On Child Sex Abuse Presence in BitTorrent Networks

Yuval Shavitt and Noa Zilberman

School of Electrical Engineering, Tel-Aviv University, Israel
shavitt,noa@eng.tau.ac.il

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 three major BitTorrent web portals. We detect and characterize child sex abuse material in the network, and also analyze different aspects of the abusers activity. We hope our results will help law enforcement teams put more focus on the BitTorrent network and ease the detection of new illicit material.

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 [8]. 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.

The effort to fight Internet child sex abuse (CSA) has increased significantly in the recent years. Many works, such as Lynn [6] and the FIVES project [2] try to detect child sex abuse content within files. A growing attention is given by law enforcement agencies peer-to-peer (P2P) networks, as a source for CSA material. The Internet Crimes Against Children (ICAC) is using the Roundup tool, developed by Liberatore *et al.* [5] to detect child sex offenders in the Gnutella, ARES and emule networks given a list of known files. Other widely used tools are Child Protection System/Peer Precision [9] and EspiaMule [13]. The Interpol manages the International Child Sexual Exploitation image database (ICSE DB) [11], which is not limited to P2P networks. New tools are being developed

[⋆] An early version of this paper appeared in PAM2012.

by the iCOP project, which also provides the most detailed review of works in this field [12]. Notable academic works are MAPAP [7], which focused on the eDonkey network, and Huges *et al.* [3], studying the Gnutella network. Liberatore *et al.* [5] discussed legal issues involved in investigating child pornography in the Gnutella and BitTorrent networks.

In this paper we present a study of child sex abuse activity in the BitTorrent network, examining behavioral patterns in both queries and downloads and studying geographical aspects and trends. 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 CSA terms can be unraveled and new illicit files can be detected.

2 Data Sets

2.1 Mininova

The Mininova website[⋆] was for a long time a 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 websites, with 1.07% of Internet users visiting it and first of torrent websites, ahead of portals such as The Pirate Bay (ranked 105), Torrentz.com (ranked 190) and isoHunt (ranked 196).

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

- Queries: Includes 453 million queries. Each registry contains the query text, a timestamp and its city of origin.
- Downloads: Holds 515 million torrent downloads, with over 1.3 million distinct torrents. An entry contains the torrent's name, torrent's subcategory, file size, timestamp and the city of origin.

Pornography was not common in the Mininova website and there is no category dedicated for it. Adult material was often placed under various categories, such as "Asian" or "Movies - Other".

2.2 The Pirate Bay

The Pirate Bay website[⋆⋆] is today the most popular BitTorrent portal. According to Alexa [1] the site is currently ranked 105 of all worldwide websites. Torrents on the website can be browsed or searched. While in the past the site hosted

[⋆] *http://www.mininova.org/*

[⋆⋆] *http://thepiratebay.se/*

actual torrent files, in 2012 it started hosting only magnetic links, which are essentially hyperlinks containing the hash code for torrents. Magnetic links are used by BitTorrent portals for legal reasons, as by using them the sites are no longer hosting files that link to copyrighted content.

The dataset used in this work was obtained from The Pirate Bay website, and takes a snapshot of all the magnetic links available on it on February 8th, 2012. It was uploaded to the website by an anonymous user, nicknamed ”allisfine”. The dataset contains information on 1.64 million magnetic links, including the torrent ID, torrent's name, file size, number of seeders and leechers and the magnetic link's hash.

2.3 BitSnoop

Bitsnoop[⋆⋆⋆] is a BitTorrent site launched on October 2009. It is globally ranked 1236 by Alexa [1] and 146 in South Korea, whose users are 17% of its visitors. Torrents on the website can be browsed or searched. The site hosts only magnetic links and is using bots to roam the Internet and find torrents.

The dataset used in this work was obtained from BitSnoop website, and takes a snapshot of all the magnetic links available on it on February 10th, 2012. It was uploaded by an unknown user. The dataset contains information on 17 million magnetic links, but includes only torrents' name and magnetic link's hash. The dataset is divided to 34 categories, such as video and games.

2.4 Data Set Limitations

The Mininova data set analysis has several limitations. The main limitation is users' anonymity, with only user's city available. This means that the activity of a specific user cannot be pinpointed, e.g. there is no clear distinction between users and activity sessions. The Mininova downloads database also lacks metadata information, making it difficult to classify the file and correlate between queries and downloads. The Pirate Bay and BitSnoop databases have no user information or metadata, thus limiting the span of the analysis. Last, there is no ground truth database of all the up to date terms used by pedophiles, as they try to update their vocabulary and hide changes from law enforcement teams. We believe that basing our dictionary assumptions on previous works (See Section 1) that corroborate researchers from multiple fields, including social sciences, provide an adequate baseline for our analysis.

3 Results

3.1 Mininova Queries Statistics

Keywords Ranking To identify CSA related material, a dictionary of related words was created. The dictionary relies on previous works in this area [15, 4]

[⋆⋆⋆] $http://bitsnoop.com$

and popular online sources [14]. We attempted to collect additional information from sources such as InHope (www.inhope.org), but failed to collaborate. The dictionary of CSA-related words used for this study initially included 47 words. Each of these words on its own has a pedophilic meaning, but in context may become innocent. For example, "Lolita" on its own versus the combination of "Lolita" and "Nabokov", referring to the known novel. A filter is applied to over 40 such known combinations. We note that the created dictionary may not be full, but show that these words alone are enough to portray a worrying picture.

Query	**Mininova**			**isoHunt**		**The Pirate**
	Occurrences	**% of Pedo.**	**% of All**	**2011**	**2012**	**Bay**
		Queries	**Queries**	**Ranking**	**Ranking**	**Ranking**
Lolita	26668	25.20%	0.0059%	135	232	170
Incest	26290	24.84%	0.0058%	143	299	151
Preteen	17910	16.93%	0.0039%	119	24	94
PTHC	10617	10.03%	0.0023%	83	30	N/A
Pedo	8406	7.94%	0.0018%	257	304	N/A
Underage	4756	4.49%	0.0010%	284	704	N/A
R@ygold	1594	1.50%	0.0003%	N/A	847	N/A
Hussyfan	1388	1.31%	0.0003%	955	510	N/A
Yamad	1325	1.25%	0.0003%	N/A	N/A	N/A
12yo	685	0.64%	0.0002%	275	580	N/A

Table 1. Statistics of Pedophilic Queries

Table 1 presents the Mininova's top-10 most used terms in CSA queries and their ranking in other BitTorrent portals. The table contains for each word the number of occurrences, percentage out of CSA related queries and percentage out of all queries. Next is the word's ranking in isoHunt[†] on April 2011 and June 2012, and its ranking in The Pirate Bay portal on June 2012. The words "lolita" and "incest" are the most popular pedophilic terms used in Mininova. While these two words may also relate to non-CSA contents, at least some of these queries are related, as we show in the next section. It is also observed that top keywords appear in one of every 25K to 100K queries out of all queries, which is considered high.

isoHunt's[‡] list does not provide information about the number of queries per term, rather it ranks them by their popularity. Furthermore, isoHunt filters some terms, with only "lolita" appearing in the unfiltered list. We see differences from Mininova's top-ranked list, as terms like "7yo", which was not amongst the top-30 queries, is being placed high in isoHunt's global search list (290 on 2011), and with the term "9yo" being ranked higher (274) than "12yo".

The Pirate Bay presents the top 500 search terms. It includes the terms Lolita, Incest and Preteen, but no other of our terms. Possibly, some filtering

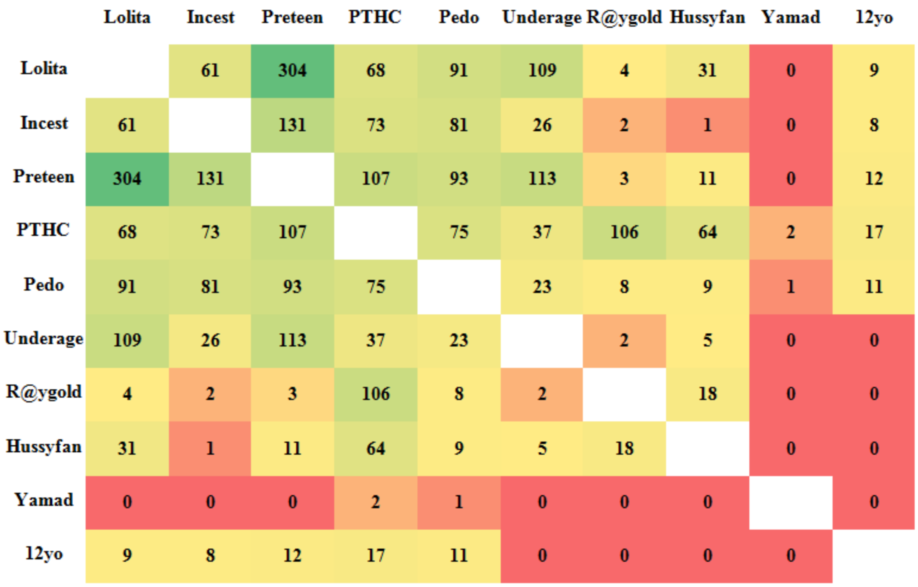
[†] *http : //ca.isohunt.com/*

[‡] isoHunt was ranked second amongst torrent websites [1] at sampling time.

is applied here too. BitSnoop provides top 100 queries and downloads, however none matches any of the dictionary's terms.

Comparing the results to MAPAP's eDonkey research [7], the term ”PTHC” is ranked first, followed by ”Pedo”, both searched considerably more than any other term. Other popular terms in BitTorrent, like ”Lolita”, are less popular in eDonkey, while terms like ”Preteen” and ”Underage” are not ranked at all.

Correlation Between Keywords Queries identified as CSA-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. The six highest ranking keywords are well connected, each one appearing tens to hundreds of times in queries with the other five keywords. We believe that such queries are being issued with the intent to find CSA torrents. The percentage of queries where two terms are used in conjunction is 3.8%, with some of the keywords collocated with other terms in over 10% of their appearances.



	Lolita	Incest	Preteen	PTHC	Pedo	Underage	R@ygold	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
Underage	109	26	113	37	23		2	5	0	0
R@ygold	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

On some occasions, connection between pedophile terms and ordinary words is possible. We rank words collocated in the same queries as pedophile terms and find that for all keywords, except for one case, there is no dominant single word appearing: never more than 10% of the keyword's queries. Three main types of words appear in conjunction with keywords: media type, pornography related, and names. Media type includes, for example, "video" and "pics". Words that fall under the category "pornography" include terms like "sex" and "porn". The last group of words includes personal names like "Vicky", "Jenny" and "Daphne", issued together with keywords as "PTSC" and "PTHC". A troubling aspect is when these words are collocated with age indication, like "9yo jenny". While this sounds as a naive query, searching this term on the web leads to tens of discussions in CSA forums with a clear description of the movie's contents. We leverage this information later to extend our CSA terms' dictionary.

Extending The Dictionary An important contribution is detecting new terms that relate to CSA, which is a hard task in an anonymous database. For this end, we analyze Mininova's queries from each city separately, 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 look 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 find that in 98.5% of the cases the length is no longer than five minutes and the number of queries is at most 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's [1] finding that the average site visit time was 4.3 minutes. For further analysis we used cities that contain only distinct SUBPs, at least 10 SUBPs and that registered at least one pedophile query. This results 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. Using natural language analysis tools [10] such as comparative frequency analysis we screen out of these words numbers, conjunctions and terms highly ranked in the global queries list (such as "Harry Potter"). This leaves 140 words. We classify those to four groups: 51 General sex related words, 29 potential victims' names, 54 CSA terms, and 7 words that may refer to either general pornography or child sex abuse. The 54 new words include 19 words spelled differently 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 validated using Urban Dictionary [14] and Google websearch, without entering any site with an illicit material. The list of ignored phrases is updated in accordance. This extends the dictionary by 115%. Four of the words in the extended dictionary are also ranked within a new top 12 pedophile query terms, each with 842 to 3317 queries.

One issue in extending the dictionary is the definition of CSA terms. As legal definitions differ between countries, it is unclear whether terms as "teensex" should be included. The heuristic discovers six such new terms relating to teen pornography.

Geographic Distribution According to Alexa [1], visitors to Mininova website arrive mostly from The US (16.8%), India (11.9%), UK (5.1%), Italy (4.2%) and France (4.1%). We use this information, combined with Mininova's dataset, to explore the geographic distribution of CSA queries. The leading countries in the absolute number of CSA queries are The US (21%), Italy (19%), India (11%), UK (6%) and France (6%), matching the top five visitor's countries mentioned above. An interesting question is from where originate the highest percentage of pedophile queries relative to the absolute number of queries per city. The city with the highest rate of pedophile queries is Chicago (0.1%), Moscow(0.05%), Islamabad, Hyderabad, Seol, Riyad and Bangkok follow (all 0.04%). Ranked lowest are cities like Paris, Singapore and Toronto with a relative percentage of 0.01%.

City	**Country**	**% of Pedo. Queries**	**% of All Queries**
Chicago	USA	0.1%	0.96%
Moscow	Russia	0.05%	0.15%
Islamabad	Pakistan	0.04%	0.27%
Hyderabad	India	0.04%	0.21%
Seoul	S. Korea	0.04%	0.4%
Riyadh	Saudi Arabia	0.04%	0.37%
Bangkok	Thailand	0.04%	0.23%

Table 2. Statistics of Pedophilic Queries by Highest Ranked Cities

3.2 Mininova Downloads

We detect in the Mininova's database only five 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 spread viruses. The five files are downloaded 1432 times within the dataset timeframe.

We take a few of the detected files and further investigate them. 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, but its content is known to include a video of nude children. The selected files are downloaded only within the dataset's timeframe, except for V4, whose first download may have occurred before we started logging.

We take these downloads and cross them back to queries generated from the same location in the time period before the file was downloaded. We find that many of the downloads 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

[§] based on filenames and web search, without viewing the actual file contents 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).

The geographic distribution of illicit downloads is spread around the world: the users downloading the four investigated torrents are located all across four continents (excluding Africa).

3.3 Pirate Bay Downloads

Running our original dictionary over the magnetic links database resulted in 1078 torrents, and over 2200 torrents when using the extended dictionary. We found the terms "young girl" or "young boy" have a high collocation score and thus used them to extend the dictionary further. This increased the number of suspected files to almost 2500.

The names of the files are many times very descriptive, indicating the expected contents of the video or image. These files can be easily classified as CSA related or not. The other extreme is files with only the name of its source website and a name of the person supposed to appear in it. In such cases it is hard to classify the file as CSA or common pornography, though the source website reputation can serve as an indicator. Another problem with classifying files is that the age of the people appearing in it is unclear: When a file name indicates that young girls are involved, their actual age may be above the legal definition for child abuse. The nature of the files is sometimes unmistakable (though they may be false). Such an example is a torrent called "2 young girls kidnapped and raped". On the other hand, a torrent called "Necropedophilia Masterbate Video" is in fact a clip of a music band.

The most popular terms used in torrents names are "teensex" (765 times), "incest" (539), "lolita" (339) and "young girl" (237). All these torrents may in fact not be CSA material, involving people above the consent age. The pedophilic terms preteen, pthc and pedo, ranked high in queries (by Mininova and isoHunt), had each three or less torrents.

The average numbers of seeders per file is 3.9 and the average number of leechers is 2.5. 1855 of the files have at least one seeder and 1812 of them has at least one leecher. 10% of the files have more than 10 seeders and 3.5% have 10 leechers or more. The most downloaded torrent on this list has 203 seeders and 104 leechers, and the runner up has 163 seeders and 67 leechers. For both torrents it is unclear whether their content is truly CSA: users comments in the torrent's webpage indicate that the girls appearing in it may vary in age from 14 to early twenties. However, users cannot know the actual content until they have downloaded the file. While the number of leechers seems to be lower than the number of seeders, this may be misleading as a leecher automatically becomes a seeder once he downloaded the file, until he removes it.

We randomly select twenty suspected torrents that are with high probability of child sex abuse, and check their status in the Pirate Bay on June-2012. In all cases, the torrents were still available through the website with at least one seeder or leecher. As the upload date of a torrent appears on the website, we note that some such torrents were uploaded at 2004 and are still active.

The Pirate Bay website refers to possible CSA material hosting, and asks to report child sex abuse to the local authorities.

3.4 BitSnoop Downloads

Running our original dictionary over Bitsnoop's database resulted in 6171 torrents, and over 7678 using the extended dictionary. Adding the terms ”young girl” or ”young boy” increased the number of suspected files to 9441.

The most popular terms used in torrents names are ”lolita” (2718 times), ”incest” (2118 times), and ”young girl” (1646 times). The term ”teensex”, ranked highest in Pirate Bay's torrents, is ranked here fourth with 649 mentions in torrents names. We note is that the term ”PTNN” - Pre-Teen Non-Nude, which was ranked low amongst queries (outside top-20), is within the top-10 torrents, with 117 torrents. For reference, in the Pirate Bay database there was one torrent with this keyword. Another finding is that keywords are often broken into segments, separated by spaces, presumably to avoid detection. This means that a terms such as ”PTHC”, ”Pedo” and ”preteen” are written as ”PTH C”, ”Ped O” and ”pr et een”, possibly as a mean to avoid automatic filtering. Identifying these torrents is an indication of the strength of our method. When we add these terms to our dictionary and discover 60 additional suspicious torrents.

As before, we randomly select a group of torrents that are with high probability CSA, and check their current status in BitSnoop. All the torrents are still available through the website, but some of them have no seeders or leechers. Also, some of the files are now provided only as a direct download link.

Last, the similarity between torrents on BitSnoop and The Pirate Bay is checked. Out of the Top-10 torrents in the Pirate Bay, in terms of seeders and leechers, only four exist on BitSnoop. The highest ranking torrent in the Pirate Bay had no seeders and one leecher in BitSnoop (tested June 2012).

BitSnoop has a policy against child pornography, stated on their website, and they claim to use filters to avoid indexing of such torrents. They also provide a contact for reporting illegal torrents and a reported link is immediately removed from the website.

4 Conclusion

This paper provides a first large-scale analysis of CSA activity in the BitTorrent network, through three large Torrent datasets and reflecting changes over a period of several years. We show that there are differences in terms and files over the portals that may be related to the cultural differences of the portal's users. The paper also shows that the BitTorrent network is used mainly for video trafficking and less for CSA image sharing, with worldwide spread usage. In addition, we show that in the common case there are more people distributing a file (seeders) than downloading it (leechers). As a user becomes a seeder (intentionally or unintentionally) he actively distributes the file, increasing by this the probability to be caught and possibly the severity of the offence. Last, we provide means to extend the dictionary of CSA terms and detect new CSA files. The method, employing natural language analysis techniques, improves previous methods by focusing only on sessions of potential CSA consumers, thus increasing the relative part of new CSA terms in the detected set of terms and reducing the chances for false positives. An advantage of our method is that it can be done while keeping the anonymity of users, therefore reducing possible legal requirements up until the stage that a suspect has to be identified. Identifying new CSA files is very important for catching their creators: only the first seeder of a file is with high probability related to its creators.

For future work, it will be important to collaborate with law enforcement agencies and to gain ground truth information that can corroborate the study's results. Law enforcement agency involvement may also provide more research flexibility by removing some legal limitations. Last, future research should study files shared by users (as opposed to queries and downloads), which is an offense sought by the police.

5 Acknowledgements

We would like to thanks Moshe Rutgaizer and Omer Vertman who took part in the early work on this project. We would also like to thank Erik Dubbelboer who provided the Mininova database.

References

1. Alexa. www.alexa.com.
2. Fives. Forensics Image and Video Examination Support. http://fives.kau.se/.
3. D. Hughes, S. Gibson, J. Walkerdine, and G. Coulson. Is deviant behaviour the norm on p2p file sharing networks? *IEEE Distributed Systems Online*, 2006.
4. M. Latapy, C. Magnien, and R. Fournier. Quantifying paedophile queries in a large p2p system. In *IEEE Infocom Mini-Conference*, 2011.
5. 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.
6. C. Lynn. *Image Recognition Takes Another Step Forward.* Seybold Report, 2004.
7. 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.*
8. K. Mochalski and H. Schulze. Ipoque internet study 2008/2009. 2009.
9. Flint Waters. Challenges and Solutions for Protecting our Children from Violence and Exploitation in the 21st Century. Testimony to United States Senate Committee On The Judiciary Subcommittee On Crime And Drugs, 2007.

10. Hughes, Danny and Rayson, Paul and Walkerdine, James and Lee, Kevin and Greenwood, Phil and Rashid, Awais and May-Chahal, Corinne and Brennan, Margaret. Supporting Law Enforcement in Digital Communities through Natural Language Analysis. In *IWCF*, 2008.
11. Interpol. Crimes against children. FactSheet, COM/FS/2009-09/THB-03, 2009.
12. Maggie Brennan and Sean Hammond. Complete Critical Literature Review. Report, iCOP, Identifying and Catching Originators in P2P Networks, 2011.
13. Paulo Fagundes. Fighting Internet Child Pornography The Brazilian Experience. *The Police Chief*, 2009.
14. Urban Dictionary. http://www.urbandictionary.com/.
15. V. Vehovar, A. Ziberna, M. Kovacic, A. Mrvar, and M. Dousak. An empirical investigation of paedophile keywords in edonkey p2p network. *tech. report*, 2009.

Yuval Shavitt received the B. Sc. in Computer Engineering (cum laude), M. Sc. in Electrical Engineering and D. Sc. from the Technion — Israel Institute of Technology, Haifa, Israel in 1986, 1992, and 1996, respectively. After graduation he spent a year as a Postdoctoral Fellow at the Department of Computer Science at Johns Hopkins University, Baltimore, MD. Between 1997 and 2001 he was a Member of Technical Stuff at Bell Labs, Lucent Technologies, Holmdel, NJ. Starting October 2000, he is a Faculty Member in the School of Electrical Engineering at Tel-Aviv University, Israel. His research interests include Internet measurements, mapping, and characterization; and data mining peer-to-peer networks.

Noa Zilberman received her B.Sc. and M.Sc. (both magna cum laude) in Electrical Engineering from Tel-Aviv University, Israel in 2003 and 2007, respectively. Since 1999 she has filled several development, architecture and managerial roles in the telecommunications industry. She is currently a Ph.D. candidate in the School of Electrical Engineering at Tel-Aviv University. Her research focuses on Internet measurements, mapping, and characterization; and data mining peer-to-peer networks.