ISSN:2231-0711

Jie Cheng et al | IJCSET |February 2013 | Vol 3, Issue 2, 38-42

The Pirate Bay Torrent Analysis and Visualization

Jie Cheng, Ryder Donahue

Department of Computer Science and Engineering
University of Hawaii at Hilo
200 W. Kawili Street Hilo, HI USA 96720

***Abstract*— Using C# as a parser, we process about 3.4 million pieces of data over 680 thousand torrents from thepiratebay.org, and create a graphical representation of the data by info-graphic. Info-graphic presents the information in an easily readable format, and also can be distributed across many web-mediums. Based on the representation/analysis of the data, we are able to determine some interesting characteristics and properties of the torrents hosted there, such as operating system (OS) specific share ratios, average file size, average torrent seeder ratios, and much more. The discovered characteristics can help torrent users to make informative decision about their torrent usage, and seeding/leeching habits.**

***Keywords*— Torrents, Pirate Bay, infographic, data analysis, C#, visualization, BitTorrent.**

I. INTRODUCTION

A. *Overview of BitTorrent*

With the explosive growth of world wide bandwidth usage and cost, file and software distributers are turning from traditional centralized file hosting to new models and platform for content distribution. The BitTorrent file distribution system [1] is a novel approach to minimize the file cost and bandwidth usage through peer-to-peer (P2P) configuration. In contrast to other P2P systems like Kazaa and Gnutella [2] [3] whose main goal is to response a file request query quickly with results, BitTorrent file distribution system is to increase file availability to users by quickly replicating the file [4]. Asynchronous upload/download rates encountered in many non-commercial high-speed connections also get addresses. Due to these advantages, BitTorrent accomplishes well in almost any network environment; therefore it has accounted for about 43% to 70% of all internet traffic by February 2009 [5].

BitTorrent architecture consists of a number of components. The use of web server is to store meta-data file with .torrent extension. Torrent files are used to specify the information of the file package that tends to be downloaded from a web server via a standard web browser. The torrent file includes information such as the name of the file package, the length of any files contained in the package, piece hash data, and the URL of the tracker. Peers in the system may either be uploading or uploading and downloading pieces of the package. Peers who provide a complete file are seeders, and the peer providing the initial copy is called the initial seeder. Peers that download the files by connecting to the seed are called leechers. Together, all peers, including seeders and leechers sharing a torrent are called a swarm. The main purpose of the tracker, which is often remotely hosted with the server and acts as the only centralized resource in the architecture, is to trace which peers are currently involved with a particular file package.

In order to have an access to a network of file sharers, there are certain operations in BitTorrent system. First, users who wish to share any digital content which can be accessed from their own computer, such as movies, music, and software, must create a torrent file and make these files available through tracker. These individual become the first seeder. The file being distributed is divided into segments called pieces.Then any other users (leechers) interested in these contents can connect to the initial seeder's file via tracker by searching a list of active torrent files from web. Besides seeders, leechers also interact with other leechers to exchange parts of the file that they do not yet obtain. While leechers get a new piece of file, it can become the source for other peers. Therefore, only one single copy of the file sent by the seeder finally can be distributed to an unlimited number of peers. Upon completion of downloading the file, leechers have two options: 1) walk away with the file and terminate the downloading/sharing process (this usually disliked by the community) 2) subsequently act as seeders. If the original seeder stop the seeding process the torrent will become dead unless at least one of the former leecher continue seeding the file to keep the torrent remain active.

B. *Overview of the Pirate Bay – BitTorrent Communities*

Since BitTorrent depends on central servers that run trackers, simple processes that monitor the users that are downloading or uploading content and mainly allow peers to search and retrieve content they are interested in by making available the addresses of the contents, there are thousands of BitTorrent trackers on the internet. Some of them not only offer content database, but also provide services such as content search and rating, forums, comments, account management, and web feed, which make newly published content easily and quickly discovered by the users. These website are very popular and surrounded by communities of millions of users, among which the Pirate Bay (TPB) is by far

the most well-known and largest public torrent-related community with around 21 million users are active at any given moment.

We aim to gain interesting and detailed insight, including characteristics and properties, about torrent files hosted on TPB. Based on this information, users will be able to make more informed decisions about their torrent usage, and seeding/leeching habits.

The paper is organized as follows: section II describes our approach for data analysis and visualization of TPB. Section III presents the related work. Data analysis and visualization of TPB is shown in section IV. Finally, the conclusions are drawn in section V.

II. Our Approach/Steps for Data Analysis and Visualization

By using C# as a parser to read and parse the data as a string, we sift through about 3.4 million pieces of data over 680 thousand torrents from TPB. It took only several seconds to compute the entire set. We then create a graphical representation of the data by infographic."*An infographic is a image that takes a large amount of information and presents it in asthetically pleasing tables, graphs, and other data visualization methods that allow the user to absorb a large amount of data quickly and enjoyably.* "[6][7]The software we used to create the infographic in was *Adobe Flash Professional CS5.5*, as it provides scalable vector tools to ensure the high quality of the final product. The final image was encoded in high quality JPG format and sizes in around 1000px by 7800px.

III. Related Work

Andrade et al. [8] presents measurement results of six communities including TPB, aiming to study file popularity, seeding, and leeching. They conclude that torrent popularity distributions are not following power-law distribution. Even though a small set of users supply most of the resources, users that contribute more resources are also those that demand more from it. They also discovered that the seeding ratio is higher in torrents with small file sizes and is higher in younger torrents for all except one site. Guo et al. [9] focus on torrent popularity, torrent life-span, and client performance variations and characterization of torrent system. They observe that the BitTorrent's policy in favor of high speed downloaders to replicate file fast in fact affect peers in downloading; therefore an encouragement mechanism is needed to motivate seeds to contribute. Renowned early measurement studies of the BitTorrent protocol are on the evolution of a torrent by Izal et al. [7], which analyzed up to a few thousands simultaneously active clients over a large period of time. The results obtained validate that BitTorrent is highly effective based on the throughput per client during download and the ability to sustain high flash crowd.

IV. Experimental Result and Analysis

A. Sample Output of C#

The following are the sample output in the end of the program by C#: Please Wait; Computing Data; data: The Pirate Bay Data.csv; Total Number of items: 679516; Music: 189278; Movies: 275763; Applications: 51173 Games: 47540; Other: 39005Porn: 63659; Sizes; TotalMusic:43125.1 average 227.84; TotalGames: 64948.06 average1366.177; TotalMovie: 450009.6 average 1631.871; ApplicationTotal: 17735.65 average 346.5822; PornTotal: 40811.07 average 641.0887; OtherTotal: 6865.408 average 176.0135; PERCENTAGES %%%% ;Music: 27.85483; Movies: 40.58227 ;Applications: 7.530801 ;Games: 6.996156; Other: 5.740115 ;Porn: 9.368286 ;Category Ratios ;Music: 3.010126 ;Movies: 1.438316 ;Games: 2.493047 ;Application: 4.493691 ; Other: 3.264016 ;Sizes 1_10KB: 901 %: 0.1325944 ;10_100KB: 1244 %: 0.1830715 ;100_1000KB: 15457 %: 2.274707 ;1MB: 9470 %: 1.393639 ;2_5MB: 21107 %: 3.106181 ;5_25MB: 55221 %: 8.126519 ;25_50MB: 46142 %: 6.790421 ;50_200MB: 170754 %: 25.12877 ;200_500MB: 106744 %: 15.70883 ;500_1,000MB: 124043 %: 18.25461;1_2GB: 53213 %: 7.831015 ;2_4GB: 23008 %: 3.385939 ;4_6GB: 37442 %: 5.510098; 6_10GB: 10772 %: 1.585246 ;10_20GB: 2500 %: 0.3679089 ;20_100GB: 1466 %: 0.2157418 ;100+GB: 31 %: 0.004562071 ;Ratios 1_10KB: 4.647742 ;10_100KB: 4.062142 ;100_1000KB: 3.980501 ;1MB: 4.082843 ;2_5MB: 4.151653;5_25MB: 3.757301 ;25_50MB: 2.848794 ;50_200MB: 2.813746; 200_500MB: 1.888859 ;500_1,000MB: 1.531591 ;1_2GB: 1.231281 ;2_4GB: 1.192464 ;4_6GB: 0.9310662 ;6_10GB: 0.7831408 ;10_20GB: 0.6954312 ;100+GB: 0.7111191 ;Average Number of Seeders/Leechers Ratio 1KB7.284238 …

B. Data Visualization and Analysis

In order to visualize these data, we are able to take all of the data and shove into one nice infographic. For illustration purpose, we decompose this one inforgraphic into different figures.

Based on our result of data processing, in December 2008, there were 679,515 unique torrents, hosted on TPB with 632.4 terabytes (TB) of file size and 5.7 Petabytes(PB) of global size, which refers to the amount of disk space being used by all the seeders. The average file is 931 Megabyte (MB) with a 1.2 seed/leecher ratio. There are about 65% active torrents in TPB; in contrast, there are 35% dead torrents. An interesting finding is that if a torrent has 0 seeders or leechers, it will have 0 seeders and leechers. This is unexpected and is possibly an amoral of this dataset. This overall statistics is shown in figure 1.

Figure 1: General Statistics of TPB

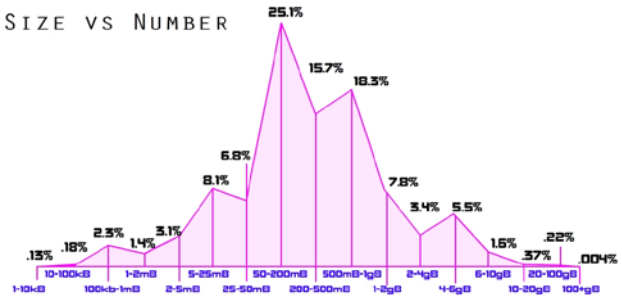

Figure 2: Distribution of the torrents in terms of file size

Figure 2 represents the distribution of torrents based on the size of the files they are related to. X-axis represents the intervals of file size and y-axis represents the ratio. It reveals that the majority of files are clustered within the 50MB-1 gigabytes (GB) range. Normally files in this range are CDs, Albums, re-encoded DVD Movies, or software.

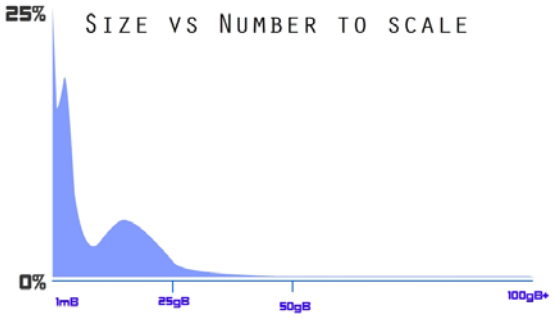

Figure 3: Distribution of the torrents in terms of file size with scaled intervals.

Figure 3 shows the same information as the figure 2, with the X-axis been scaled evenly from 0-100GB, instead of intervals used in the figure 2. Clearly, large percentage of files shared by peers concentrates on the size smaller than 25 GB. Above 99% of files has file size smaller than 10 GB.

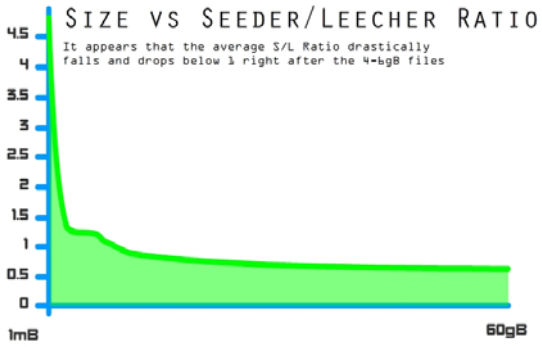

Figure 4: Seeder/Leecher ratio based on the torrents size

The seeder/leecher ratio represents the number of seeders per leecher, and therefore hints an idea of supply vs. demand in swarms. Figure 4 depicts the seeder/leecher ratio based on the torrents size. The ratio dramatically drops below 1 as the files size increase to 4 MG and larger. Even with the smaller file size, TPB had only 2-4 seeders per leecher on average. It is shown that leechers tend not to continue seeding large files after they have finished downloading them. This could be because the file is normally not kept in the same location or deleted after it is downloaded, either of which would prevent the file from being seeded.

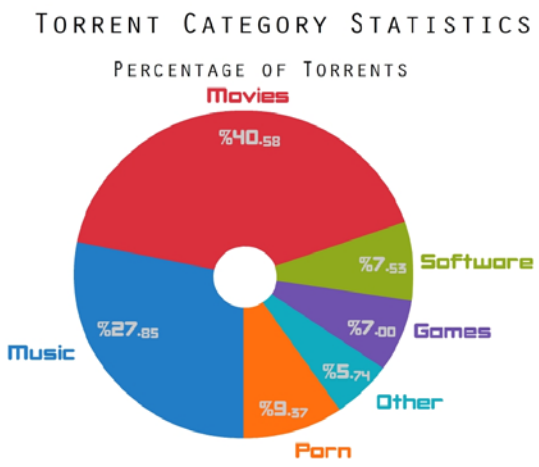

Figure 5: Torrent Category Statistics

Figure 5 shows that torrents for movie account for the largest percentage with 40.58%. Peers in TPB like to share more about movies and music than software, games, and porn.

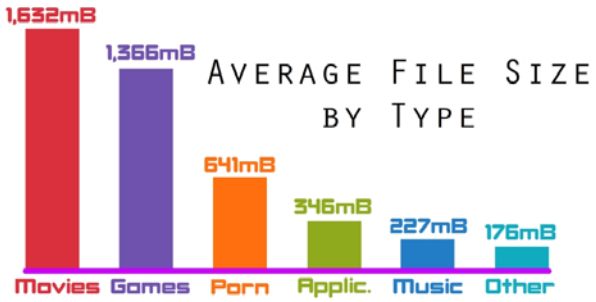

Figure 6: Average file size by type

Figure 6 shows that the average file size for different categories of files, such as movies, games and application. The average size of the movie with 1,622MB is the largest and followed by the games.

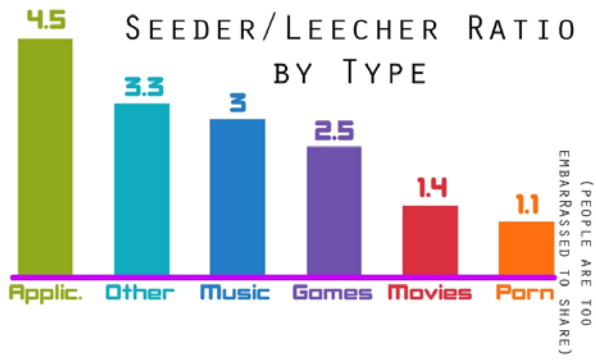

Figure 7: Seeder/leecher ration by type

Figure 7 represents the seeder/leecher ratio based on type, ranging from application to porn. In average, there are 4.5 seeders per leecher for application files and 1.4 seeders per leecher for movies. It hints that users tend to continue seeding application files after they have finished downloading them. In other words, it is likely that requests for application files from an arriving leecher tend to get granted faster by seeders than those requests for movies and porn. The low ratio of movies is possibly due to the copyright of the movie.

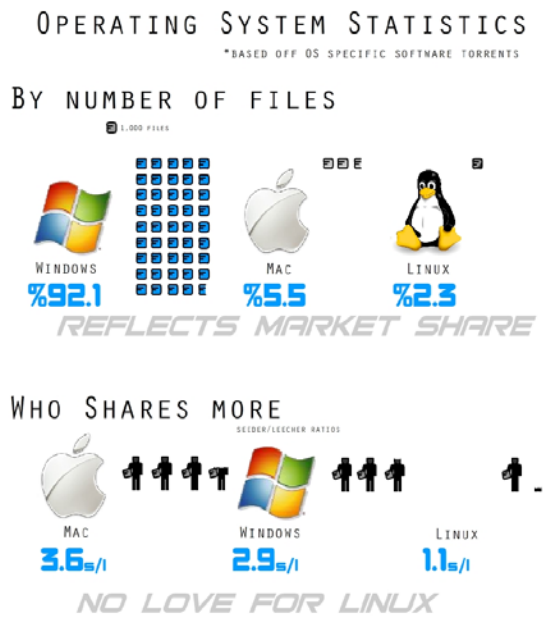

Figure 8: Operating System Statistics

Based on the operating system specific software torrents, Figure 8 shows that 92.1% torrents are windows software; in contrast only 2.3% are Linux software. This statistics conforms to the market share of these three operating systems. Based on the seeding and leeching ratio, Mac users are more willing to share mac application than windows and Linux users. Mac application has 3.6 seeders per leecher; window application has 2.9 seeders per leecher; Linux application has 1.1 seeders per leecher. It is clear that Mac software are intended to get shared the most, followed by the windows and Linux as the last.

V. CONCLUSIONS

Our most important finding by analyzing 3.4 million data on TPB over 680 thousand torrents are shown as follows: 1) In December 2008, there were 679,515 unique torrents, hosted on TPB with 632.4 TB of file size and 5.7 PB of global size; 2) Above 99% of files has file size smaller than 10 GB; 3) majority of files are clustered within the 50MB-1 GB range; 4) Even with the smaller file size, TPB had only 2-4 seeders per leecher on average; 5) There are 4.5 seeders per leecher for application files and 1.4 seeder per leecher for movies. Porn has the lowest seeder/leecher ratio of 1.1. 6) 92.1% of application torrents are windows software and 55% are Mac; in contrast only 2.3% are Linux software. This statistics conforms to the market share of these three operating systems.

REFERENCES

[1] B. Cohen. Incentives Build Robustness in BitTorrent. In Workshop on *Economics of Peer-to-Peer Systems*, Berkeley, USA, May 2003. http://bittorrent.com.

[2] D. Milojičič, et al, Peer-to-Peer Computing. *Technical Report HPL-2002-57*, HP Laboratories, Palo Alto, CA, USA, 2002.

[3] S. Androutsellis-Theotokis, and D. Spinellis, A Survey of Peer-to-Peer Content Distribution Technologies. *ACM Computing Surveys* (CSUR), Vol. 36, No. 4, pp 335-371, 2004.

[4] M. Izal, G. Urvoy-Keller, E. W. Biersack, P. Felber, A. A. Hamra, and L. Garc´es-Erice. Dissecting bittorrent: Five months in a torrent's lifetime. In PAM, pages 1–11, 2004.

[5] H. Schulze and K. Mochalski (2009). "Internet Study 2008/2009". Leipzig, Germany: ipoque. *http://www.ipoque.com/sites/default/files/mediafiles/documents/internet-study-2008-2009.pdf*.

[6] D. Newsom and J. Haynes (2004). *Public Relations Writing: Form and Style*. p.236.

[7] M. Smiciklas, The Power of Infographics: Using Pictures to Communicate and Connect with Your Audience, 2012.

[8] N. Andrade, M. Mowbray, A. Lima, G. Wagner, and M. Ripeanu. Influences on cooperation in bittorrent communities. In *P2PECON'05*, pages 111–115, New York, NY, USA, 2005. ACM.

[9] L. Guo, S. Chen, Z. Xiao, E. Tan, X. Ding, and X. Zhang. Measurements, analysis, and modeling of bittorrent-like systems. In IMC'05, pages 4–4, Berkeley, CA, USA, 2005. USENIX Association.

[10] M. Izal, G. Urvoy-Keller, E. W. Biersack, P. Felber, A. A. Hamra, and L. Garc´es-Erice. Dissecting bittorrent: Five months in a torrent's. In *PAM*, pages 1–11, 2004.