

b) Verify whether URL exists on the cyberlocker sever or not. If it does not exist, repeat a). If it does exist, continue to c).

c) Retrieve the file's information of file name and file size.

As shown in Table I, we can calculate the possible name space of URLs for Rapidgator Speedyshare,

[1] We sampled Rapidgator on January 2013. At that time they still applied the old id form. Now they have changed with a new form made up by 30 letters and numbers.

1Fichier and Megashares, which are $10^{!}$, $62^{!}$ ($9.2\times 10^{!}$), $36^{!}$ ($=2.2\times 10^{!}$) and $62^{!}$ ($=3.5\times 10^{!"}$) respectively. Based on this large quantity of possible URLs sets of each name space, this could avoid the repeatability of generating the same URL. And comparing to crawling over the forum sites, it also avoids the content bias to the users' preference and habit. For the number of files to be sampled on the four cyberlockers, (1) shows a classic relation between the size n of the sampled set, the desired level of precision e and the confidence interval [9-10]. In order to get a confidence interval of 95% and a desired level of precision smaller than 0.03, the sampled number should be 1200. Thence for each cyberlocker, it has to collect 1200 files.

$$n = \frac{! \times !(!!!)}{!!} \quad (1)$$

Our sampling process is running over network TOR with many different IP addresses located in different countries. Table II below shows the data collection on the four cyberlockers. In the end of our experiment, we sample 1200 files on each cyberlocker.

Table II. Data Collection

Type	Collected Date	Generated URLs
Rapidgator	January 2013	13,355
Speedyshare	May 2013	598,717
1Fichier	May 2013	1,128,574
Megashares	May to June 2013	6,061,271

3.2. File Analysis Methodology Estimate the File Number and the File Size.

We already calculate name space $N_{!"}$ of each cyberlockers in the section A, which are $10^{!}$ for Rapidgator, $62^{!}$ for Speedyshare, $36^{!}$ for 1Fichier and $62^{!}$ for Megashares. And we take $N_{!}$ as the number of sampled files, $N_{!}$ as the number of totally generated URL links. We can estimate the number of hosted files N on each cyberlockers with (2).

$$N = N_{!"} \times N_{!} / N_{!} \quad (2)$$

$$S = S_{!"\#\$} \times N \quad (3)$$

We then use (3) to estimate the total volume of the hosted files on each cyberlocker $S_{!"\#\$}$, which is the average size of sampled files and N is the result in (2).

File Classification. In [5-6], they mixed file form and file content for the statistic study of different file types. In order to avoid this, in our study we applied file form classification and file content classification respectively to analyse the file characteristics. The file form is based on whether a file is a compressed file or not. The file content is based on the different content types. The following describes the two classifications.

File Form Classification

Single Archive Files: Compressed files that are not split. They are normally classified according to the file extensions such as .zip, .rar and .7z, and where there is no splitting information in the file title.

Split Archive Files: Compressed files that are split. They are normally classified according to the file extensions such as .zip, .rar and .7z, and where is splitting information in the file title such as "part".

Raw Files: Regular files. File extensions that do not equal to .zip, .rar and .7z.

File Content Classification

Audio: Files corresponding to music, concert and other audio record. They are normally classified according to the file name or the file extension such as .mp3, .ma4, .wav, .flac, etc.

Document: Files corresponding to eBooks, magazines, all document formats and programming code. They are normally classified according to the file name or the file extension such as .txt, .pdf, .doc, .xml etc.

Picture: Files corresponding to all image formats. They are normally classified according to the file name or the file extension such as .jpg, .bmp, .jpeg etc.

Software: Files corresponding to software, executable files and video games. They are normally classified according to the file name or the file extension such as .exe etc.

Video: Files corresponding to videos. They are normally classified according to the file name or the file extension such as .avi, .mov, .mkv, .mp4 etc.

Others: Files that cannot set with any of the content types above. This is caused by non-sense names of the sampled files.

For the second classification, in order to better understand the files content of each type and also in order to have a detailed distribution of each content type, we divide each content type into several sub-types. The Table III shows the detail of sub-types of each content type.

Table III. The Sub-Types of Each Content Type

Type	Sub-types Name
Audio	Music-Album Full/Part, Music-Song, Music-Clips, Music-Concert, Others
Document	Book, Magazine, TextFile, Code
Software	Software Full/Part, Video Game Full/Part
Video	Film Full/Part, Film-Porn Full/Part, Film-Animation Full/Part, Series Full/Part, Series-Animation Full/Part, Media Full/Part, Amateur, Tutor Full/Part, Others

4. Result Analysis

In this section, we analyse the sampled files on the four cyberlockers in order to find the general characteristics of cyberlockers and some specific properties for each cyberlocker.

4.1. Files Number and File Size

Firstly we would like to estimate the number of files hosted on each cyberlockers and the total volume of each cyberlocker. In the section III, we show the equation to estimate the total files number from the tested URLs and the sizes of the name space of this cyberlocker. And Table II shows the number of tested URLs on each cyberlocker. Thence according to (2), we get the estimated number of files hosted on Rapidgator, Speedyshare, 1Fichier and Megashares is 18.7 million, 2.37 million, 2.32 million and 697 million respectively. With (3), we can estimate the total volume on each cyberlocker. The average sizes of the sampled files are as below: Rapidgator 370 MB, Speedyshare 75.6 MB, 1Fichier 429 MB and Megashares 254 MB. Therefore, the estimated volume of Rapidgator, Speedyshare, 1Fichier and Megashares is 6.93 PB, 0.179 PB and 0.993 PB and 177 PB respectively.

However there are some points that should draw attention:

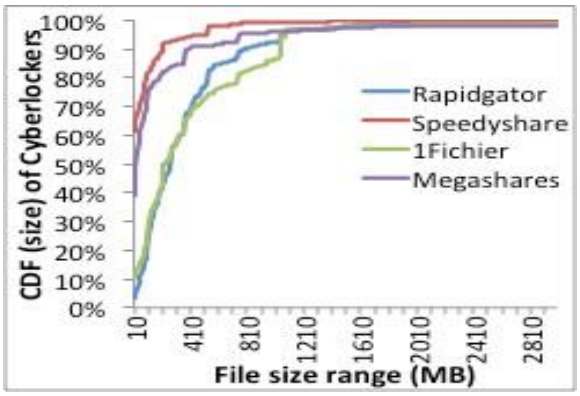
- The same file can be posted several times by different users and have different URLs. Therefore, there is content repetition in the hosted files on cyberlockers.
- There exist different versions for the same file content.
- The split archive files take a relative large part, which means that there are many content files for decompressing one useful file.
- The expired files are deleted by the cyberlockers.

Thence, the real file number and file size should be smaller than these calculated values.

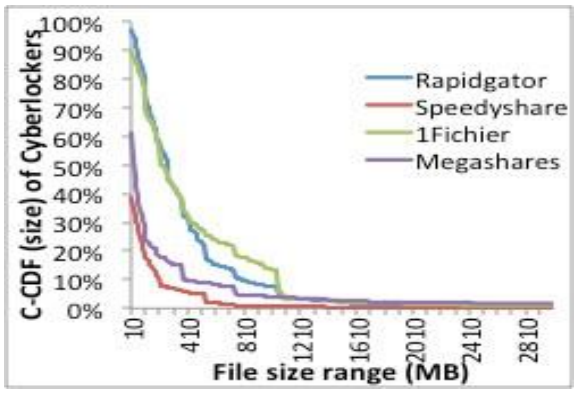
4.2. File Size Distribution

In this section we analyse the file size distribution on the cyberlockers. Fig. 1a and Fig. 1b are the Cumulative Distribution Function (CDF) and Complementary CDF (C-CDF) of files size sampled from the four cyberlockers. From Fig. 1a we can see that on Speedyshare there are about 60% of files whose size is no larger than 10 MB. On Megashares, there are about 40% of files whose size is no larger 10 MB. However this value is much smaller on Rapidgator and 1Fichier, which is 3.17% and 10.32% respectively. On Rapidgator, we can find that the curve of CDF increases fast for the file size no larger than 530 MB, which takes 82% of all Rapidgator files. And there are about 60% of files no larger than 330 MB. On Speedyshare, we find that its CDF increases largely for the size no larger than 140 MB, the remaining part increases slowly. Additionally, we find that there are about 80% of files on Speedyshare that are no larger than 140 MB. There are two inflection points on the CDF of 1Fichier. The first one is at 370 MB, which takes 65% of files. The second one is at 1 GB that from 1 GB to 1.05 GB the rise is about 8%. The increasing of the curve of Megashares is similar to that of Speedyshare. From 10 MB to 190 MB, the rise is about 40%. Then the curve increases slowly, from 190 MB to 530 MB, the CDF only increases 10%.

Fig. 1b shows C-CDF of the four cyberlockers. We can see the proportion of size larger than 1 GB on Rapidgator, Speedyshare, 1Fichier and Megashares is 7.5%, 0.75%, 13.57% and 3.92%. For the proportion of file size larger than 2 GB, Rapidgator is at 0.92%, 1Fichier is at 1.33% and Megashares is at 2.17%.



a.CDF of the files size on the four cyberlockers



b.C-CDF of the files size on the four cyberlockers

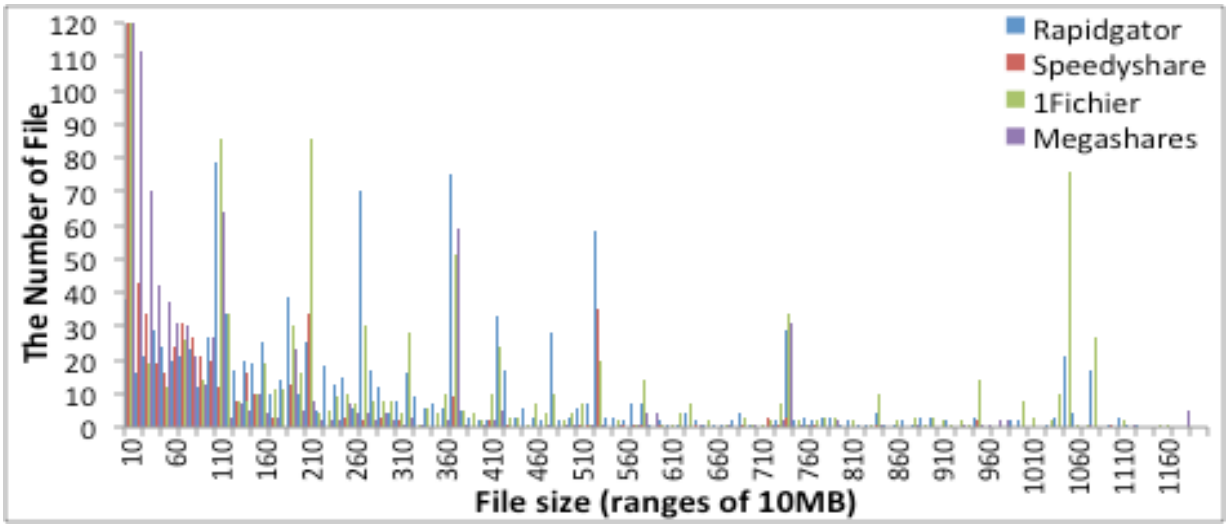


Figure 1. c.The distibution of the number of files with different size on the four cyberlockers

We mentioned in the last section that the average size of Rapidgator, Speedyshare, 1Fichier and Megashares are respectively of 370 MB, 75.6 MB, 429 MB and 254 MB. With the result of Fig. 1a, we can firstly conclude that there are a high proportion of small files hosted on Speedyshare. On Rapidgator, there are more than 30% files between 300 MB and 700 MB; on Megashares, there are more than 80% files no larger than 300 MB, which leads to a medium average value on the two cyberlockers. On 1Fichier, between 1 GB and 1.05 GB, the proportion is as high as 8%. That explains why the average size on 1Fichier is larger than the others. This also shows that there are many big files hosted on 1Fichier. With the result of Fig. 1b, we can deduct that the files size distribution of the four cyberlockers is long-tailed for the large files. Especially on 1Fichier and Megashares, the files size is also heave-tailed.

Then we take a look at the peaks in different file sizes on the four cyberlockers. Fig. 1c is the distribution of the number of files with different sizes. On Rapidgator, we can see most file sizes drop before 530 MB, especially at 10 MB, 110 MB, 120 MB, 190 MB, 270 MB, 370 MB, 420 MB and 530 MB. On Speedyshare, most file sizes drop before 210 MB, especially at the point of 10 MB, the peak value is 463[2]. The other peaks on Speedyshare are at 20 MB, 30 MB, 210 MB and 530 MB. On 1Fichier, most of file sizes drop before 740 MB. The peaks exist at 10 MB, 110 MB, 120 MB, 190 MB, 210 MB, 270 MB, 370 MB, 740 MB and 1.05 GB. On Megashares, most of file sizes drop before 110 MB. The peaks are at 10 MB with a value of 469, 20 MB, 30 MB, 40 MB, 50 MB, 60 MB 70 MB, and 110 MB. There are also two peaks larger than 110 MB, which are at 370 MB and 740 MB. From the result of Fig. 1c we find there are many peaks of files in common on the four cyberlockers. Thence in the following section C and D we try to figure out what those peaks represent.

[2] In order to better observe the file sizes, which do not have many files, we take 120 as the maximum of the y-axis.

4.3. Files Form Analysis

In this part we analyse the files hosted on cyberlockers by applying the File form classification. Table IV and Fig. 2 show the file number and file size proportions of the different file forms on Rapidgator, Speedyshare 1Fichier and Megashares. We first take a look at the file number distribution. On Rapidgator the proportions of single and split archive files are almost the same, which is about 10% higher than that of raw files. On Speedyshare raw files take more than 60% of all files, while split archive files take no larger than 10%. On 1Fichier raw files take more than 40% of all files, split archive files take about 38%, while the single archive files take the smallest part. On Megashares, it is opposite to 1Fichier, which both raw files and single archive files take more than 40% of all the total file number while split archive files take about 12%. We then take a look at the file size distribution. Among the four cyberlockers, except Megashares, split archive files take the largest portion of the total file size on the other three cyberlockers. While on Megashares, raw files take more than 80% of the total size. Both single and split archive files on Megashares take less than 10% of the total size. Additionally, raw files on Rapidgator and 1Ficher also take a larger proportion of the total file size than that of single archive files. While on Speedyshare it is single archive files, which take the second large part of the total file size.

Table IV. Proportions of File Number and File Size in File Form Cliassificatio of the Four Cyberlockers

CLS	Single Archive Files		Split Archive Files		Raw Files	
	N[3]%	*S[4]%*	*N%*	*S%*	*N%*	*S%*
Rapidgator	36.6	24.4	37.3	43.2	26.1	32.4
Speedyshare	28.2	32.9	9.7	38.0	62.1	29.1
1Fichier	15.4	10.5	38.6	50.1	46.0	39.4
Megashares	40.3	9.8	12.3	5.4	47.4	84.8

[3] N here represents the percentage of file number.
[4] S here represents the percentage of file size.

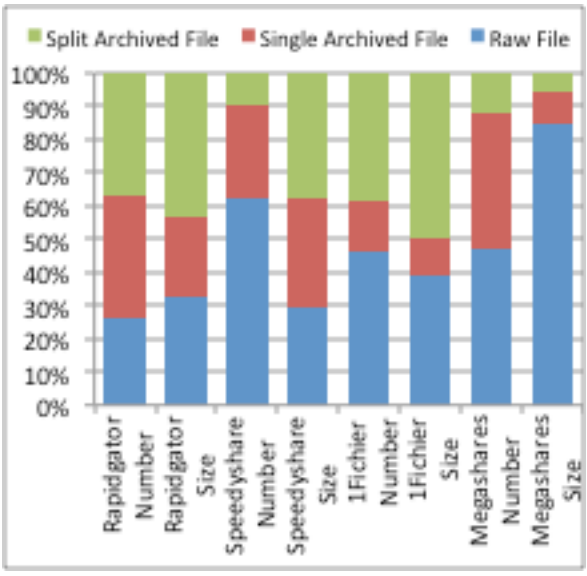


Figure 2. The distributions of file number and file size in different file forms on the four cyberlockers

With the information shown in Table IV and Fig. 2, we can tell that there is not a common file size and file number distribution in the file form classification. On most of the four cyberlockers, there are not many split archive files but raw files do take a relative large part of the file number distribution. For file size distribution, we can conclude that except on Megashares, on the other three cyberlockers, the split archive files take the largest portion in the file size distribution, which the biggest volume is taken by raw files on Megashares. From these two distributions, we can infer that most of single archive files may have a small size, and most of split archive files could have a medium or a larger size, which may be caused by the different limits on uploading files or caused by the different user behaviours on the four cyberlockers.

In the second part we take a look at the distribution of the number of files with different size via file form classification. Fig. 3 shows this distribution on the four cyberlockers. Compared with Fig. 1c, on Rapidgator the peak of 10 MB is from single archive files, the peaks of 110 MB and 120 MB are from single and split archive files. The peaks of 190 MB, 270 MB, 420 MB, and 530 MB are from split archive files. The peak of 370 MB is from single archive and raw files. On Speedyshare, the peaks of 10 MB and 30 MB are from single archive and raw files. The peaks of 210 MB and 530 MB are from split archive files. And the peak of 20 MB is from raw files. On 1Fichier the peaks of 10 MB, 190 MB, 370 MB, 740 MB are from raw files. The peaks of 110 MB, 120 MB, 210 MB, 270 MB and 1.05 GB are from split archive files. And on Megashares, the peak of 10 MB is from single archive files, split archive files and raw files, which number is 265. The peaks of 20 MB, 30

MB, 40 MB, 50 MB, 60 MB and 70 MB are from single archive files. The peak of 110 MB is from split archive files. And the peaks of 370 MB and 740 MB are from raw files.

From this we can see that on most cyberlockers, it is the split archive files that cause the most peaks, except on Megashares, which are the single archive files that cause the most peaks. We also find that the single archive files always a small size no larger than 110 MB, which mostly is between 10 MB to 30 MB. This explains why single archive files take a relative larger number portion but with a smaller size portion. And split archive files have several common sizes on different cyberlockers, which are 110 MB, 120 MB, 210 MB, 270 MB and 530 MB. Additionally, on 1Fichier, the split archive files are long-tailed, which is around 1.05 GB. This shows that split archive files normally have a medium or a large size, which explain why on three of the cyberlockers split archive files take the largest size portion. From the four cyberlockers we can find that the size of raw files are mainly around 10 MB, 370 MB and 740 MB. The large size of raw files and split files show that cyberlockers servers have large tolerance for the uploading size, which allows users to upload files larger than 1 GB. However, there may exist a bandwidth limit for the free users that explains why there are many medium-size split archive files on cyberlockers. We also can tell that the compression tools chosen by users normally split large files to archives around 100 MB, 200 MB, 270 MB or 500 MB.

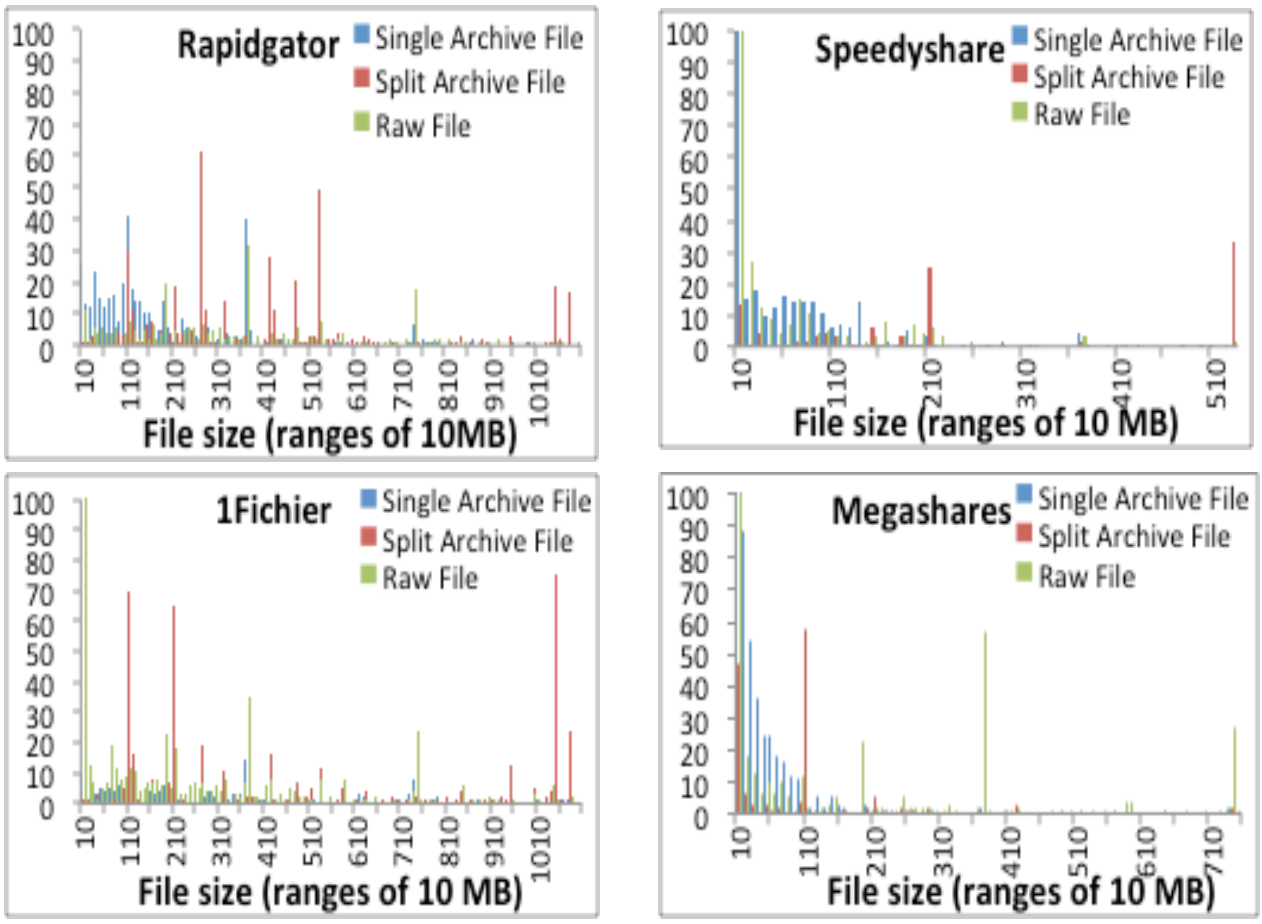


Figure 3. The distribution of the files number with different sizes on the four cyberlockers by file form classification

4.4. File Content Analysis

In this section we analyse the files hosted on cyberlockers by file content classification. Both Table V and Fig. 4 show the file number and the file size proportions of different content types on the four cyberlockers. First we take a look at the file number distribution. We can find that both on Rapidgator and 1Fichier the video files take more than 60% of all the files. On Speedyshare, the proportions of all the content types are no larger than 30%. The two largest content types are picture files with 24.50% and video files with

24.08%. The file number distribution on Megashares is similar to that on Speedyshare. The largest portion is software content with a percentage of 33.50%. And the second one is video content with a percentage of 32%. However, for the file size proportion, it is video files that take the largest part on all of the four cyberlockers. As the largest content type on Megashares, software file just take a proportion of 20% over the total file size. While on the other three cyberlockers, software files take about 10% of the total size. For audio content type, on all the four cyberlockers, we also find that the file number portion is larger than its file size portion.

From the results of Table V and Fig. 4, first we can tell that on different cyberlockers the distributions of the content types are different. This shows the different users' preferences on the four cyberlockers. And cyberlockers do not mainly store the video content, for example the most hosted content type on Megashares is software, and on Speedyshare is picture. As the video content take the largest proportion of file size distribution on all the four cyberlockers, we can suppose that video files mainly have a medium or a large size. And audio files hosted on the cyberlockers are mainly the small-size files, which means probably most of audio files have a low compression quality.

Table V. Proportions of File Number and File Size in File Content Cliassificatio of the Three Cyberlockers

CLS	Audio	Doc[5]	Oth-ers	Pic-ture	Soft-ware	Video
Rapid[6]N[7]	20.4	3.4	3.3	1.8	6.8	64.3
Rapid S[8]	9.3	1.0	2.6	0.2	9.2	77.7
Speed N	17.8	15.6	6.5	24.5	11.5	24.1
Speed S	14.5	3.5	7.3	0.7	9.8	64.2
1Fich-ier N	3.9	8.8	15.3	1.5	8.9	61.6
1Fich-ier S	1.2	0.3	16.8	0.1	11.7	69.9
Mega N	12.8	13.00	3.4	5.3	33.5	32.0
Mega S	3.0	0.5	1.8	0.1	20.0	74.6

[5] Doc here represents the content type Document.
[6] Rapid is short for Rapidgator, Speed is short for Speedyshare and Mega is short for Megashares.
[7] N here represents the percentage of file number.
[8] S here represents the percentage of file size.

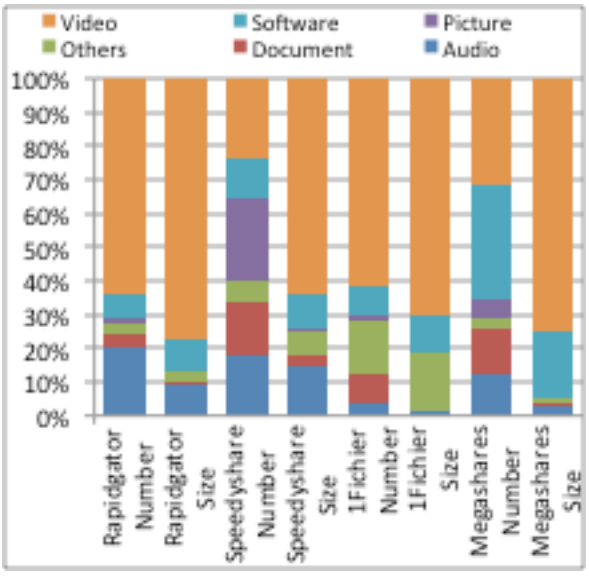


Figure 4. The distribution of file number and file size of different content types on the four cyberlockers

Then we take a look at the distribution of the number of files with different sizes via file content classification. Fig. 5 shows this distribution on the four cyberlockers. Compared with Fig. 1c, we can figure out the peaks in different content types of each cyberlocker. First of all, from Fig. 5 we can tell that on the four cyberlockers it is the video content that causes most of the peaks. On Rapidgator the peak of 10 MB is from document and picture files. The peaks of 110 MB and 120 MB are from audio and video files. The remaining peaks of 190 MB, 210 MB, 270 MB 370 MB, 420 MB and 530 MB are all from video files. On Speedyshare all the content types are numerous at the peak of 10 MB, especially for picture content, which has 290[9] files no larger 10 MB. The peak of 20 MB is from audio files and the remaining peaks of 210 MB and 530 MB are all from video files. On 1Fichier the peak of 10 MB is from document files. The peaks of 110 MB and 210 MB are from others files and video files. And all the remaining peaks of 120 MB, 190 MB, 270 MB, 370 MB, 740 MB and 1.05 GB are from video files. On Megashares, all the content types are also numerous at the peak of 10 MB, especially for software content, which has 214 files no larger 10 MB. The peaks from 20 MB to 40 MB are from software files. The peaks of 50 MB and 60 MB are from audio files and software files. And the remaining peaks of 110 MB, 370 MB and 740 MB are from video files.

Fig. 5 confirms our supposition that video files normally have large size. Compared with the result in section C, we can infer that the files of 210 MB, 270

[9] In order to better observe the file sizes, which do not have many files, we take 70 as the maximum of the y-axis.

MB, 420 MB, 530 MB and 1.05 GB mostly are part compressed video files. The files of 370 MB and 740 MB are mostly the completed uncompressed video files. Software content is always stored in the complete compressed files between 10 MB and 60 MB. And for audio content, it exists complete compressed and uncompressed files of 10 MB, 50 MB and 60 MB. And it also exists complete compressed files of 110 MB and 120 MB for audio. For document and picture content, normally are the complete compressed or uncompressed files no larger than 10 MB. From this result, first we can tell that video content is stored mostly as uncompressed files or part compressed files on cyberlockers. The various sizes of video content let us infer that there exist different encoding formats for the video files. Those with the size larger than 1 GB are probably of high definition as Blu-ray. For the video content, in the beginning of the section, we suppose that most audio files are low-compressed. After this further analysis, we can tell that there are also high-compressed audio files.

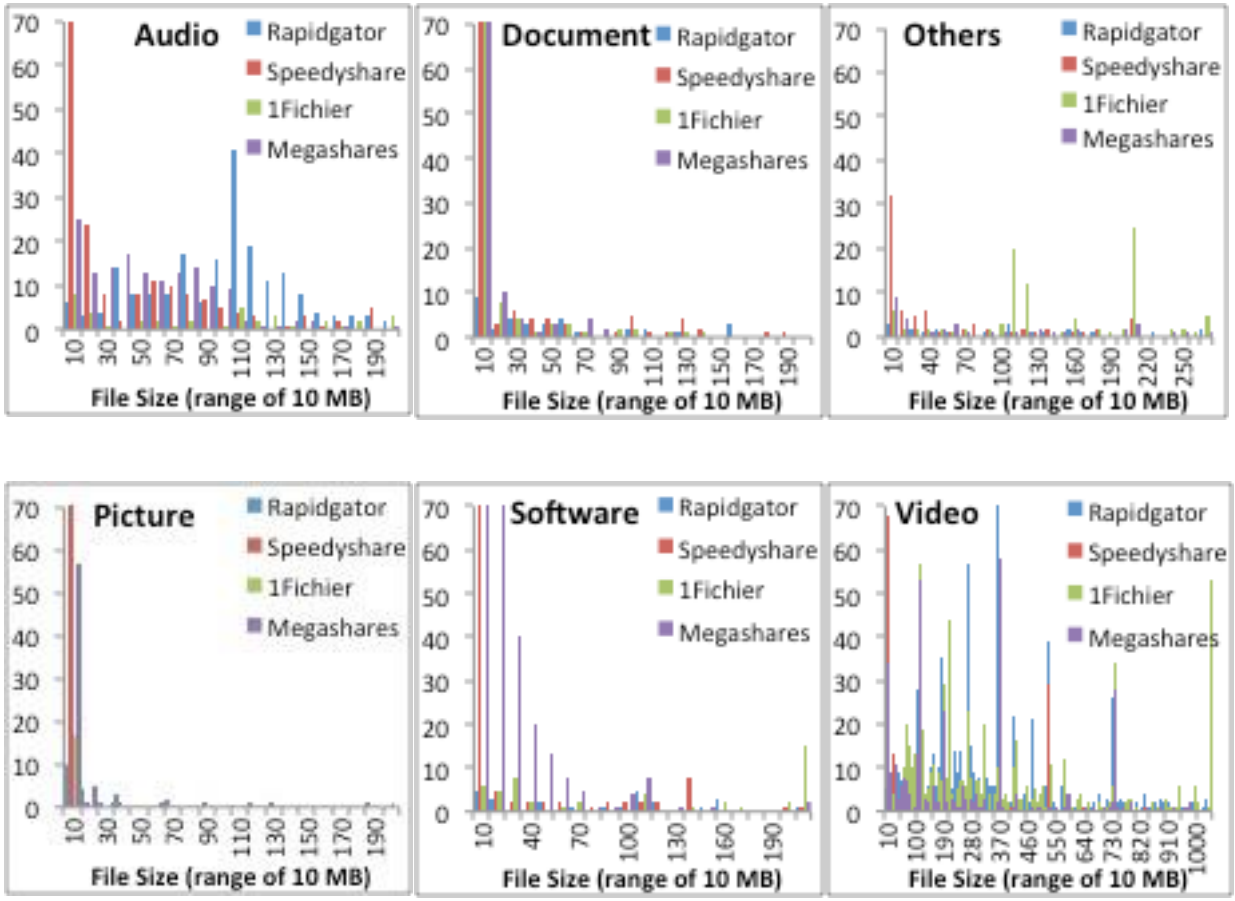


Figure 5. The distribution of the number of files with different sizes in each content type on the four cyberlockers

In the third part, we take a look at the file number and file size distribution in sub-content types. Fig. 6 shows the file number and file size distributions of each sub-type. On Rapidgator the files of music album, full and part of pornographic film and series take the bigger proportions of the total file number. For the file size distribution, the files of full and part film, full and part pornographic film and series have the most parts of the total file size. On Speedyshare files of picture take about 24% of all the total file number. While files of music album, part film, part video tutorial and series take the most part of the total size. Especially for the part video tutorial files, they are only 3% of the total file number but take more than 18% of the total size. On 1Fichier the files of others, part others, part film, series and series animation have evident high proportion of the total file number. And files of part video game, full and part film and series take the most part of the total size. On Megashares, the files of music album, book, video game, software and series take the high proportions of the total file number. And files of video game, film and series take the most part of the total size.

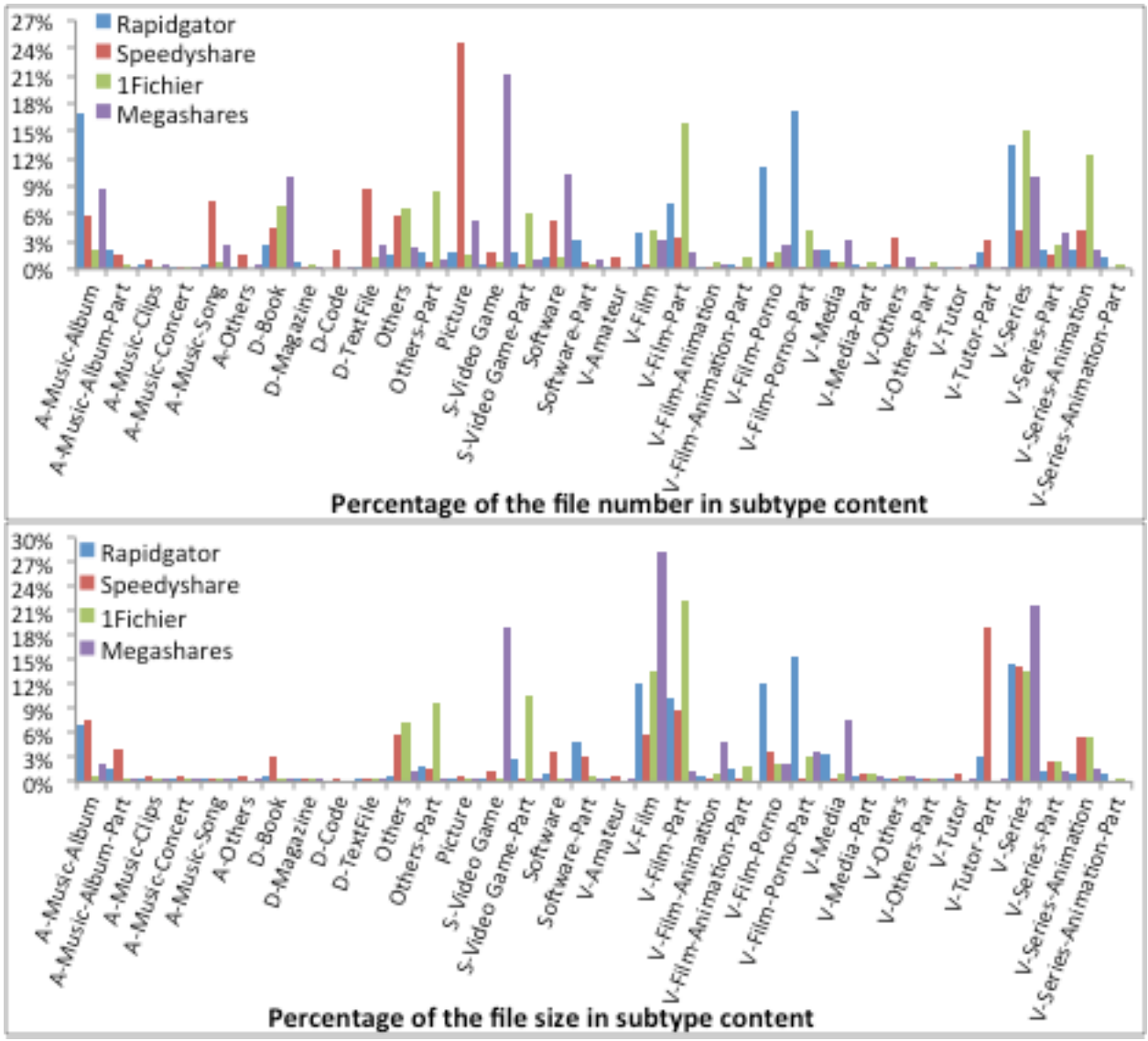


Figure 6. The file number and file size distribution sub-content types on the four cyberlockers

From Fig. 6 we can deduct that most files in video content are not necessarily films, there are also many series in video content. And normally the uploaded films and series files on cyberlockers have a medium or a large size. Additionally, Fig. 6 can also show us the different behaviours of the content generators on the four cyberlockers. Rapidgator users like to upload music, film, series and pornography. Speedyshare users like to upload songs, papers and pictures. 1Ficher users like to upload books, films, series and animations. As on 1Fichier the content of others takes a relative large part both in the number and size distributions. We can infer that those unclassified files could be unpublished personal files, for which users just take the server to store them. And on Megashares, users like to upload video game, software and series. Furthermore, with the content generators' behaviours, we can infer the different preferences of usages for the four cyberlockers. Rapidgator is mostly for the entertainment usage. Speedyshare and 1Fichier are mostly for entertainment and personal usage; Megashares is mostly for the entertainment and professional usage.

5. Conclusion and Future Work

In this paper we present our recent work about the characteristics of the hosted files on cyberlockers. We applied a statistic method to study the content properties on Rapidgator, Speedyshare, 1Fichier and Megashares. We estimated the total file number and the total file size on the four cyberlockers. There are more than 2 millions files hosted on Speedyshare and 1Fichier, more than 18 millions files hosted on Rapidgator and more than 690 millions files hosted on Megashares. The

possible total size is 6 PB on Rapidgator, 0.18 PB on Speedyshare, 1 PB on 1Fichier and 177 PB on Megashares. Then we focused on analysing the file size and file number distribution in different file form and file type. We find that the split-compressed files do not take the largest portion of the total file number, while they do really take the largest portion of the total file size on the other three cyberlockers except Megashares. Compared to prior studies, we find that there exist a relative high proportion of raw files on the four cyberlockers. Especially on Megashares, the raw files take the largest part both in file number and file size distributions. We also find the correspondences between the file form and the file content. We infer that files of 370 MB and 740 MB mostly are uncompressed video files. Files of 210 MB, 270 MB, 420 MB, 530 MB and 1.05 GB mostly are part compressed video files. Files no larger than 60 MB could be compressed softwares. In our study, we also find that the different cyberlockers have different user behaviours. The distribution of the sub-types of content uploaded by users is not same. But at least, it seems that users of the four cyberlockers prefer to store the TV series on the cyberlockers. With the content type study we also can conclude that Speedyshare and 1Ficher are mostly for entertainment and personal usage. Rapidgator is mostly for entertainment usage, Megashares is mostly for entertainment and professional usage, on which the files are always published on websites for sharing and downloading.

For the future work, first we will continue to study over the detailed characteristics of the files of the video type, for example to find out which kind of encoding is most used for film files and series files. Then we would like to figure out the relation between the cyberlockers and the file-sharing and file-downloading forums. We also like to insert an intelligent classification method for the cyberlocker files classification.

6. References

[1] File_Hosting_Service; http://en.wikipedia.org/wiki/File_hosting_service (16 January 2014).

[2] Hendrik Schulze and Klaus Mochalski (2007) "Internet Study 2007", Ipoque Report; http://www.ipoque.com/en resouces/internet-studies (16 January 2014).

[3] Hendrik Schulze and Klaus Mochalski (2009) "Internet Study 2008/2009", Ipoque Report; http://www.ipoque. com/en/resources/internet-studies (16 January 2014).

[4] Demetris Antoniades, Evangelos P. Markatos and Constantine Dovrolis (2009) "One-Clike Hosting Services: A File-sharing Hideout", in Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference, ACM Press: New York, NY, USA, pp. 223-224.

[5] Aniket Mahanti (2011) "Measurment annd Analysis of Cyberlocker Services", in Proceedings of the 20th International Conference Companion on World Wide Web, ACM Press: New York, NY, USA, p. 373–378.

[6] Aniket Mahanti, Carey Williamson, Niklas Carlsson, Martin Arlitt and Anirban Mahanti (2011) "Characterizing the File Hosting Ecosystem: A View from the Edge", Performance Evaluation of the ACM 68 (11), pp. 1085-1102.

[7] Piracy Intelligence (2011) "An Estimate of Infringing Use of the Internet", Envisional Report; http:// docouments.envisional.com/docs/Envisional-Internet Usage-Jan2011.pdf (16 January 2014).

[8] Micheal K. Bergman (2001) "The Deep Web: Surfacing Hideout Value", Journal of Electronic Publishing 7 (1).

[9] Confidence Interval; http://en.wikipedia.org/wiki/ confidence_intercal (16 January 2014).

[10] Glenn D.Israel (1992) "Determing Sample Size", University of Florida IFAS Ectension Publication; http://edis.ifas.ufl.edu/pd006 (16 January 2014).