A Week in the Life of the Most Popular BitTorrent Swarms

Mark Scanlon, Alan Hannaway and Mohand-Tahar Kechadi
UCD Centre for Cybercrime Investigation, School of Computer Science & Informatics,
University College Dublin, Belfield, Dublin 4, Ireland
{mark.scanlon, alan.hannaway, tahar.kechadi}@ucd.ie

Abstract—The popularity of peer-to-peer (P2P) file distribution is **consistently increasing since the late 1990's. In 2008, P2P traffic accounted for over half of the world's Internet traffic. P2P** networks lend themselves well to the unauthorised distribution of copyrighted material due to their ease of use, the abundance of material available and the apparent anonymity awarded to the downloaders. This paper presents the results of an investigation conducted on the top 100 most popular BitTorrent swarms over the course of one week. The purpose of this investigation is to quantify the scale of unauthorised distribution of copyrighted material through the use of the BitTorrent protocol. Each IP address, which was discovered over the period of the weeklong investigation, is mapped through the use of a geolocation database, which results in the ability to determine where the participation in these swarms is prominent worldwide.

I. INTRODUCTION

IN 2008, Cisco estimated that P2P file sharing accounted for 3,345 petabytes of global Internet traffic, or 55.6% of the total usage. Cisco forecast that P2P traffic would account for 9,629 petabytes globally in 2013 (approx. 30% of the total usage) [3]. While the volume of P2P traffic is set to almost triple from 2008-2013, its proportion of total Internet traffic is set to decrease due to the rising popularity of content streaming sites and one-click file hosting sites such as Rapidshare, Megaupload, etc. BitTorrent is the most popular P2P protocol used worldwide and accounts for the biggest proportion of Internet traffic when compared to other P2P protocols. Depending on region, Ipoque GmbH has measured BitTorrent traffic to account for anything from 20%-57% of that region's total Internet usage in 2009 [11].

The BitTorrent protocol is designed to easily facilitate the distribution of files to a very large number of downloaders with minimal load on the original file source [2]. This is achieved through the downloaders uploading their completed parts of the entire file to other downloaders. A BitTorrent swarm is made up of seeders, i.e., peers with complete copies of the content shared in the swarm, and leechers, i.e., peers who are downloading the content. Due to BitTorrent's ease of use and minimal bandwidth requirements, it lends itself as an ideal platform for the unauthorized distribution of copyrighted material, which typically commences with a single source sharing large sized files to many downloaders.

To commence the download of the content in a particular BitTorrent swarm, a metadata ".torrent" file must be downloaded from an indexing website. This file is then opened using a BitTorrent client, which proceeds to connect to several members of the swarm and download the content. Each BitTorrent swarm is built around a particular piece of content which is determined through a unique identifier based on a SHA-1 hash of the file information contained in this UTF- 8 encoded metadata file, e.g., name, piece length, piece hash values, length and path.

Each BitTorrent client must be able to identify a list of active peers in the same swarm who have at least one piece of the content and is willing to share it, i.e., that has an available open connection and has the bandwidth available to upload. By the nature of the implementation of the protocol, any peer that wishes to partake in a swarm must be able to communicate and share files with other active peers. There are a number of methods that a client can attempt to discover new peers who are in the swarm:

1. Tracker Communication – BitTorrent trackers maintain a list of seeders and leechers for each BitTorrent swarm they are currently tracking. Each BitTorrent client will contact the tracker intermittently throughout the down- load of a particular piece of content to report that they are still alive on the network and to download a short list of new peers on the network.
2. Peer Exchange (PEX) – Peer Exchange is a BitTorrent Enhancement Proposal (BEP) whereby when two peers are communicating, a subset of their respective peer lists are shared during the communication.
3. Distributed Hash Tables (DHT) – Within the confounds of the standard BitTorrent specification, there is no intercommunication between peers of different BitTorrent swarms. Azureus/Vuze and µTorrent contain mutually exclusive implementations of distributed hash tables as part of the standard client features. These DHTs maintain a list of each active peer using the corresponding clients and enables cross-swarm

Co-funded by the Irish Research Council for Science, Engineering and Technology and Intel Ireland Ltd. through the Enterprise Partnership Scheme.

communication between peers. Each peer in the DHT is associated with the swarm(s) in which he is currently an active participant.

The most popular BitTorrent indexing website, according to Alexa, is The Pirate Bay [9]. In January 2010, The Pirate Bay held the Alexa global traffic rank of number 99 and is the 91st most popular website visited by Internet users in the United States [1]. For the purpose of the investigation outlined in this paper, the top 100 torrents listed on The Pirate Bay were chosen to be investigated due to the popularity of the website.

II. Specifications of the Investigation

The steps involved in the execution of this investigation are:

1. Connect to The Pirate Bay and download the ".torrent" metadata files for the top 100 torrents.
2. Connect to each swarm sequentially and identify each of the IP addresses currently active in the swarm until no new IPs are found.
3. Once the active IPs are found for the entire 100 torrent swarms, the process is repeated for the next 24 hours.
4. After 24 hours, the process was begun again at step 1.

The investigation was conducted using a single dedicated server, which sequentially monitored each torrent swarm until all the IPs in the swarm were found. Over the course of the seven day investigation, a total of 163 different torrents were investigated. None of the content appearing in these torrents was found to be distributed legally; each torrent swarm is distributing copyrighted material without any documented authorization.

A. Investigation Methodology

For a regular BitTorrent user, the starting point to acquiring a given piece of content is to visit a BitTorrent indexing website, e.g., The Pirate Bay [9]. The indexing site serves as a directory of all the content available to the user. The user finds the specific content of interest by browsing the website or through keyword search. He must then download the ".torrent" metadata file which contains information specific to the content desired such as name, size, filenames, tracker information, hash value for completed download, chunk size, hash values for each chunk, etc.

TABLE I
Breakdown of the Information Contained in the ".torrent" Metadata Files

	Maximum	Minimum	Average
Content Size	38.37GB	109.69MB	1.62GB
Chunk Size	4MB	128KB	1.3MB
Number of Chunks	19,645	346	1,251
Number of Files	322	2	24
Number of Trackers	57	4	20

This ".torrent" file is then opened with a BitTorrent client, e.g., Azureus/Vuze, µTorrent, etc., which in turn contacts the tracker or the DHT for a bootstrapping list of IP addresses to get started in the swarm.

The main goal of the methodology used for this investigation is to gather information in an intelligent and novel manner through the amplification of regular client usage. This is in order to collect the complete list of IPs involved in any given swarm as efficiently as possible. For example, in a large swarm of >90,000 IPs, the software tool developed for this experiment is capable of collecting the active peer information in as little as 8 seconds. However, the precise design and specifications of the software used during the investigation outlined is beyond the scope of this paper.

B. Specifics of the Content Investigated

From the analysis of the daily top 100 swarms, video content was found to be the most popular being distributed over BitTorrent. Movie and television content amounted for over 94.5% of the total, as can be seen in Fig. 1, while music, games and software amounted for 1.8%, 2.5% and 1.2% respectively. One highly probable explanation for the popularity of television content is due to the lag between US television shows airing in the US and the rest of the world.

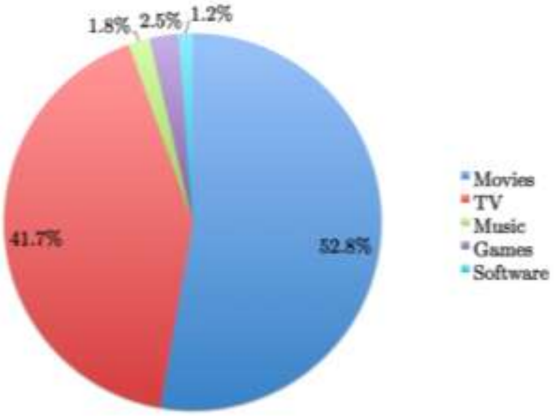

Fig. 1. BitTorrent swarms investigated by category

III. Results and Analysis

Over the week long investigation, the total number of unique IP addresses discovered was 8,489,287. On average, each IP address detected was active in 1.75 swarms, or, almost three out of every four IP addresses were active in at least two of the top 100 swarms during the week. The largest swarm detected peaked at 93,963 active peers and the smallest swarm shrunk to just 1,102 peers. The time taken to capture a snapshot of 100 swarms investigated varies due to the increase and decrease in the overall size of the swarms. The average time to collect all the peers' information for each swarm is 3.4 seconds.

A. File Information

Notably, 50.6% of the files contained in the top 100 torrents are split into smaller chunks for distribution, as can be seen in Figure 2. 38.9% of the files were also compressed into RAR files to save on the overall size of the content. The partitioning of large files into numerous smaller files is consistent with file distribution through one-click file hosting websites and newsgroups where the distribution of small files greatly improves the overall throughput of the system.

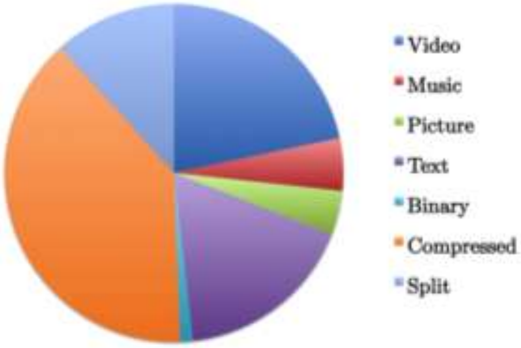

Fig. 2. Breakdown of file types discovered during the investigation

Video files are distributed as AVI, MP4 or MKV files and are typically grouped with screenshots, subtitles and sample videos. Music files are generally distributed as MP3 files and are grouped with associated album playlists and artwork.

B. Worldwide Distribution

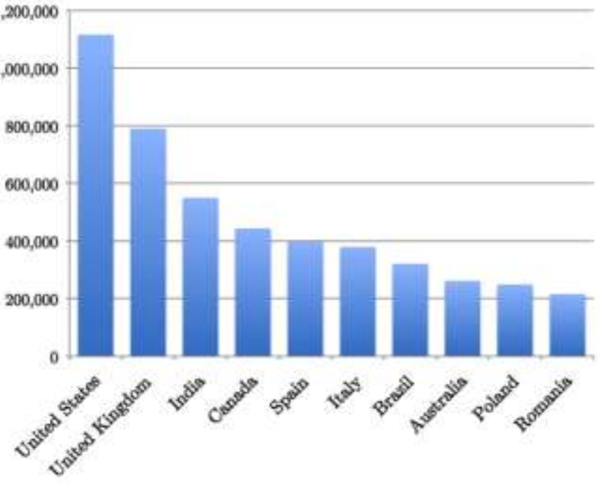

Fig. 1. Top 10 countries detected.

There are a number of assumptions that had to be made for the purposes of the geolocation of the IP addresses and for the count of the total number of BitTorrent users involved in the swarms investigated:

1. The country, city and ISP level geolocation databases used are as accurate as possible. In January 2010, MaxMind state that their geolocation databases are 99.5% accurate at a country level, 83% accurate at a US city level within a 25 mile radius and 100% accurate at the ISP level [7].
2. For a week long investigation, each IP address found to be participating in the swarms investigated is assumed to only ever be allocated to one end user. Due to a typical DHCP lease from an Internet service provider lasting somewhere in the order of 2-7 days, dynamic IP address allocation may result in the reallocation of the same IP address to two or more end users during the investigation. Should this have occurred during the investigation, it is ignored for the interpretation of the results outlined below. It is deemed technically infeasible to identify precisely when this may occur on a global level within the scope of this paper.
3. No anonymous proxy servers or Internet anonymity services, e.g., I2P [5], Tor [12], etc., are used by the IP addresses discovered.
4. It is infeasible for users on dial-up Internet connections to download the very large file that typically justifies distribution via the BitTorrent protocol. The average content size for the swarms investigated was 1.62GB, which would take a typical 56kbps dial-up user over 69.5 hours to download, assuming no other Internet traffic and achieving the maximum theoretical dial-up connection speed. For the purposes of the analysis presented in section III below, it was assumed that a negligible amount of dial-up users participate in the BitTorrent swarms.

For each IP address detected during the investigation, the geolocation is obtained using MaxMind's GeoIP databases [7], which results in information such as city, country, latitude, longitude, ISP, etc., being resolved. This information is then gathered and plotted as a heatmap to display the distribution of the peers involved in copyright infringement on a world map, seen in Figure 4. The most popular content tends to be content produced for the English speaking population, which is reflected in the heatmap, i.e., countries with a high proportion of English speaking population are highlighted in the results.

TABLE II
NUMBER OF BROADBAND SUBSCRIBERS DISCOVERED DURING THE INVESTIGATION (ASSUMING A NEGLIGIBLE AMOUNT OF DIAL-UP USERS)

Country	Number of IP Addresses Discovered	Broadband Subscription Count [6]	Percentage of Broadband Subscriptions Discovered
United States	1,116,111	73,123,400	1.53%
United Kingdom	790,162	17,276,000	4.57%
India	549,514	5,280,000	8.70%
Canada	443,577	9,842,300	4.51%
Spain	397,892	8,995,400	4.42%
Italy	378,892	8,995,400	3.36%
Brazil	320,829	10,098,000	3.18%
Australia	261,433	5,140,000	5.09%
Poland	248,731	4,792,600	5.19%
Romania	215,403	2,506,000	15.4%

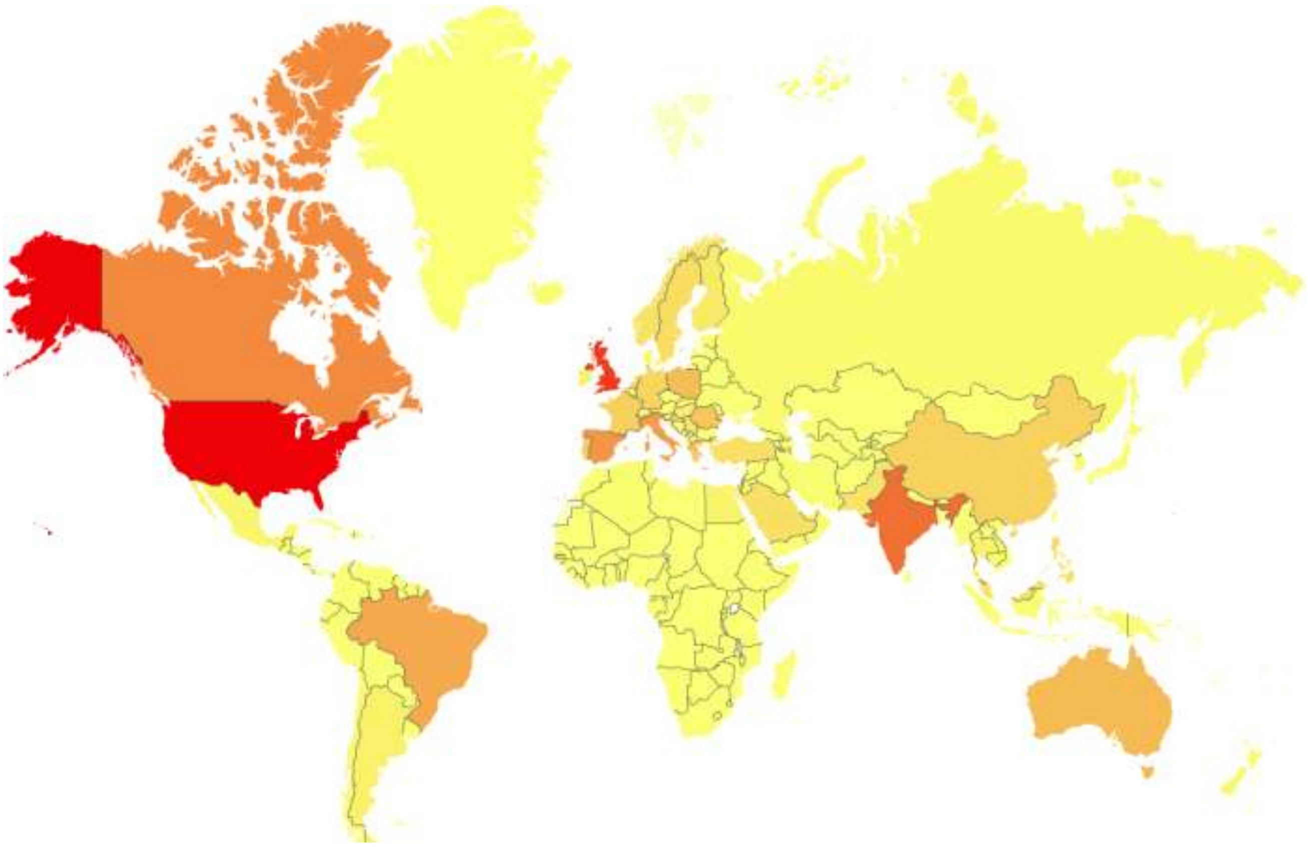

Fig. 2. Heatmap Showing the Worldwide Distribution of Peers Discovered.

A significant percentage of the worldwide broadband subscribers were detected during the investigation. 2.43% of the 349,980,000 worldwide broadband subscriptions was discovered during the investigation [6]. The percentages of broadband subscribers detected in the top 10 countries are outlined in Table II. The top ten countries detected account for over 53.6% of the total number of IPs found.

C. United States

The United States is the most popular country detected with over 1.1 million unique IP addresses, which accounted for 13.15% of all the IP addresses found. While accounting for the largest portion of the results obtained in this investigation, this relatively low percentage suggests that BitTorrent has a much more globally dispersed user base in comparison to other large P2P networks. For example, a 10 day investigation conducted on the Gnutella network in 2009, it was found that "56.19% of all [worldwide] respondents to queries for content that is copyright protected came from the United States" [4]. When the IP addresses detected during this investigation are geolocated and graphed onto a map, the population centers can be easily identified, as can be seen in Figure 5. The state of California accounted for 13.7% of the US IPs found, with the states of Florida and New York accounting for 7.2% and 6.8% respectively.

D. Extrapolated Results

If the total amount of worldwide public BitTorrent activity is considered to be the summation of the number of active IP addresses in all BitTorrent swarms at any given moment served by the world's largest trackers, the percentage of overall BitTorrent activity analyzed as part of this investigation can be estimated. From analyzing the scrape information available from two of the largest trackers, OpenBitTorrent [8] and PublicBitTorrent [10], it is estimated that the most popular 100 torrents at any given moment accounts for approximately 3.62% of the total activity.

IV. Conclusion and Future Work

The objective of this investigation was to attempt to identify the scale of the unauthorized distribution of copyrighted material worldwide using the BitTorrent protocol. 2.43% of the broadband subscriber base was detected over the course of one week in the 163 torrents monitored. This number is far greater than the number of subscribers that could possibly be prosecuted for their actions. The number of end users involved in illegal downloading is undoubtedly much higher than this due to the relatively small scale of this investigation. Some network factors will also have a negative effect over the results achieved, such as two or more end-users appearing as a

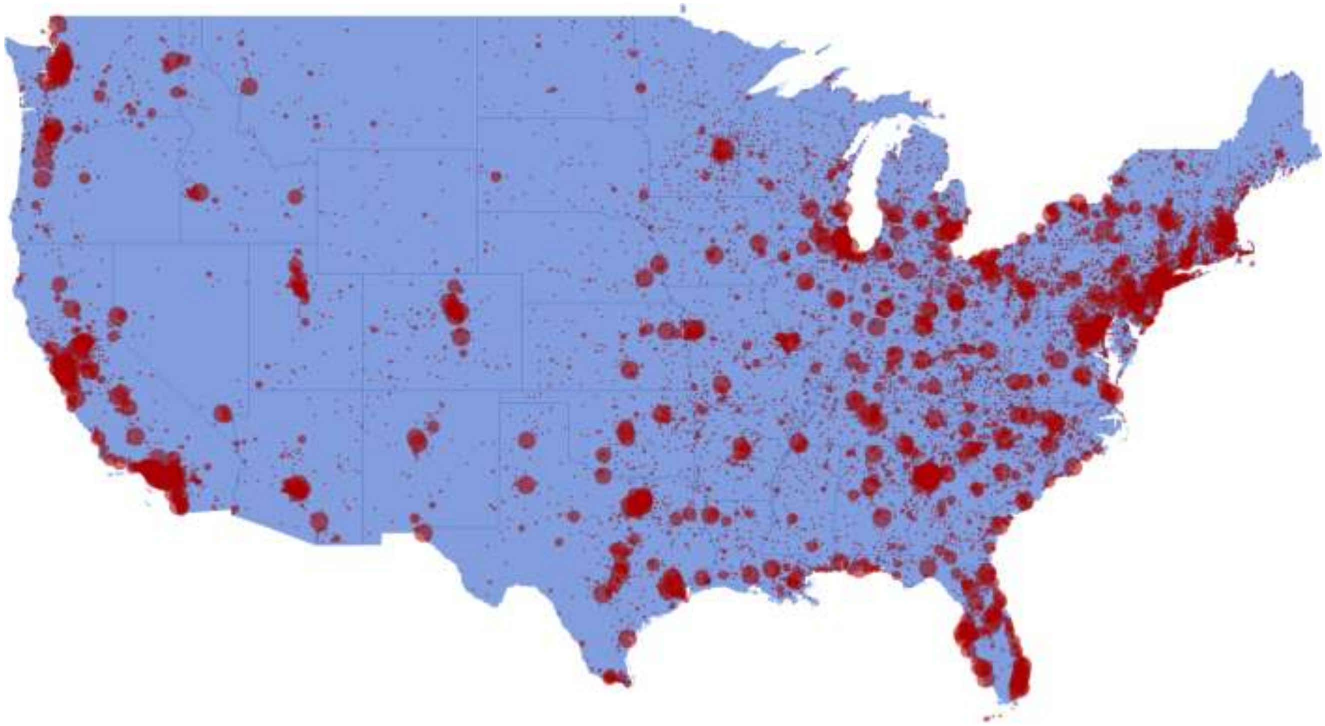

Fig. 5. Distribution of IP addresses in mainland United States. The largest circles represent locations with greater than 1500 IPs, while the smallest circles may only represent a single IP at a given location

single Internet IP address though Internet connection sharing, proxy services etc.

To further improve the results outlined above, a number of additional steps would be considered:

1. Overcome dynamic IP address reallocation – While a number of the IPs discovered during the investigation may have been allocated to more than one end user, this multiple allocation is not possible to identify using the standard BitTorrent protocol. However, the identification of individual users through metadata examination and heuristic approaches is possible. This would increase the overall accuracy of the results obtained and give an accurate measure of the total BitTorrent population.
2. Conduct a larger scale analysis – In each 24 hour period of this investigation, only 100 swarms were monitored. Increasing the scale of the investigation will yield to better results. The scaled-up analysis of all BitTorrent swarms will identify some interesting statistics, e.g., legal vs. illegal distribution, total worldwide BitTorrent population, etc.
3. Identification of peers using anonymous Internet services – By comparing the IP addresses discovered during the investigation with a list of known Internet traffic proxy or pass-through services, such as that maintained by MaxMind [7], the quality of the results collected can be greatly improved.
4. Identification of Internet connection sharing – Through the analysis of the peer metadata, such as client information, downloaded content types and categories, etc., multiple users sharing a single Internet connection can be identified.

REFERENCES

[1] Alexa Information on The Pirate Bay, Downloaded January 2010, http://www.alexa.com/siteinfo/thepiratebay.org

[2] The BitTorrent Protocol Specification, Downloaded January 2010, http://www.bittorrent.org/beps/bep 0003.html

[3] Cisco Systems, Inc., Cisco Visual Networking Index: Forecast and Methodology, 2008-2013. Retrieved from: http://www.cisco.com/en/ US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white paper c11- 481360.pdf. 2009.

[4] A. Hannaway and M-T. Kechadi, An Analysis of the Scale and Distribution of Copyrighted Material On The Gnutella Network International Conference on Information Security and Privacy, Orlando, USA, 2009.

[5] I2P - The Anonymous Network, http://www.i2p2.de

[6] International Telecommunication Union, United Nations, Report on Internet, Report Generated January 2010, http://www.itu.int

[7] MaxMind Inc., GeoLite Country Database, Downloaded January 2010, http://www.maxmind.com

[8] OpenBitTorrent, Aggregated Scrape File. Downloaded January 2010, http://openbittorrent.com/all.txt.bz2

[9] The Pirate Bay, World's Largest BitTorrent Tracker, Total Top 100, Downloaded January 2010, http://thepiratebay.org/top/all

[10] PublicBitTorrent, Aggregated Scrape File. Downloaded January 2010, http://www.publicbt.com/all.txt.bz2

[11] H. Schulze, K. Mochalski, Internet Study 2008/2009. Ipoque GmbH. http://www.ipoque.com/resources/internet-studies/ 2009.

[12] The Tor Project, http://www.torproject.org/