Detecting Pedophile Activity in BitTorrent Networks

Moshe Rutgaizer, Yuval Shavitt, Omer Vertman, and Noa Zilberman

School of Electrical Engineering, Tel-Aviv University, Israel

Abstract. The wide spread of Peer-to-Peer networks makes multimedia files available to users all around the world. However, Peer-to-Peer networks are often used to spread illegal material, while keeping the source of the data and the acquiring users anonymous. In this paper we analyze activity measurements in the BitTorrent network and examine child sex abuse activity through the Mininova web portal. We detect and characterize pedophilic material in the network, and also analyze different aspects of the abusers activity. We hope our results will help law enforcement teams detecting child molesters and tracking them down earlier.

1 Introduction

Peer-to-peer networks are being widely used around the world by millions of users for sharing content. The anonymity provided by these networks makes them prone to sharing illegal contents, from simple copyright protected material to highly dangerous material, as will be discussed next.

The BitTorrent file sharing network was responsible for 27% to 55% of internet traffic (depending on geographic location) in 2009 [13]. The BitTorrent protocol allows to download large files without loading a single source computer, rather the downloading users join a group of hosts that download and upload from each other, simultaneously. Every BitTorrent file is uniquely defined by a descriptor file called a torrent, which is distributed via email or http websites. This torrent file allows the downloading and uploading users, called leechers and seeders, to share the content file.

Pornography is one of the major content consumption area in the Internet. In 2006, over $2.84 billion were spent in the United States alone on Internet pornography, with 4.2 million websites dedicated for this purpose alone, accounting for 12% of all websites at that time [14]. Child pornography is a subset of this activity, earning over three billion dollars a year (including non-online activity as well), with over 100K websites offering child pornography and with over 116K daily queries in Gnutella network for "child pornography" [14, 15].

Many works try to fight Internet child sex abuse. The most common approach is CBIR, Content Based Image and video Retrieval, which tries to detect and retrieve visual files based on previously studied characteristics from similar files. The retrieval techniques have been thoroughly studied in many works, such as [10]. Chopra *et al.* [2] have tried to address the problem at the network level. They suggested adapting classification techniques to allow network infrastructure, such as routers, to detect illegal file transfer. Projects, such as FIVES

[3], combine efforts on multiple domains, from efficient file fragment matching, through means to evaluate large amounts of data, to improved capabilities of linking new illegal multimedia material to old one. For video analysis they use not only image features but also motion information [7]. Research related to child sex abuse in P2P networks is scarce,with few exceptions such as MAPAP [11], which focuses on the eDonkey network, or Huges *et al.* [5], studying the Gnutella network. Liberatore *et al.* [9] discussed legal issues involved in investigating child pornography in the Gnutella and BitTorrent networks. They also developed Roundup, used by Internet Crimes Against Children (ICAC) Task Forces to detect child sex abusers in the Gnutella network, given a list of known child pornography files.

In this paper we present a study of child sex abuse activity in the BitTorrent network, examining behavioral patterns in both queries and downloads. The results presented in this paper may be employed by law enforcement forces to detect and track pedophiles in the BitTorrent network, e.g. using the given analysis new illicit files can be detected.

2 MiniNova Data Set

The Mininova website [12] was for a long time a very popular BitTorrent portal, until a court order forced it to remove all copyrighted torrents at the end of 2009. According to Alexa [1] at the end of 2009 the site was ranked 90 of all worldwide websites, with 1.07% of all internet users visiting it, and first of all torrent websites, ahead of portals such as The Pirate Bay (ranked 105),Torrentz.com (ranked 190) and isoHunt (ranked 196). The average visitor to the website is a young male without children, age 18 to 34, browsing from home and staying in the website for 4.3 minutes (based on [1]). Torrents on the website can be found by browsing, or more frequently by searching the website. The torrents are located under categories and subcategories on the website, with over 950 subcategories. A new torrent uploaded to the website is placed under a subcategory selected by the originator; the website moderators rarely change a torrent's location.

The dataset used in this work was obtained from the Mininova website, and covers two time periods in 2009: the first from September 2nd to September 25th, and the second from October 15th to December 7th; a total of 67 days. The dataset was anonymized before it was provided to us, with the users IP addresses removed from the dataset. The dataset is comprised of queries and downloads.

- Queries: The query dataset holds 453 million queries. A query registry contains the query text, a timestamp, and its city of origin.
- Downloads: The downloads dataset holds 515 million torrent downloads, with over 1.3 million distinct torrents. A download entry contains the torrent's name, torrent's subcategory, file size, a timestamp, and the city of the user.

The most popular subcategories (by downloads) on the Mininova website are action and comedy movies, games for Windows, TV shows (miscellaneous) and ebooks. Pornography is not very common in the Mininova website and there is no subcategory dedicated for such torrents. Adult material is often placed under different subcategories, such as "Asian" or "Movies - Other".

2.1 Data Set Limitations

The Mininova data set analysis has several limitations. First of all, this data set covers only one Torrents website. While it can be argued that this site was clearly the leading Torrents site[1] at the sample time, it might not be representative of child sex activity in the entire BitTorrent network. Another difficulty is users anonymity, with only user's city available. This means that the activity of a specific user can not be pinpointed, e.g. there is no clear distinction between users and activity sessions. The downloads database also lacks metadata information, making it difficult to classify the file and correlate between queries and downloads. Last, there is no ground truth database for child sex abuse that can be referenced. We believe that basing our dictionary assumptions on previous work (See Section 1) that corroborate researchers from multiple fields, including social sciences, provide an adequate baseline for our analisys.

3 Results

3.1 General Statistics

We divide the queries and downloads in the database to six groups: movies, music, pictures, applications, documents, and unknown. The distinction is based on keywords for queries, and keywords and torrent category for downloads. Discarding queries and downloads of an unknown type, most of the downloads from Mininova, over 80%, are of movies. Next are music files (about 14%) and pictures (3%). Considering general queries, 53% are for movies, 21% for programs, 14.4% music and 8.5% pictures. Looking at pedophile material, the queries for them divide approximately two thirds movies and one third pictures. The amount of detected distinct pedophile files in the database is too small to set a baseline for pedophile downloads statistics.

3.2 Collected Queries Statistics

Keywords Ranking To identify pedophile related material, a dictionary of related words is created. The dictionary relies on previous works in this area [17, 8] as well as popular online sources [16]*. The dictionary of pedophile-related words that was used for this study includes 47 words. Each of these words on its own has a pedophilic meaning, but in context may become innocent. For

* We attempted to collect additional information from sources such as InHope (www.inhope.org), but failed to collaborate or retrieve information

example, "Lolita" on its own versus the combination of "Lolita" and "Nabokov", which refers to the known novel. In all our results, we apply a filter to all such known combinations, which add up to over 40 combinations. We note that the created dictionary may not be full, but we show that these words alone are enough to portray a worrying picture.

Query	**Occurrences**	**% of Pedo. Queries**	**% of Queries**
Lolita	26668	25.20%	0.0059%
Incest	26290	24.84%	0.0058%
Preteen	17910	16.93%	0.0039%
PTHC	10617	10.03%	0.0023%
Pedo	8406	7.94%	0.0018%
Underage	4756	4.49%	0.0010%
R@ygold	1594	1.50%	0.0003%
Hussyfan	1388	1.31%	0.0003%
Yamad	1325	1.25%	0.0003%
12yo	685	0.64%	0.0002%

Table 1. Statistics of Pedophilic Queries

Table 1 presents the top-10 most used terms in pedophilic queries. The table contains for each word, the number of occurrences, percentage out of pedophilic queries, and percentage out of total number of queries. The words "lolita" and "incest" are the most popular terms used. We recognize that these two words may also relate to non-pedophilic contexts, but claim that at least some of these queries are still related, as we show in the next section. We see that about 50% of all queries are the top 2 words, which we suspect not to be completely filtered, still only four or five more words are dominant in the queries. It is also observed that these words appear in one of every 25K to 100K queries out of all queries, which is considered high.

We compare these results with Gish *et al.*[4] which looked at popular queries in the Gnutella network. The term "PTHC" ranked fifth in their most popular constant phrases, appearing in approximately 0.1% of all queries, while the term incest ranked tenth, appearing in 0.05% of all queries. In the Mininova database these two terms are not as frequent as in Gnutella, however the use of Gnutella and Mininova by users is not identical; While in the Gnutella network the most popular queries are pornographic or music related [4] in Mininova the popular categories are movies and applications. The same observation also applies when comparing the results to MAPAP's eDonkey based research [11]: the term "PTHC" is ranked first, with the second term being "Pedo", both searched considerably more than any other term. Other popular terms in BitTorrent, such as "Lolita", are less popular in eDonkey, while terms such as "Preteen" and "Underage" are not ranked at all.

The most frequent queries are also compared with isoHunt's top searches list [6]**. isoHunt's list does not provide information about the number of queries

** IsoHunt was ranked second amongst torrent websites [1] at the sampling time.

per term, rather it ranks them by their popularity. In addition, isoHunt provides different ranking for filtered and unfiltered terms, with all pedophilic terms, except "lolita", being filtered out. Compared to the Mininova top-10 list, the term "PTHC" is ranked highest (83), followed by "Preteen"(119), "Lolita"(135), "Incest"(143), "Pedo"(257), "12yo"(275), "Underage"(284), and "Hussyfan"(955). The terms "R@ygold" and "Yamad" are not amongst the top 1000 searches. We see an additional difference from our top-ranked list, as terms such as "7yo", which was not amongst our top-30 queries, being placed high in the global search list (290), and with the term "9yo" being ranked higher(274) than "12yo".

Another aspect that should be considered here is the time that passed between datasets collections: two years between the Gnutella and eDonkey collection to the Mininova dataset, and two additional years from Mininova to isoHunt.Over this time, the awareness to P2P networks usage for ill purposes has grown, thus users "vocabulary" has widened and altered in order to avoid tracking.

Correlation Between Keywords Queries identified as pedophile-related often include more than a single term that is pedophilic in nature. Figure 1 presents a heatmap of keywords appearing together in the same queries. Only the highest-ranked keywords are shown. It is evident from the figure that the six highest ranking keywords are well connected: each one of them appears tens to hundreds of times in queries with the other five keywords. We strongly believe that such queries are being issued with the intent to find torrents of child pornography. The percentage of queries where two terms are used in conjunction is only 3.8%, with some of the keywords, such as PTSC and Hussyfan, co-occurring with other term in over 10% of their appearances. The keyword Yamad, on the other hand, appears in only 3 queries together with other terms.

	Lolita	Incest	Preteen	PTHC	Pedo	Under age	R@y gold	Hussyfan	Yamad	12yo
Lolita		61	304	68	91	109	4	31	0	9
Incest	61		131	73	81	26	2	1	0	8
Preteen	304	131		107	93	113	3	11	0	12
PTHC	68	73	107		75	37	106	64	2	17
Pedo	91	81	93	75		23	8	9	1	11
Under age	109	26	113	37	23		2	5	0	0
R@y gold	4	2	3	106	8	2		18	0	0
Hussyfan	31	1	11	64	9	5	18		0	0
Yamad	0	0	0	2	1	0	0	0		0
12yo	9	8	12	17	11	0	0	0	0	

Fig. 1. Heatmap of keywords appearance in the same queries

File Indicator	Total Downloads	Looked Up	Lifetime [Hours]
P1	948	397	194.5
P2	136	25	1727.9
P3	44	2	1.4
V4	2446	29	1740.1

Fig. 2. Pedophilic Torrents Downloads

On some occasions, connection can be made between pedophile terms and ordinary words. By ranking words that co-appear in the same queries as pedophile terms, some interesting insights surface. For all keywords, except for one case, there is no dominant single word that appears with them: a word never appears in more than 10% of the queries were a keyword appears. We distinguish between 3 main types of words that appear together with keywords: media type, pornography related words, and names. Words that fall under the category "media type" include, for example, "video", "pics" and "stickam", representing three type of media files: movies, still pictures and streaming media. We note that the term "video" is most common amongst such queries, as can be expected from torrents. Words that fall under the category "pornography" include terms such as "sex","xxx", and "porn". When a keyword occurs together with one of these words in a query, the ill intent of the issuing user is clear, for example, co-occurrences of "lolita" and "porn" or of "12yo" with "sex". The last group of words includes personal names, is of highest concern. This category includes names such as "Vicky", "Jenny" and "Daphne", issued together with keywords "PTSC" and "PTHC". The most troubling aspect is when these words appear together in queries with age indication, like "9yo jenny". While this may sound as a naive query, a quick search of this term on the web leads to tens of pedophile forums discussions with a clear description of the movie contents as well as other sources that include the illicit content. We thus deduce that this method can be used also outside the BitTorrent network to track and discover pedophile contents.

Extending The Dictionary An important contribution is detecting new terms that relate to child sex abuse, which is a hard task in an anonymous database. For this end, we analyze separately queries from each city, and define **a busy period** as a sequence of queries with no gaps longer than a given threshold. In large cities with many users the busy period is an aggregation of many users and may be quite long. We are looking for cities with sparse accesses to Mininova, where the probability that two user sessions will fall into the same busy period is negligible.

We analyze the busy periods length in cities with an average of 500 queries a day or less and found that in 98.5% of the cases the length is no longer than five minutes and the number of queries is no more than ten. We thus assume that these busy periods are due to a single user activity and define **a single user busy period (SUBP)** as a busy period up to five minutes long and with up to ten queries . This is in line with Alexa [1] finding that the average site visit time was 4.3 minutes. For further analysis we used only cities that contain only distinct SUBPs, at least 10 SUBPs and that registered at least one pedophile query. This resulted in 692 cities.

We find the SUBPs where pedophile terms were used in queries and create a list of potential new keywords. This list has initially about two thousand words. We screen out of these words numbers, conjunctions and terms that are highly ranked in the global queries list (such as "Harry Potter"). This process was also accompanied by a manual inspection, in order to avoid filtering required terms.

This leaves us with 140 words. We classify those to four groups: 51 General sex related words, 29 names of potential victims, 54 pedophile keywords, and 7 words that may refer to either general pornography or child sex abuse. The 54 new words include 19 words that have either a spelling error or a different spelling than an existing keyword in the database, such as "lolyta", 18 familiar terms that are written a bit differently, e.g., 10yr or kingspass, and 17 completely new terms. The new terms were checked using Urban Dictionary [16] and Google websearch, without entering any site with an illicit material. The list of ignored phrases is updated in accordance. This thus extended the dictionary by 115%. Four of the words in the extended dictionary are also ranked within a new top 12 pedophile query terms, with 842 to 3317 queries each.

One issue in extending the dictionary is the definition of child sex abuse terms. As definitions differ between countries, it is unclear whether terms such as "teensex" should be added or not. The heuristic discovers six such new terms, that relate to teens pornography.

Frequency The frequency of pedophile queries is relatively high, approximately 0.04% of all queries to the database. The queries are distributed across all hours of the day, with least queries being sent at 6am and most queries sent at 1am (user's local time). The pedophile queries frequency graph generally follows the global queries graph, with minor deviations, mainly caused by a slightly higher rate of pedophile queries during night hours and early morning. The high rate of pedophile queries is also interesting as the site scarcely contains child sex abuse torrents, yet we did not observe a decline in queries rate over time, as may be expected when pedophiles find out the contents of the site.

3.3 Downloads Analysis

Distinct Pedophile Downloads We detect in the Mininova database only 5 files (out of over a million) that include in their filename keywords taken from our dictionary and are not of a legal nature. These files are also manually checked and verified to be potential illicit material and not innocent ones[***]. We note that some files, such as torrents called "PTHC" are often used to target their leechers and spread viruses, however our focus here is the leechers and not the seeders.

The distinct 5 files contain five of the words included in our dictionary, and they are downloaded 1432 times within the dataset timeframe.

Correlation Between Downloads And Queries Tracking down pedophile activity in the Mininova dataset is a hard task, both in identifying child sex abuse material and processing the large database. A different challenge stems from the fact that while the dataset provides a torrent name, the meta data connected to this torrent, such as a description of the file and users comments, are not visible to us. As a result, it is difficult to correlate between queries and downloads, since the connection may reside in the hidden data. We use three

[***] based on filenames and web search, without viewing the actual file contents

parameters to overcome this obstacle: time, city and repetition pattern. We say that the query and download correlate if they both originate from the same city and the download is seen shortly after the query. If a suspected filename includes one of the dictionary terms, or if the set of downloads resulting from the restrictions on time and city include only a single entry, then the task is simple. however, this is not the common case. We thus say that a file is included in a set of suspected pedophile files if for multiple pedophile queries it is included in the resulting downloads set. Using our heuristic, we find a ratio of 1:30 between queries and downloads, while using a set of over 90K pedophile queries.

The first observation is that 7.2% of the queries result in no matched download.Another indication of success of the heuristic is that it detects downloads of the pedophile torrents with known keywords (as described in Section 3.3). We note in this group of downloads the reoccurrence of keywords with kids names, such as a torrent called "pthc 9yo jenny". It also detected pedophile torrents that contain in their filenames words with sexual connotation. We note that the majority of these files is pictures and not other types of multimedia.

We last detect a group of files with innocent names, that can easily be tracked back to pornographic material. An attempt to discover child pornography files with innocent names has failed so far.

We take some of the files with pedophile related keywords in their filenames and further investigate them. while most considered torrents include distinctive keywords, some torrents known to include child pornography may be the result of an occasional pornography search (for example, a nudist family movie whose content was verified). Figure 2 shows the number of downloads of four of these torrents as a result of a pedophilic query, compared to their overall number of downloads. The first three files, marked P1 through P3, have distinct pedophilic words in their names, such as PTHC and Raygold. File V4 is pedophilic in the wide sense, meaning its name includes pornographic but not pedophilic keywords in it, but its content is known to include a video of nude children. The selected files are downloaded only within the duration that we check the database, meaning their first and last downloads are in the timeframe our dataset was collected. V4 is the only exception, as its first download may have occurred before we started logging.

We take these downloads and cross them back to the queries generating from the same location in the time period before the file was downloaded. We find that many of the downloaded files are as a result of direct access to the page. For P3 only two queries were submitted that contain a pedophilic or a sex related word. For other files, we see that most of the downloads are also the result of direct access, either because no query was submitted from the origin city before the download time or because no pedophilic or pornographic related query was issued. For all four torrents, 23% to 67% of the downloads had no prior query, 15% to 34% of the downloads followed a query with a word from the torrent's name and 5% to 14% of the downloads can be related to a pedophilic or pornographic keyword in a previous query (except for P3). As in large cities,

such as Chicago, Paris or London, there are tens of queries every minute, we can not track the query directly to its source despite filtering out innocent queries.

Figure 3 shows for the same four files the download time distribution. Each sub-figure shows for one of the files the number of downloads every hour since the file was first downloaded. As the behavior differs significantly between files, the axis values are different. For file P1 and P3, the downloads peak in the first hour that the file is distributed, and decline afterwards (P3 is downloaded over 2 hours only). File P2 and V4 are downloaded over a long period of time, at a relatively constant low rate, with the gap during this time period caused by a gap in the data. interestingly, File V4 peaks after 3-4 days since the measurement begins, as opposed to the previous cases.

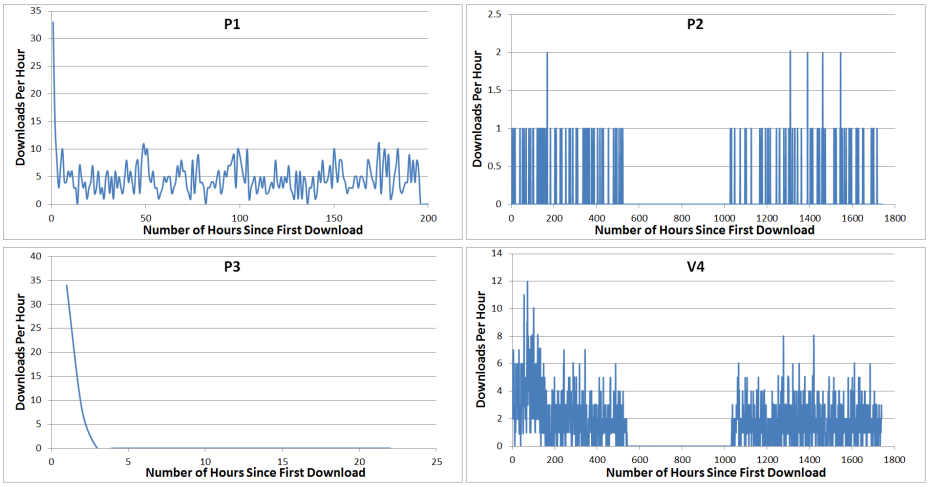

Fig. 3. Number of File Downloads per Hour

Geographic Aspects The geographic distribution of illicit downloads is spread around the world, in all continents. We take the four files discussed in the previous section and further investigate their download pattern. Following their trend in time, the downloads are spread across four continents from the time of the first download to the last.

Another interesting result checks the time difference between downloads from the same city. While the lifetime of P3 is not long enough to examine this, in P1 the time gap between two downloads from the same city is less then one day, for cities with more than a couple of downloads. In torrents P2 and V4, on the other hand, the gap between two downloads is often over a week and even over a month. The density of downloads from the same city is lower as well. We assume that this may be as the contents of P1 and P3 may be of "high quality", while P2 and V4 may be bogus or of lower interest, hence this result.

3.4 Behavior Analysis

We use the small cities heuristic described in 3.2 to explore the behavior of pedophile users in the Mininova database. We note that small cities behavior

may be different than a global view, but we believe that due to the large number of cities included in this analysis, it has a value.

The average number of queries in a standard visit to Mininova is measured (in the set of small cities) to be 2.6. In comparison, a visit which includes a pedophile term in it has on the average only 1.5 queries. The average gap between queries in such a visit is 32 seconds.

4 Conclusion

In this paper we presented an analysis of pedophilic activity in Mininova, a portal used by the BitTorrent network. We discussed how child pornography is spread through multimedia files and how the files can be detected on the BitTorrent network. The paper focused on the characteristics of the molesters looking for illicit material, as they manifest in their web activity, by time and content. We also suggest a way to expand the list of known pedophile keywords, and succeed to more than double our initial list. A repetitive run of this heuristic on recent peer-to-peer databases can assist law enforcement teams to detect pedophiles more efficiently.

References

1. Alexa. www.alexa.com [Accessed: November 17,2009].
2. M. Chopra, M. V. Martin, L. Rueda, and P. C. Hung. Toward new paradigms to combating internet child pornography. In *CCECE'06*, pages 1012 –1015, 2006.
3. Fives. Forensics Image and Video Examination Support. http://fives.kau.se/.
4. A. S. Gish, Y. Shavitt, and T. Tankel. Geographical statistics and characteristics of p2p query strings. In *In IPTPS07*, 2007.
5. D. Hughes, S. Gibson, J. Walkerdine, and G. Coulson. Is deviant behaviour the norm on p2p file sharing networks? *IEEE Distributed Systems Online*, 7, 2006.
6. isoHunt. isoHunt Zeitgeist. *http : //ca.isohunt.com/*[Accessed: April, 2011].
7. C. Jansohn, A. Ulges, and T. Breuel. Detecting pornographic video content by combining image features with motion information. In *Proceedings of the International Conference on Multimedia*. ACM, 10 2009.
8. M. Latapy, C. Magnien, and R. Fournier. Quantifying paedophile queries in a large p2p system. In *IEEE Infocom Mini-Conference*, 2011.
9. M. Liberatore, R. Erdely, T. Kerle, B. N. Levine, and C. Shields. Forensic Investigation of Peer-to-Peer File Sharing Networks. In *Proc. DFRWS Annual Digital Forensics Research Conference*, August 2010.
10. C. Lynn. *Image Recognition Takes Another Step Forward*. Seybold Report, 2004.
11. MAPAP. Measurement and Analysis of P2P activity Against Paedophile content. *http : //ec.europa.eu/information_society/activities/sip/projects/completed/illeg_content/index_en.htm*.
12. Mininova. http://www.mininova.org/.
13. K. Mochalski and H. Schulze. Ipoque internet study 2008/2009. 2009.
14. J. Ropelato. Internet pornography statistics. *TopTenReviews*, 2007.
15. TopTenReviews. Porn industry statistics. http://www.toptenreviews.com/2-6-04.html, Feb. 6 2004.
16. Urban Dictionary. http://www.urbandictionary.com/ [Accessed: February, 2011].
17. V. Vehovar, A. Ziberna, M. Kovacic, A. Mrvar, and M. Dousak. An empirical investigation of paedophile keywords in edonkey p2p network. *tech. report*, 2009.