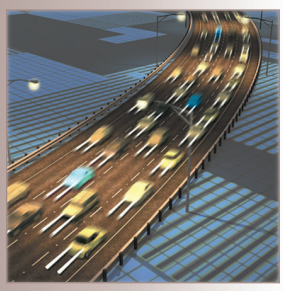

On the Presence of Child Sex Abuse in BitTorrent Networks

Peer-to-peer networks' widespread use makes multimedia files available to users worldwide. However, such networks are often used to spread illegal material while the data source and acquiring users remain anonymous. This article analyzes activity measurements in the BitTorrent network and examines child sex abuse activity through three major BitTorrent Web portals.

Yuval Shavitt and Noa Zilberman
Tel-Aviv University

Millions of users employ peer-to-peer (P2P) networks worldwide to share content. Although much of this activity is legal, the anonymity such networks provide enables some users to share illegal content, from simple copyright-protected works to highly dangerous material.

The BitTorrent file-sharing network was responsible for 27 to 55 percent of Internet traffic (depending on geographic location) in 2009.[1] BitTorrent lets users download large files without being a burden on a single source computer; rather, 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*, that's distributed via email or HTTP websites. This torrent file lets downloading and uploading users — or *leechers* and *seeders*, respectively — share the content.

Efforts to fight Internet child sex abuse (CSA) have increased significantly in recent years. Many efforts try to detect CSA content within files (see http://fives.kau.se/).[2] Law enforcement agencies are paying increased attention to P2P networks as a source for CSA material. The Internet Crimes Against Children (ICAC) Task Force uses the Roundup tool[3] to detect child sex offenders in the Gnutella, ARES, and eMule networks given a list of known files. Other widely used tools include the Child Protection System/Peer Precision[4] and EspiaMule.[5] Interpol manages the International Child Sexual Exploitation image database (ICSE DB),[6] which isn't limited to P2P networks. The Identifying and Catching Originators in P2P Networks (iCOP) project is developing new tools and also provides the most detailed review of works in this field.[7] Notable academic works include the Measurement and Analysis of P2P Activity against Paedophile Content (MAPAP) project (http://ec.europa.eu/information_society/activities/sip/projects/completed/illeg_content/index_en.htm), which focuses on the eDonkey network, and research from Danny Hughes and his colleagues that studies the Gnutella network.[8] Marc Liberatore and his colleagues discuss the legal issues involved in investigating child pornography in the Gnutella and BitTorrent networks.[3]

Published by the IEEE Computer Society
1089-7801/13/$31.00 © 2013 IEEE

Here, we present a study of CSA activity in the BitTorrent network, examining behavioral patterns in both queries and downloads and studying geographical aspects and trends. Law enforcement forces can employ our results to detect and track pedophiles — for example, using the given analysis, they can unravel new CSA terms and detect new illicit files.

Datasets

We use three datasets, collected from popular BitTorrent portals, in this work. We obtained one directly from the portal's owners, and downloaded the other two from the portals themselves. Because we collected these datasets from top BitTorrent portals, they're representative of the BitTorrent network.

Mininova

The Mininova website (www.mininova.org) was a popular BitTorrent portal for a long time, until a court order forced it to remove all copyrighted torrents at the end of 2009. At that time, the site was ranked 90th in the world, according to Alexa (www.alexa.com), with 1.07 percent of Internet users visiting it; it was the top torrent website, ahead of portals such as the Pirate Bay (ranked 105), Torrentz.com (ranked 190), and isoHunt (ranked 196).

We obtained the Mininova dataset we use here from the Mininova team; it covers two time periods in 2009: 2 September to 25 September, and 15 October to 7 December. The dataset was anonymized before we received it, with users' IP addresses removed. The dataset comprises two types of information:

- 453 million queries, each with a registry that contains the query text, a time stamp, and its city of origin.
- 515 million torrent downloads, with more than 1.3 million distinct torrents. An entry contains the torrent's name, subcategory, the file size, a time stamp, and the city of origin.

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

The Pirate Bay

The Pirate Bay (http://thepiratebay.se/) is the most popular BitTorrent portal today. Alexa reports that the site is currently ranked 105th worldwide. On the Pirate Bay, users can browse or search torrents. Although in the past, the site hosted actual torrent files, in 2012 it began hosting only magnetic links, which are essentially hyperlinks containing the hash code for torrents. BitTorrent portals use magnetic links for legal reasons because, by doing so, the sites no longer host files that link to copyrighted content.

The dataset we use here came from the Pirate Bay website and is essentially a snapshot of all the magnetic links available on it on 8 February 2012. This snapshot 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, the file size, the number of seeders and leechers, and the magnetic link's hash.

BitSnoop

BitSnoop (http://bitsnoop.com) is a BitTorrent site launched in October 2009. Alexa ranks it at 1,236 globally and at 146 in South Korea, whose users comprise 17 percent of its visitors. Users can browse or search torrents on the site, which hosts only magnetic links and uses bots to roam the Internet and find torrents.

We obtained the dataset — a snapshot of all the magnetic links available on 10 February 2012 — from the BitSnoop website. It was uploaded by an unknown user. The dataset contains information on 17 million magnetic links, but includes only the torrent name and magnetic link's hash. The dataset was divided into 34 categories, such as video and games.

Dataset Limitations

Our Mininova dataset analysis has several limitations, mainly surrounding user anonymity because only the user's city is available. This means that we can't pinpoint a specific user's activity — for instance, no clear distinction exists between users and activity sessions. The Mininova downloads database also lacks metadata information, making it difficult to classify files and correlate between queries and downloads. The Pirate Bay and BitSnoop databases have no user information or metadata, thus limiting the span of our analysis. Finally, no ground-truth database exists for all the up-to-date terms pedophiles use; they often try to update their vocabulary and hide changes from

Table 1. Statistics on pedophilic queries.

Query	Mininova			IsoHunt ranking		Pirate Bay ranking
	Occurrences	Pedophilic queries (%)	All queries (%)	2011	2012	
Lolita	26,668	25.20	0.0059	135	232	170
Incest	26,290	24.84	0.0058	143	299	151
Preteen	17,910	16.93	0.0039	119	24	94
PTHC	10,617	10.03	0.0023	83	30	N/A
Pedo	8,406	7.94	0.0018	257	304	N/A
Underage	4,756	4.49	0.0010	284	704	N/A
R@ygold	1,594	1.50	0.0003	N/A	847	N/A
Hussyfan	1,388	1.31	0.0003	955	510	N/A
Yamad	1,325	1.25	0.0003	N/A	N/A	N/A
12yo	685	0.64	0.0002	275	580	N/A

law enforcement teams. We believe that basing our dictionary assumptions on previous works that corroborate research from multiple fields, including social sciences, provides an adequate baseline for our analysis.

Results

Here, we introduce the results of our dataset analysis. We begin by analyzing queries in the Mininova dataset and continue with an analysis of downloads from all three datasets.

Mininova Query Statistics

Our analysis of queries in the Mininova database considers several aspects, including popular keywords and the correlation between them, as well as geographic distribution. We used this information to extend the dictionary of CSA terms.

Keyword ranking. To identify CSA material, we created a dictionary of related words relying on previous work in this area[9,10] and popular online sources such as Urban Dictionary (www.urbandictionary.com). We attempted to collect additional information from sources such as InHope (www.inhope.org) with no success. Our dictionary of CSA-related words initially included 47 terms, each of which has a pedophilic meaning on its own, but can become innocent in context — for example, "lolita" on its own versus the combination of "lolita" and "nabokov," referring to the well-known novel. We applied a filter to more than 40 such known combinations. Although our dictionary might not be complete, it demonstrates that these words alone are enough to paint a worrying picture.

Table 1 presents Mininova's top 10 most-used terms in CSA queries and their ranking in other BitTorrent portals. For each word, the table contains the number of occurrences, its percentage among all CSA-related queries, and its percentage among all queries. We also show the word's isoHunt ranking (http://ca.isohunt.com) in April 2011 and June 2012, and its ranking in the Pirate Bay portal on June 2012. The words "lolita" and "incest" were the most popular pedophilic terms used in Mininova. Although they might also relate to non-CSA content, at least some of these queries are related, as we show in the next section. We also observed that each of the top keywords appeared in one out of every 25,000 to 100,000 total queries, which was high.

The isoHunt list doesn't provide information about the number of queries per term but rather ranks them by their popularity (Alexa ranked isoHunt second among torrent websites at sampling time). Furthermore, isoHunt filters some terms, with only "lolita" appearing on the unfiltered list. We see differences from Mininova's top-ranked list, with terms such as "7yo" — which wasn't among Mininova's top 30 queries — placed high in isoHunt's global search list (290 in 2011), and the term "9yo" having a higher rank (274) than "12yo."

The Pirate Bay provided its top 500 search terms, including the terms "lolita," "incest," and "preteen," but no other terms from our list. Possibly, some filtering was applied here, too. BitSnoop provided its top 100 queries and downloads, but none matched any of our dictionary's terms.

Comparing the results to MAPAP's eDonkey research, the term "PTHC" is ranked first, followed

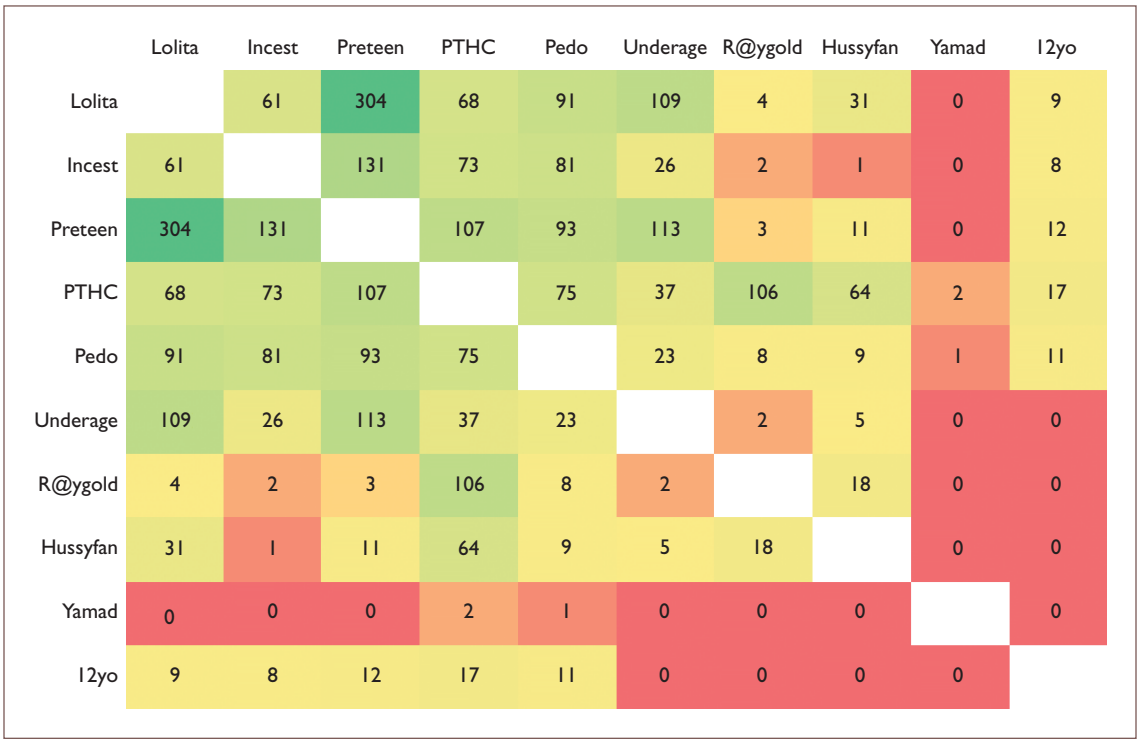

Figure 1. Heatmap of keyword appearance in the same queries.

by "pedo"; both were searched considerably more than any other term. Other popular terms in BitTorrent, such as "lolita," were less popular in eDonkey, whereas terms such as "preteen" and "underage" weren't ranked at all.

Correlation between keywords. Queries identified as CSA-related often include multiple terms that are pedophilic in nature. Figure 1 presents a heatmap of keywords that appeared together in the same queries. It shows only the highest-ranked keywords. The six highest-ranking keywords were well connected, each one appearing tens to hundreds of times in queries with the other five. We believe that users were issuing such queries to find CSA torrents. Two terms were used in conjunction in 3.8 percent of the queries, with some keywords collocated with other terms in more than 10 percent of their appearances.

On some occasions, we can connect pedophilic terms and ordinary words. We ranked words collocated in the same queries as pedophilic terms and found that for all but one keyword, no dominant single word appeared — that is, appeared in more than 10 percent of the keyword's queries. Three main types of words appeared in conjunction with keywords: media type, pornography-related, and names. Media type included "video" and "pics," whereas pornography-related words included terms such as "sex" and "porn." The last group of words included personal names such as "Vicky," "Jenny," and "Daphne," issued together with keywords such as "PTSC" and "PTHC." A troubling aspect emerged when these words were collocated with age indication, such as "9yo jenny." Although this sounds like a naive query, searching this term on the Web leads to tens of discussions in CSA forums with a clear description of, for example, a video's contents. We leveraged this information later to extend our CSA term dictionary.

Extending the dictionary. An important contribution is detecting new terms that relate to CSA, which is a hard task in an anonymous database. To this end, we analyzed Mininova's queries from each city separately, and defined 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 was an aggregation of multiple users and could be quite long. We looked for cities with sparse accesses to Mininova, where the probability that two user sessions would fall into the same busy period was negligible.

We analyzed the busy periods' length in cities with an average of 500 queries a day or less.

Table 2. Pedophilic queries by highest-ranked cities.

City	Country	Pedophilic queries (percent)	All queries (percent)
Chicago	US	0.1	0.96
Moscow	Russia	0.05	0.15
Islamabad	Pakistan	0.04	0.27
Hyderabad	India	0.04	0.21
Seoul	South Korea	0.04	0.4
Riyadh	Saudi Arabia	0.04	0.37
Bangkok	Thailand	0.04	0.23

We found that, in 98.5 percent of the cases, the length was no longer than 5 minutes and the number of queries was at most 10. We thus assume that these busy periods occurred due to a single user's activity and defined a single user busy period (SUBP) as one up to 5 minutes long and with up to 10 queries. This is in line with Alexa's finding that the average site visit time was 4.3 minutes. We used cities that contained at least 10 distinct SUBPs and registered at least one pedophilic query, resulting in 692 cities.

We then found the SUBPs that used pedophilic terms in queries and created a list of potential new keywords. This list initially had about 2,000 words. Using natural language analysis tools,[11] such as comparative frequency analysis, we screened out numbers, conjunctions, and terms highly ranked in the global queries list (such as "Harry Potter"). This left 140 words. We classified those into four groups: 51 general sex-related words, 29 potential victims' names, 54 CSA terms, and seven words that might refer to either general pornography or CSA. The 54 new words included 19 words spelled differently than an existing keyword in the database, such as "lolyta"; 18 familiar terms that were written a bit differently, as in "10yr" or "kingspass"; and 17 completely new terms. We validated these new terms using Urban Dictionary and Google Web Search, without entering any site with illicit material. We updated the list of phrases to ignore in accordance. This extended the dictionary by 115 percent. Four of the words in the extended dictionary are also ranked within a new list of top 12 pedophilic query terms, each with 842 to 3,317 queries.

One issue in extending the dictionary is defining CSA terms. Because legal definitions differ among countries, it's unclear whether terms such as "teensex" should be included; our heuristic discovered six similar new terms related to teen pornography.

Geographic distribution. According to Alexa, visitors to the Mininova website arrive mostly from the US (16.8 percent), India (11.9 percent), the UK (5.1 percent), Italy (4.2 percent), and France (4.1 percent). We used 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 were the US (21 percent), Italy (19 percent), India (11 percent), the UK (6 percent), and France (6 percent), matching the top five aforementioned visitor countries. Given this, an interesting question arises: Where do the highest percentage of pedophilic queries originate from relative to the absolute number of queries per city? We found that the city with the highest rate of such queries, as indicated in Table 2, was Chicago (0.1 percent), followed by Moscow (0.05 percent); Islamabad, Hyderabad, Seoul, Riyadh, and Bangkok were next at 0.04 percent each. Ranked lowest were cities such as Paris, Singapore, and Toronto, with a relative percentage of 0.01 percent.

Mininova Downloads

In the Mininova database, we detected only five files (out of more than a million) that included in their filename keywords taken from our dictionary and thus contain illegal content. We also manually checked these files (based on filenames and Web search, not by viewing the actual contents) and verified that they were potentially illicit material. Some files, such as torrents called "PTHC," are often used to spread viruses. The five files we detected were downloaded 1,432 times within the dataset time frame.

We further investigated a few of the detected files. The first three files, P1 through P3, had distinct pedophilic words in their names, such as "PTHC" and "Raygold." File V4 was pedophilic in a broad sense, meaning its name included pornographic but not pedophilic keywords, but its content was known to include a video of nude children. The selected files were downloaded only within the dataset's timeframe, except for V4, whose first download might have occurred before we started logging.

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

occurred as a result of direct access to the page. For P3, only two queries were submitted that contained a pedophilic or sex-related word. For the other files, we found that most of the downloads were 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 percent of the downloads had no prior query, 15 to 34 percent followed a query with a word from the torrent's name, and 5 to 14 percent could be tied back to a pedophilic or pornographic keyword in a previous query (except for P3).

The geographic distribution of illicit downloads is spread across four continents.

Pirate Bay Downloads

Running our original dictionary over the magnetic links database from the Pirate Bay resulted in 1,078 torrents and more than 2,200 torrents when we used the extended dictionary. We found the terms "young girl" or "young boy" to have a high collocation score and thus used them to extend the dictionary further. This increased the number of suspected files to almost 2,500.

File names are often very descriptive, indicating a video or image's expected contents. We can easily classify these files as CSA-related or not. The other extreme is files with only the name of the source website and the person supposed to appear in it. In such cases, classifying the file as CSA or common pornography is difficult, although the source website's reputation can serve as an indicator. Another problem with classifying files is that the age of those appearing in it is unclear: when a file name indicates that young girls are involved, their actual age might be above the legal definition for child abuse. The files' nature is sometimes unmistakable (although they might be false). One 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 torrent names are "teensex" (765 times), "incest" (539), "lolita" (339), and "young girl" (237). All these torrents might 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) — each had fewer than four torrents.

The average number of seeders per file was 3.9, and the average number of leechers was 2.5; 1,855 files had at least one seeder and 1,812 at least one leecher. Ten percent of the files had more than 10 seeders, and 3.5 percent had 10 leechers or more. The most downloaded torrent on this list had 203 seeders and 104 leechers, and the runner up had 163 seeders and 67 leechers. For both torrents, it was unclear whether their content was truly CSA: user comments on the torrent's webpage indicated that the girls appearing in it might vary in age from 14 to their early 20s. However, users can't know the actual content until they've downloaded the file. Although the number of leechers seemed to be lower than the number of seeders, this might be misleading because leechers automatically become seeders once they've downloaded a file and remain so until they remove it.

We randomly selected 20 suspected torrents that had a high probability of being CSA and checked their status in the Pirate Bay database as of June 2012. In all cases, the torrents were still available through the website with at least one seeder or leecher. Because a torrent's upload date appears on the website, we could tell that some such torrents were uploaded in 2004 and are still active. The Pirate Bay website does refer to possible CSA material hosting and asks users to report it to the local authorities.

BitSnoop Downloads

Running our original dictionary over BitSnoop's database resulted in 6,171 torrents, and more than 7,678 when we used the extended dictionary. Adding the terms "young girl" or "young boy" increased the number of suspected files to 9,441.

The most popular terms used in torrent names were "lolita" (2,718 times), "incest" (2,118 times), and "young girl" (1,646 times). The term "teensex," which ranked highest in the Pirate Bay's torrents, was ranked 4th here, with 649 mentions in torrent names. The term "PTNN," which was ranked low (outside the top 20) among queries, was within the top 10 torrents, with 117. For reference, in the Pirate Bay database, one torrent had this keyword. Another finding was that keywords are often broken into segments and separated by spaces, presumably to avoid detection. This means that terms such as "PTHC," "pedo," and "preteen" are written as "PTH C," "Ped O," and "pr et een," possibly as a way to avoid automatic filtering. That we can identify

these torrents indicates our method's strength. When we added these terms to our dictionary, we discovered 60 additional suspicious torrents.

As before, we randomly selected a group of torrents that had a high probability of being CSA and checked their current status in BitSnoop. All the torrents were still available through the website, but some had no seeders or leechers. Also, some files were provided at that point only as a direct download link.

Finally, we compared the similarity between torrents on BitSnoop and the Pirate Bay. Out of the top 10 torrents on the Pirate Bay in terms of seeders and leechers, only four existed on BitSnoop. The highest-ranking torrent in the Pirate Bay database had no seeders and one leecher in BitSnoop (tested June 2012).

BitSnoop has a policy against child pornography, stated on its website, and it claims to use filters to avoid indexing such torrents. The site also provides a contact for reporting illegal torrents, and a reported link is immediately removed from the website.

Our results show that the BitTorrent network, which has worldwide usage, is used mainly for video trafficking and less for CSA image sharing. Commonly, more people distribute a file (seeders) than download it (leechers). As a user becomes a seeder (intentionally or unintentionally), he or she actively distributes the file, which increases the probability of detection and possibly the severity of the offense. One advantage of our method is that it can be enacted while maintaining user anonymity, therefore reducing possible legal requirements up until the stage that a suspect must be identified. Identifying new CSA files is very important for catching their creators: only a file's first seeder has a high probability of being 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 our study's results. Such agency involvement could also provide more research flexibility by removing some legal limitations. Finally, future research should study files that users share — rather than simply query and download — which is an offense that the police frequently prosecute.

Acknowledgments

We thank Moshe Rutgaizer and Omer Vertman, who took part in the early work on this project. We also thank Erik Dubbelboer, who provided the Mininova database. An early version of this article appeared in the *Proceedings of the 2012 Passive and Active Measurement Conference*.

References

1. K. Mochalski and H. Schulze, *Internet Study 2008/2009*, Ipoque, 2009.
2. C. Lynn, *Image Recognition Takes Another Step Forward*, tech. report, Seybold, 2004.
3. M. Liberatore et al., "Forensic Investigation of Peer-to-Peer File Sharing Networks," *Proc. Digital Forensic Research Workshop Ann. Digital Forensics Research Conf.*, Elsevier, 2010; doi:10.1016/j.diin.2010.05.012.
4. F. Waters, "Challenges and Solutions for Protecting our Children from Violence and Exploitation in the 21st Century," testimony to US Senate Committee on the Judiciary, Subcommittee on Crime and Drugs, 2007.
5. P. Fagundes, "Fighting Internet Child Pornography: The Brazilian Experience," *The Police Chief*, Sept. 2009.
6. "Crimes against Children," Interpol fact sheet COM/FS/2009-09/THB-03, 2009.
7. M. Brennan and S. Hammond, *Complete Critical Literature Rev.*, tech. report, Identifying and Catching Originators in P2P Networks (iCOP), 2011.
8. D. Hughes et al., "Peer-to-Peer: Is Deviant Behavior the Norm on P2P File Sharing Networks?" *IEEE Distributed Systems Online*, vol. 7, no. 2, 2006.
9. V. Vehovar et al., *An Empirical Investigation of Paedophile Keywords in eDonkey P2P Network*, tech. report, Measurement and Analysis of P2P Activity against Pedophile Content Project, 2009.
10. M. Latapy, C. Magnien, and R. Fournier, "Quantifying Paedophile Queries in a Large P2P System," *Proc. IEEE Infocom Mini Conf.*, IEEE, 2011.
11. D. Hughes et al., "Supporting Law Enforcement in Digital Communities through Natural Language Analysis," *Proc. 2nd Int'l Workshop Computational Forensics*, Springer, 2008, pp. 122–134.

Yuval Shavitt 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. Shavitt has a DSc from the Technion — Israel Institute of Technology, Haifa. Contact him at shavitt@eng.tau.ac.il.

Noa Zilberman is a PhD candidate in the School of Electrical Engineering at Tel-Aviv University, Israel. Her research focuses on Internet measurements, mapping, and characterization, and data mining peer-to-peer networks. Zilberman has an MSc in electrical engineering from Tel-Aviv University. Contact her at noa@eng.tau.ac.il.