Investigating Cybercrimes that Occur on Documented P2P Networks

Mark Scanlon*, Alan Hannaway, Mohand-Tahar Kechadi

UCD Centre for Cybercrime Investigation,
School of Computer Science and Informatics,
University College Dublin, Ireland.

Email: {mark.scanlon, alan.hannaway, tahar.kechadi}@ucd.ie

ABSTRACT

The popularity of Peer-to-Peer (P2P) Internet communication technologies being exploited to aid cybercrime is ever increasing. P2P systems can be used or exploited to aid in the execution of a large number of online criminal activity, e.g., copyright infringement, fraud, malware and virus distribution, botnet creation and control, etc. P2P technology is perhaps most famous for the unauthorised distribution of copyrighted materials since the late 1990's, with the popularity of file-sharing programs, such as Napster, etc. In 2004, P2P traffic was accounted for 80% of all Internet traffic and in 2005, specifically BitTorrent traffic accounted for over 60% of the world's P2P bandwidth usage. This paper outlines a methodology for investigating a documented P2P network, BitTorrent, using a sample investigation for reference throughout. The sample investigation outlined was conducted on the top 100 most popular BitTorrent swarms over the course of a one week period.

Keywords: *Computer Forensics, Cybercrime, Peer-to-Peer, Investigation, Methodology, Internet, Communication, Protocol, BitTorrent.*

INTRODUCTION

The efficiency, ease of use, negligible cost (zero cost once one has a computer and an Internet connection) and the perceived anonymity of individual peers lends P2P networks well to being used for a growing number of cybercrimes, ranging from copyright infringement and virus distribution to botnet creation and control.

The content producing industry report that revenue figures are steadily declining as a result of online piracy. The IFPI's Digital Music Report 2011 states that legitimate digital music distribution is up 1000% from 2004 to 2010, although total global recorded music revenues are down 31% over the same period. The report cites Internet piracy as having a significant impact on their sales. The report cites a study from 2010 entitled "Piracy, Music and Movies: A Natural Experiment" which found that physical sales would be up 72% with the abolishment of piracy in Sweden.

In 2008, P2P traffic accounted for over half of the world's Internet traffic. P2P networks especially lend themselves well to the unauthorised distribution of copyrighted material due to the abundance of material available to the downloaders. This paper

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

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 the unauthorised distribution of copyrighted material through the use of the BitTorrent protocol.

BITTORRENT

Based on global bandwidth usage, BitTorrent is the most popular P2P network in use today. In 2005, D. Erman measured BitTorrent traffic was to account for over 60% of the world's bandwidth usage. The BitTorrent protocol is designed to easily facilitate the distribution of files to a potentially large number of interested parties, i.e., other peers, with minimal load on the original file source, as outlined in the BitTorrent protocol specification. This is achieved through the following steps:

1. The file is split up into a number of uniformly sized pieces or chunks – with typical chunk sizes generally ranging from 128kB to 4MB.
2. The initial source of the file creates a UTF-8 encoded ".torrent" metadata file, which includes unique SHA-1 hash values for the entire file and each of the file chunks, along with other required file information, e.g., filenames, chunk size, total file size, path information, client information, comments etc.
3. This metadata file is then shared by the creator with other users interested in acquiring the original content – either through direct distribution, e.g., email, instant messaging etc., or through the much more common method of uploading onto a torrent indexing website, such as ThePirateBay.org.
4. Users interested in downloading the available content must then download this metadata file and open it using a BitTorrent client, such as Azureus/Vuze or µTorrent.
5. The BitTorrent client is then tasked with identifying other peers who are sharing the file uniquely identified in the metadata file, i.e., other peers in the swarm. This includes identifying seeders, i.e., peers with complete copies of the content shared in the swarm, and other leechers, i.e., peers who are currently downloading the content, but are sharing the completed chunks with others. This peer discovery is achieved through a variety of methods including tracker communication, distributed hash tables and peer exchange.

The success of the BitTorrent protocol can be attributed to uploaders incurring no additional cost besides their Internet connectivity costs to share a file with many users. In practice, the original uploader need only stay connected to the swarm until a sufficient number of leechers have one full copy of the file between them. This is made possible through the leechers uploading their completed chunks of the entire file to other downloaders. Due to BitTorrent's ease of use, minimal bandwidth requirements and perceived Internet anonymity, it lends itself well as an ideal platform for the unauthorised distribution of copyrighted material, which typically has a single original source for sharing large sized files between many downloaders.

BitTorrent Peer Discovery Methods

Each leecher must be able to identify a list of active peers in the same BitTorrent swarm which has at least one chunk of the content and is willing to share it, i.e., the peer has an available open connection and has enough bandwidth to upload. The protocol is implemented in such a manner that any peer who wishes to download content from a particular swarm, must be able to communicate and share file chunks with other active peers. There are a number of methods that a peer 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 download of a particular piece of content – both 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, is mutually shared during the communication.
3. Distributed Hash Tables (DHT)– Within the specification of the standard BitTorrent protocol, there is no intercommunication between peers of different BitTorrent swarms. Azureus and µTorrent contain mutually exclusive implementations of distributed hash tables as part of the standard client features. These DHTs maintain a list of all recently active peers using each client and enable cross-swarm communication. Each peer in the DHT is associated with the swarm(s) in which he is currently an active participant.

It is common for BitTorrent clients to use more than one method of peer discovery to ensure the highest possible download speed, i.e., continuous access to a fresh list of peers with the desired parts of the original content.

INVESTIGATION METHODOLOGY

The initial step in investigating a cybercrime occurring on a documented P2P network involves determining the identifying factor that uniquely groups the required peers together. In a copyright infringement investigation scenario, the unique identifying factor is the content being investigated. However, in deciding to investigate a particular album or movie, etc., it must be decided how many different variations of the same content should be investigated. A predetermination should also be made on how peers are classified, who may have part of the content, e.g., one track from an album, a number of chunks of a given torrent, etc.

For the purpose of the investigation outlined as part of this paper, it was decided to investigate the most popular BitTorrent swarms, irrespective of the content being shared. The most popular BitTorrent indexing website, according to Alexa, is ThePirateBay.org. In November 2010, The Pirate Bay held the Alexa global traffic rank of number 95 and is the 79th most popular website visited by Internet users in the United States. As a result, the top 100 torrents investigated were taken from those listed on The Pirate Bay overall top 100 list.

Assumptions

As with any computer forensic investigation, it is of upmost importance to identify the assumptions required to ensure accurate comprehension of any results achieved. The following list of assumptions hold true for the majority of Internet focused investigations:

1. An IP address may not uniquely identify a specific peer's computer connected to the Internet. The employment of Internet connection sharing, web proxies, virtual private networks or web anonymity services such as Tor and I2P can all result in inaccuracies in the ultimate IP address identification.
2. Geo-location and ISP identification based on a given IP address is only as accurate as the databases used for lookup. For the investigation outlined in this paper, MaxMind Inc. stated that the geolocation databases used were 99.5% accurate at a country level, 83% accurate at a United States city level within a 25 mile radius and 100% accurate at the ISP level.
3. It is infeasible to detect the end-user churn rate on a given IP address due to the inconsistent dynamic IP address allocation employed by worldwide ISPs without factoring in additional available information,

e.g., the end user's connection speeds, client information and more heuristic information such as an IP address downloading content from an inconsistent category.

4. Based on the size of the content being shared, it may be safe to assume that it is infeasible for users of slow Internet connections, e.g., 56kbps dial-up Internet subscribers, to partake in a swarm. This assumption can be useful for the analysis of the proportion of Internet subscribers detected throughout the investigation.

RESULTS AND ANALYSIS

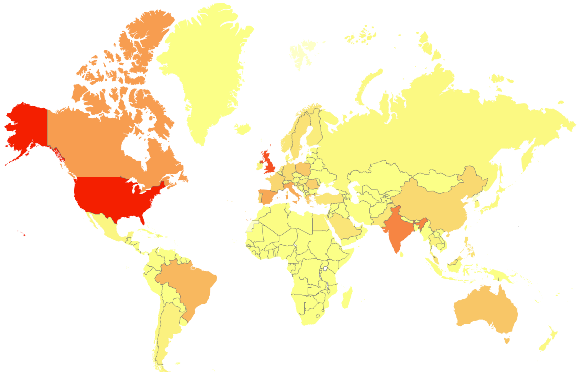

Figure 1. Heatmap displaying the worldwide distribution of IP addresses discovered.

For each IP address detected during the investigation, the IP geolocation databases developed and maintained by MaxMind Inc. are used to get the IP specific information such as city, country, latitude and 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 1.

Worldwide Results

The most popular content indexed by The Pirate Bay tends to be produced for the English speaking worldwide population, which is reflected in the heatmap in Figure 1, i.e., countries with a high proportion of English speaking population are highlighted in the results. As can be seen in Figure 2, the top ten countries detected account for over 53.6% of the total number of IPs found.

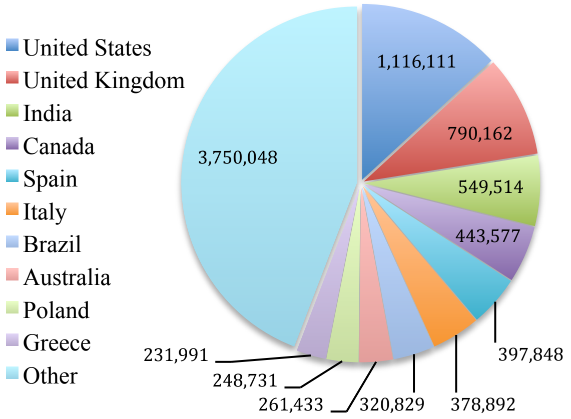

Figure 2. Top ten countries account for over 53.6% of the total worldwide activity.

Country	*Broadband Subscribers*	*Percentage Discovered*
United States	73,123,000	1.53%
United Kingdom	17,276,000	4.57%
India	5,280,000	8.70%
Canada	9,842,000	4.51%
Spain	8,995,000	4.42%
Italy	11,283,000	3.36%
Brazil	10,098,000	3.18%
Australia	5,140,000	5.09%
Poland	4,792,000	5.19%
Greece	1,506,000	15.40%

Table 1. Percentage of broadband subscribers found in each of the top ten countries.

The assumption was made that a negligible number of dial-up users were involved in the swarms investigated due to the average required download time for the file sizes involved would have been over 69.5 hours, assuming an optimal performance dial-up connection. As a result, the percentage of worldwide broadband subscribers detected during the investigation can be easily calculated based on the latest broadband subscription count from the International Telecommunication Union. 2.43% of the 349,980,000 worldwide broadband subscriptions were discovered during the investigation. The percentages of broadband subscribers detected in the top 10 countries are outlined in Table 1 above.

United States Results

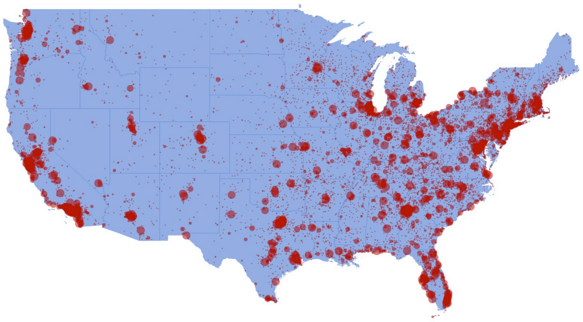

Figure 3. IP addresses discovered during the weeklong investigation plotted onto a map of the mainland United States.

The United States was the most popular country detected with over 1.1 million unique IP addresses, which accounted for 13.15% of worldwide activity. 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 users' base in comparison to other large P2P networks. For example, a 10 day investigation conducted on the Gnutella network in 2009 by Hannaway et. al, it was found that "56.19% of all [worldwide] respondents to queries for content that is copyright protected came from the United States".

When the IP addresses detected during this investigation are geolocated and graphed onto a map, the large population centres can be easily identified, as can be seen in Figure 3. 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. 14,202 IP addresses were identified in the most active city in the USA; Los Angeles. Chicago, New York, and Brooklyn were also identified as cities having more than then thousand unique peers identified each.

United Kingdom Results

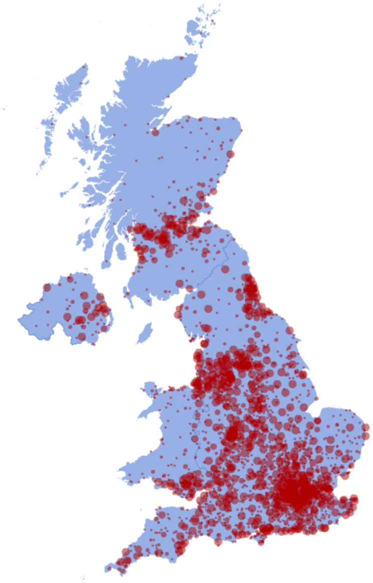

Figure 4. Geographic distribution of IP addresses found in the United Kingdom.

790,162 unique IP addresses were identified in the United Kingdom. London was found to be the city with the largest number of identified IP addresses in the world and accounted for 17.2% of the total activity in the United Kingdom. Manchester, Birmingham, Brighton, Halifax, Bristol, Glasgow and Leeds were all cities with more than ten thousand IP addresses discovered. The distribution of discovered IP addresses were mainly identified in England, with significantly fewer IP addresses discovered in Wales, Scotland and Northern Ireland, as can be seen in Figure 4.

Irish Results

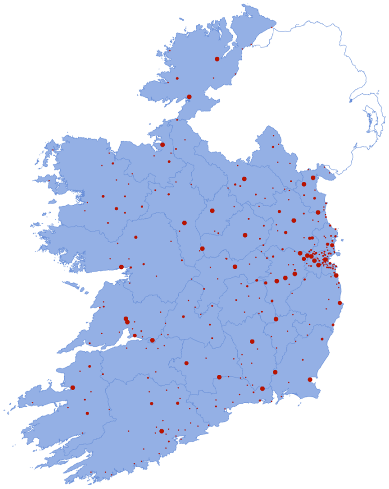

Figure 5. Geolocation of Irish IP addresses.

Of the 62,153 IP addresses identified in Ireland, 35,532 of those were located in Dublin. The cities of Limerick, Galway and Cork each had greater than 2000 unique IP addresses.

CONCLUSION

The objective of this investigation is to attempt to identify the scale of the unauthorised distribution of copyrighted material worldwide. 2.43% of the world's broadband subscriber base was detected over the course of the weeklong investigation. The actual number of P2P users involved in the unauthorised distribution of copyrighted material 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 single internet IP address though internet connection sharing, proxy services, anonymity services, etc.

ACKNOWLEDGEMENT

The authors wish to acknowledge the support of Amazon.com for supporting part of this research with cloud computing time and storage through an Amazon Web Services research grant.

REFERENCES

Alexa Internet, Inc., ThePirateBay.org Site Info. Retrieved November 17, 2010 from http://alexa.com/siteinfo/thepiratebay.org

BitTorrent Protocol Specification. Retrieved November 17, 2010 from http://bittorrent.org/beps/bep_0003.html

Erman, D. (2005). BitTorrent Traffic Measurements and models, Licentiate thesis, Blekinge Institute of Technology, Sweden.

Hannaway, A., & Kechadi, M-T. (2009). An Analysis of the Scale and Distribution of Copyrighted Material on the Gnutella Network. *In Proceedings of the International Conference on Information Security and Privacy*, Orlando, FL, USA.

IFPI (International Federation of the Phonographic Industry) Retrieved February 1, 2010 from

International Telecommunication Union, United Nations (January, 2010). *Report on Internet*, Retrieved November 17, 2010 from http://www.itu.int

MaxMind Inc., GeoLite Country Database, Retrieved January 2010 from http://www.maxmind.com

Scanlon, M., Hannaway, A., & Kechadi, M-T. (2010). A Week in the Life of the Most Popular BitTorrent Swarms. *In Proceedings of the 5th Annual Symposium on Information Assurance (ASIA '10), academic track of the 13th Annual New York State Cyber Security Conference* (pp. 32-36), Albany, NY, USA.

The Pirate Bay, World's Largest BitTorrent Tracker, Total Top 100, Retrieved January 2010 from http://www.thepiratebay.org/top/all