Child Abuse & Neglect 33 (2009) 560–568

Contents lists available at ScienceDirect

Child Abuse & Neglect

Child pornography in peer-to-peer networks

Chad M.S. Steel*

Virginia Polytechnic Institute and State University, Department of Computer Science, Falls Church, VA 22040, USA

ARTICLE INFO

Article history:
Received 11 June 2008
Received in revised form 9 December 2008
Accepted 18 December 2008
Available online 13 September 2009

Keywords:
Child pornography
Peer-to-peer
Internet

ABSTRACT

Objective: The presence of child pornography in peer-to-peer networks is not disputed, but there has been little effort done to quantify and analyze the distribution and nature of that content to-date. By performing an analysis of queries and query hits on the largest peer-to-peer network, we are able to both quantify and describe the nature of querying by child pornographers as well as the content they are sharing.
Method: Child pornography related content was identified and analyzed in 235,513 user queries and 194,444 query hits.
Results: The research confirmed a large amount of peer-to-peer traffic is dedicated to child pornography, but supply and demand must be separated for a better understanding. The most prevalent query and the top two most prevalent filenames returned as query hits were child pornography related. However, it would be inaccurate to state child pornography dominates peer-to-peer as 1% of all queries were related to child pornography and 1.45% of all query hits (unique filenames) were related to child pornography, consistent with a smaller study (Hughes et al., 2008).

In addition to the above, research indicates that the median age searched for was 13 years old, and the majority of queries were gender-neutral, but of those with gender-related terms, 79% were female-oriented. Distribution-wise, the vast majority of content-specific searches are for movies at 99%, though images are still the most prevalent in availability.
Conclusions: There is no shortage of child pornography supply and demand on peer-to-peer networks and by analyzing how consumers seek and distributors advertise content we can better understand their motivations.
Practice implications: Understanding the behavior of child pornographers and how they search for content when contrasted with those sharing content provides a basis for finding and combating that behavior. For law enforcement, knowing the specific terms used allows more timely and accurate forensics and better identification of those seeking and distributing child pornography. For Internet researchers, better filtering and monitoring is possible. For mental health professionals, understanding the preferences and behaviors of those searching supports more effective treatment.

Published by Elsevier Ltd.

Introduction

Child pornography on the Internet is an ongoing problem effecting society and represents an important link in the chain of child victimization. In most countries, distribution and even possession of child pornography is categorized as criminal behavior. In the USA, it is a federal offense at the felony level to produce, possess and/or distribute child pornography (Child Protection and Obscenity Enforcement Act, 1988).

* Corresponding author address: 5417 Graywing Court, Columbia, MD 21045, USA.

0145-2134/$ – see front matter. Published by Elsevier Ltd.
doi:10.1016/j.chiabu.2008.12.011

With the advent of the Internet, and specifically peer-to-peer networking, the distribution of child pornography has become easier. No longer the purview of mail-based providers and the back rooms of adult bookstores, child pornography can now be semi-anonymously shared with an Internet connection and one of any number of peer-to-peer clients.

In this paper we look at the prevalence of child pornography on the Gnutella peer-to-peer network. The Gnutella network is the ideal peer-to-peer network to analyze for multiple reasons. First, the clients perform no default filtering of queries as occurs in some other peer-to-peer networks, providing an unadulterated look at the actual terms used by requestors. Second, the Gnutella network allows a system to self-identify as a Ultrapeer, allowing others to connect directly to that system and route queries through it. Third, the Gnutella network is widely used by many popular clients including as Phex, Limewire, and BearShare. Finally, the Gnutella network represents over 40% of all peer-to-peer file sharing installations, making it the most popular peer-to-peer client (Resnikoff, 2007). As a note, peer-to-peer prevalence measurement is a highly contentious issue (Karagiannis, Broido, Brownlee, & Faloutsos, 2004) and volatile based on external influences (Jacob, 2007).

For the purposes of this research, child pornography is defined as the depiction of real, naked children under the age of 18. While this represents a combination of the legal definitions of child pornography and child erotica, separating the two is not possible without viewing the contents.

This paper represents the first broad study of the *demand* for child pornography as well as the supply. Second, not only the prevalence of child pornography is determined but the meta-data associated with the content is analyzed. Finally, the quantifications found provide a baseline for future research into the detection and prevention of peer-to-peer child pornography.

A side result was the generation of a set of words associated exclusively and conclusively with child pornography. This is of tremendous benefit to those performing forensic keyword searching (as well as monitoring and filtering), as potentially ambiguous words commonly used such as "teen" may have a high false positive rate due to news stories, advertisements, and so forth.

Method

Prior art

The most widely cited study of peer-to-peer child pornography was conducted by the US Government Accountability Office in 2003 of the KaZaA network. The study identified 42–44% of the content on the network as being child pornography (GAO, 2003c). It has been cited in other GAO reports, and received significant media attention (GAO, 2003a, 2003b). The study only focuses on the supply, not the demand, however, and makes no attempts to analyze the content itself. Additionally, there are gray areas within the methodology—specifically, the use of "at least one word with a sexual connotation and an age-related keyword indicating that the subject is a minor" (GAO, 2003c). Some of this is necessary, given the nature of the research, but it makes difficult to determine if there is a bias present in the methodology, and whether all of the results using the 12 keywords would also have been confirmed as child pornography had they been viewed.

A previous study found 15% of the content on Usenet was related to child pornography (Mehta, 2001). The Mehta study was comprehensive at the time, but like the GAO study it did not focus on the demand, only the supply (Mehta, 2001). Mehta also points out flaws as well as similarities with an earlier study by Rimm (1995), which has been widely questioned (Hoffman & Novak, 1999; Sigel & Sauer, 1999).

Specific to the ages of children portrayed, a Department of Justice grant study found a median age of 13, in line with the findings of this paper (Klain, Davies, & Hicks, 2001). A second study, focusing on those arrested, categorized the content as related to age, gender, and level of violence present. The predominant age ranges (6–17) were consistent with the findings of this paper, the predominance of females depicted, and the ties to violence were similar, though categorized differently (Wolak, Finkelhor, & Mitchell, 2005).

The closest study to that performed here was done by Hughes et al. and looked specifically at Gnutella and the illicit content present, the same as this study. First, that study comprised a much smaller subset of queries ($n = 10{,}000$) and second, that study used human classifiers. The primary issue with human classification in this realm is that the subculture noted by Hughes et al. use terms of art specific to the subculture, many of which would appear innocuous to non-domain expert (Hughes, Walkerdine, Coulson, & Gibson, 2006). A follow-on study by Hughes et al. used an automated association analysis on the same dataset, and their prevalence results are consistent with this study. They do not, however, examine the contents of child pornography related queries or classify them beyond prevalence (Hughes et al., 2008).

Notes on research

Because the possession of child pornography is illegal in the USA, no actual images and/or movies were downloaded. While this may mean certain query hits labeled as child pornography do not contain illicit material, there was no way to validate this without violating the law. It does mean the query hit results form an upper bound, and are indicative of what those sharing the content have advertised. This does not affect the research into the user-generated queries, however, as that shows a more direct link to intent.

A second note on this research is warranted regarding the inclusion of keywords related to child pornography. There are a series of keywords that are "terms of art" for those involved in trading child porn. With one commonly known exception,

"PTHC," the remaining words have been sanitized and/or not explicitly included in this paper. This was a difficult decision, but the potential downside of providing a "cheat sheet" of keywords to child pornographers outweighs the research benefits of providing the words. The specific words used will be shared with those in law enforcement upon direct request to the author.

Experimental setup

The dataset used represents the results of monitoring user queries over the course of several weeks on the Gnutella peer-to-peer network. Gnutella refers to the protocol used by various software clients to communicate with each other over the Internet for the purposes of sharing files, and is used in this paper to refer to both the protocol and the network of machines connected to each other using that protocol.

On the Gnutella network, an individual user's client software uses a seed list of known machines to make connections to several other clients upon starting. Each of these clients shares a set of files, which are made available for searching and downloading to other clients on the network by way of hash tables. There are two types of clients on the network—Ultrapeers and leaf nodes. Leaf nodes connect to a small number of Ultrapeers, and route their queries through those Ultrapeers. Ultrapeers may act as leaf nodes and originate queries, but are also responsible for routing queries to other Ultrapeers and to connected leaf nodes as appropriate. Clients running on lower bandwidth connections, such as those used by dial-up users, generally operate as leaf nodes and higher bandwidth clients, such as those running on cable modems, become Ultrapeers.

When a user wants to find particular content on the Gnutella, they issue a query to the Ultrapeers with which their client is connected. This query is then propagated (with a limited number of hops) to other Ultrapeers, which search their hash tables for filenames containing the keywords used in the search. The search results are returned to the requesting client, and that client then chooses which files to download. Clients download files directly from one or more machines sharing those files and do not need to go through an Ultrapeer, hence the peer-to-peer nature of the network (Matei, Iamnitchi, & Foster, 2002).

The data was collected by using a customized version of the Phex 3.0.2 client, an open source application with an integrated query monitoring capability. To obtain the data, the Phex client was set to act as an Ultrapeer.

The Gnutella network offers a modicum of anonymity by hiding the original IP addresses associated with queries. As a query is passed through an Ultrapeer, the originating IP address is hidden and the Ultrapeer's IP address is tagged to the query as the originator. Because of this, for more accurate geolocation of the queries (and to avoid double-counting) only those that were associated with a leaf directly connected to the Ultrapeer used for this research were captured. To prevent a localized skew of the data, the connections made to other Ultrapeers were globally diverse.

All of the individual queries were logged to text files and then imported into a Microsoft SQL Server database for pre-processing and mining of the data. The pre-processing contained steps to clean, tokenize, and categorize the queries.

Cleaning. There was an initial cleansing of the data to remove several query results that were not relevant to this research. The steps involved removing Uniform Resource Name (URN) queries, removing presence and empty queries, parsing punctuation, and removing stopwords.

First, all URN queries with SHA1 hashes were removed. Gnutella uses a SHA1 hash value to uniquely identify each file on the network, and to perform searches for the same file with different names. While an analysis of known-bad hashes against the SHA1 values would be interesting research, it was not within the scope of this effort.

Second, several clients use a presence query consisting of a nonsense phrase. These queries are not user typed, and were eliminated from the dataset. Similarly, any empty queries (those containing no alphanumeric characters) were removed from the dataset.

Third, punctuation was parsed out and/or substituted as a space for tokenization where appropriate. A simple removal of all punctuation could not be performed because of the relevance to some of the terms used by child pornographers—one common term uses an "@" in place of the character "a". Similarly, many files available advertising child pornography place the relevant keywords inside of parentheses and/or brackets. Because of this, some punctuation was left in, some was removed, and the rest was replaced with spaces. All remaining letters were converted to uppercase to make matching easier.

Fourth, stopwords were removed from the dataset. Because many users attempt to utilize natural language queries (which are not supported by their clients), many terms with little informational value needed to be removed. The list of stopwords contained 319 total entries and included the most commonly used English language words (List of Common Stopwords).

One data cleansing activity that was not done initially was the removal of file extensions. Though extensions have been removed in other research and have been excluded from some of the result sets below as noted, they are essential for differentiating what content types—movie or image—child pornographers were looking for.

Tokenization. After cleaning the data, it was tokenized to generate a list of words corresponding to a given query and/or query hit. The word lists were loaded into a separate table, and used for the categorization of the content. If a query or query hit contained multiple instances of the same keyword, only one was recorded for analysis so as not to double count entries.

Classification. Based on their keyword content, results were classified as either being child pornography related or not child pornography related. Those related to child pornography were defined as those containing keywords which were associated with child pornography and had no other use. More specifically, a group of seeds terms known to be exclusively associated with child pornography were used as a baseline, then query expansion was performed using a similarity thesaurus generated from the words strongly associated with the initial words from the dataset. The resulting words were manually confirmed as being related to child pornography based on past casework. The categorization of content based solely on ambiguous terms such as "teen" or "young" was avoided. This was done to sacrifice recall in relation to precision in the resultant data and avoid conflating the results. Finally, word combinations associated exclusively with child pornography were identified by applying the query expansion to the query list, to form a resultant grouping of words and phrases uniquely associated with child pornography.

Analysis. Following the categorization, analyses were run against the database to identify the association of the results with various topics of interest. The specific analyses run and any special processing are discussed in the results section below.

Dataset

A large number of (N = 429,957) usable queries were collected by the system. The final words (N = 2,566,660) were identified after cleansing and tokenizing these queries. These represent both queries and query hits. After breaking the results up, the words (N = 649,247) from user queries (N = 235,513) were identified, with the remaining words associated with query hits. The resulting unique words (N = 36,690) used in the queries had an approximate average query size of 4.4 words per query *before removing extensions*. These results show slightly higher words per query than other research, but this is likely due to the handling of punctuation in queries and the resultant splitting of terms (Makosiej, Sakaryan, & Unger, 2004).

Queries were analyzed as they show what the actual end user typed to find content. The query hits used were more representative of the supply side of content than the results of contrived searches as they represent actual file results returned to querying users as opposed to "dormant" files never matched.

Results

Overall

Looking at the overall queries, the most shocking result comes from query prevalence list. Looking at query prevalence, the most common query was PTHC, shorthand for preteen hardcore. Approximately .2% of all queries contained this term, which is clearly and unequivocally associated with child pornography. The queries did not originate from the same location and represent a breadth of countries, indicating a transnational prevalence of term usage.

The other most frequent phrases in the top ten queries dealt with popular movies at the time of the acquisition. These phrases are likely attempts to download copyrighted material, but are not relevant to this research.

In total, just under 1% of all queries were child pornography related. Of the remaining child pornography queries (aside from PTHC), the top 10 were all single words. This shows users searching for child pornography look for broad categories of content, as opposed to the specificity used by those seeking popular movies. The distribution of child pornography related terms was Zipfian, corresponding with the distribution of general terms.

Going contrary to popular belief, the word "teen" only appeared in a relatively small number (N = 535) of queries, fewer in total than the single term "PTHC" (though it was more popular in file names). This indicates a level of sophistication in those searching for child pornography on peer-to-peer networks in using terminology specific to the subculture. A more detailed discussion of the use of the word "teen" in child pornography is included below.

Looking at the prevalence of child pornography in query result hits, the top two file results were child pornography related and included numerous terms associated with each file name. The remaining files from the top 10 were all songs from popular bands. Songs were more prevalent in query hits as they tended to be searched by more broad terms (artist name, song name, keywords) than movies, which used specific title searches and had fewer returns.

Of the total number of query hits, 16% of the total hits were child pornography related, representing 2,770 uniquely named files. This shows a large number of actual files available, significantly higher than the number of queries, but must be tempered by the fact that each file was present an average of just over 11 times, significantly more redundancy than with non-child pornography files. If only unique files are counted, the query hit volume that is child pornography related is 1.45%. Within those filenames, the top 10 words not explicitly associated with child pornography (but potentially associated in phrases) are shown in Table 1. These indicate a need for more advanced phrase-based detection of child pornography instead of simple word matching. A more detailed look at the associations is found in a later section below. There were 463 hosts sharing child pornography, or approximately 7% of all hosts. Five hundred and sixty four, or 3%, of all searching hosts looked for child pornography.

Looking at query and filename composition, the average file name contained 8.6 words (including the extension) and the average query contained 4.4 words. The average query length for child pornography related queries was 3.2, whereas the average file length was 14.5 words. Looking at the query composition, a large percentage of the queries for child pornography were single word queries, almost 50%, whereas the length distribution for all queries peaked at 4 words. For filenames, the

Table 1
Non-explicit words associated with child pornography files.

Word	Count
Child	9,382
Teen	8,754
Porn	7,846
Sexo	5,999
Kids	5,975
Kid	5,379
Anal	5,201
Sex	4,318
Anos	4,226
New	3,621

average filename size for child pornography files showed multiple peaks, with more terms used on average than the number used for all files.

These results are consistent with the queries used by each group of individuals. Child pornographers have a tendency to search for broader categories, which are covered by a smaller number of words, compared to users searching for songs, movies, or other items. This indicates a preference by the child pornographer for a higher recall versus precision, while the opposite holds for most other users. Because of this, those distributing child pornography have opted for including a higher number of terms than average, indicating they are attempting to maximize the number of users who find their files, whereas other files appear to have fewer but more relevant keywords, indicating a focus on precision.

In terms of the file types available, movies were the predominant content type explicitly searched for at 99%, however it may be that those queries that did not specify a format (the majority) were searching for images. Conversely, images made up the largest proportion of files available, at 82%, followed by movies at 17%. This is likely due to the presence of series of images, where a single host has numerous images of the same class. Additionally, there are fewer available movies and they require more storage and bandwidth to make available. Ninety nine percent of the images made available were JPEGs, whereas MPEG was the most popular movie type at 72%, followed by AVI at 26% and WMV at 2%.

Age

One prevalent theme that is found in much of the current literature is that the average age of children represented in child pornography is getting younger (CBS, 2007), often with little cited as backup. In one excellent overall study, as many as 83% of individuals arrested were found to have prepubescent child pornography pictures, defined as those between the ages of 6 and 12. The average age cannot be deduced from this study, however, as it was by its nature skewed toward younger children (individuals are more likely to be arrested with prepubescent images) and it did not count the number of images in each bracket, only if that bracket exists (Wolak et al., 2005).

By evaluating the queries generated, we determined the average age searched for by those looking for underage content. The ages were obtained by mining the query results for age-related terms that co-occurred with image and/or movie requests. The results are shown in Fig. 1.

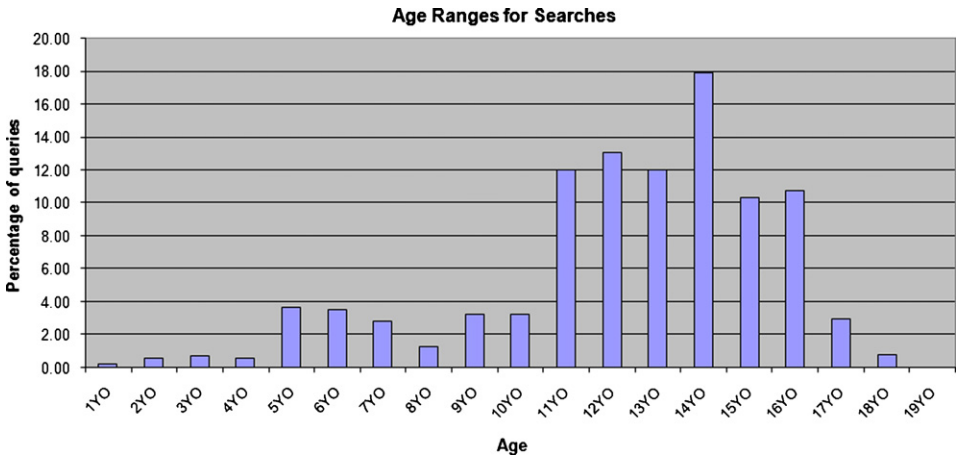

Fig. 1. Age ranges present in query strings. The ages of individuals sought in movies and/or images as used in query strings. Fourteen is the most common age sought, while 13 is the median.

As shown in the figure, the predominant age searched for was 14 years old, with a median age of 13 years of age (mean = 12.3). While this does not always correspond to the ages retrieved, it does represent the perceived ages sought by offenders. This result is consistent with earlier work and does not show a short-term downward progression (Klain et al., 2001). The predominant ages sought are between 11 and 16, representing 76% of age-specific searching. While only age-terms between 1 and 19 were compared, terms related to older ages were extremely infrequent.

Because there is a direct correlation between ages searched for and file names returned and because many files contained multiple ages, this was not segregated out as a separate result set.

Geographic location

The geographic location of those searching for child pornography was determined by geolocating the source IP address. The source IP was able to be accurately determined because the query set was limited to directly connected leaf nodes. Both the sources of the queries (demand) and the query hit locations (supply) were broken out explicitly.

For demand, the predominant location for the origination of requests was the USA at 29%, followed by Malaysia at 16% and Brazil at 12%. The overall table of results is shown in Fig. 2.

For a more normalized look at traffic, the top countries per percentage of traffic dedicated to child pornography related searching were identified. The results show a large amount of traffic for Malaysia and Thailand that is relevant—52% and 23%, respectively. The drive from Thailand may be due to monitoring (Initiative, 2008) and filtering on traditional web usage. Malaysia does not currently implement filtering on a broad scale, though strong state controls may have driven the networks onto a peer-to-peer platform. The perceived and actual enforcement of laws in these countries may also play a factor. In contrast, the US traffic shows approximately .5% of query traffic is directly related to child pornography.

In terms of overall query hits returned with child pornography related terms, the numbers are substantially different. Approximately 90% of all query hit returns originated in Brazil, with only 6% in the US and 1% in the UK. This shows the supply side of the relationship and is heavily focused on Brazilian servers. A deeper analysis of the Brazilian hosts shows 68 independent IP's serving up this content, and diversity in the content provided. This may be due to limited enforcement in Brazil of possession offenses and their legal definition of distribution.

Common assertions

Several common assertions associated with child pornography were tested. First, the assertion that certain keywords such as "Teen" and "Young" do not imply a search for child pornography was tested. Second, the assertion that there is an association between those who seek child pornography and violence was tested. Finally, the assertion that child pornography is predominantly of females was evaluated against the dataset.

Innocuous queries?. One common assertion in child pornography cases by defense counsel is that an individual's searching for "teens," "underage," or "young girls" does not indicate malicious intent. Specifically, an individual may be searching for information on teen pregnancy, underage drinking, or music popular with young girls, none of which would be directly

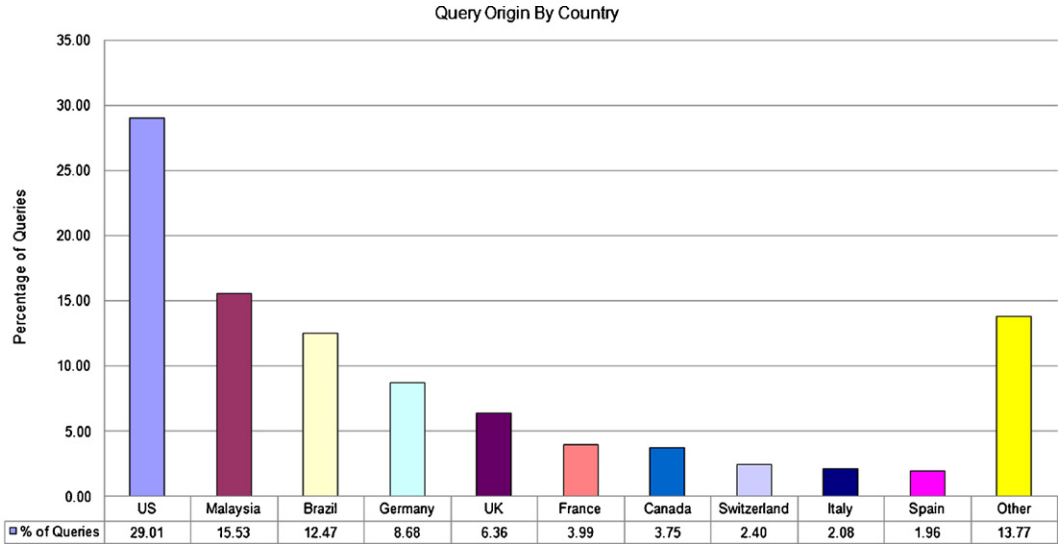

Fig. 2. Originating countries of child pornography related queries. The USA generated the most queries, with almost 1 out of 3 queries related to child pornography originating there.

Table 2
Top 10 co-occurring words.

Underage		Young		Teen	
Word	Similarity	Word	Similarity	Word	Similarity
Pedo	.17	Love	.54	Hentai	.3
Little	.13	Fight	.08	XXX	.28
Girls	.13	Jeezy	.05	Rape	.23
Girl	.11	Fuck	.05	Fuck	.22
Preteen	.11	Buck	.04	Sex	.21
Preteens	.09	Night	.04	Nude	.18
PTHC	.09	Underneath	.04	Pre	.17
Sexy	.09	Nameless	.04	Sexo	.15
Sister	.09	Close	.04	Pedo	.15
Rape	.09	Eyes	.04	Foda	.15

related to child pornography. To test this assertion, a similarity thesaurus was generated between the potentially ambiguous words and other words in the same queries. For these results, a basic similarity test was calculated as follows:

$$S_{ij} = \frac{|K_i \cap K_j|}{|K_j|}$$

Using the three popular words above, the top co-occurrence frequencies were calculated. The top 10 co-occurring words with "Teen" are seen in Table 2. As shown, the top co-occurring word was Hentai, a form of adult Japanese anime. The remaining words are all directly related to sexual activity ("foda" is Portuguese and roughly translates to "fuck" in English). In addition to these results, 15% of the queries have just the word "Teen." Based on these results, the presence of the word "teen" in a peer-to-peer query is a likely indicator the user that generated the query is looking for sexually explicit content, but it cannot be used to determine if the searchers were looking for underage teens or those of legal age.

For the word "Young," the results are a bit more ambiguous as shown in Table 2. It is initially evident that the ambiguous word "Love" co-occurs with "Young" a large number of times. The second and third most common, "Fight" and "Jeezy" are significantly lower. Looking at "Young" and "Fight," all of the results were from a repeated query of "young young love fight too to." The meaning of this query is not readily understandable. "Jeezy" refers to the name of a musical artist. The remaining results appear to be similarly mixed. While associated with some words that are indicative of illicit content, overall "Young" appears to be a more ambiguous indicator than "Teen" of suspicious activity.

The final word, "Underage," appears to be as good as "Teen" as an indicator of suspicious activity. As shown in Table 2, the most common co-occurring word is "Pedo," which is self-explanatory. Queries containing the second, "Little," with "Young" were all for underage content based on the full query wording. Similarly, the remaining terms are highly indicative of illicit material, making "Underage" another good indicator of a suspicious term in peer-to-peer queries. Unlike "Teen," the term "Underage" implies knowledge of illegality and an explicit search for content that is known to be contraband.

Based on the results, "Teen" and "Underage" would both be good indicators of suspicious activity, whereas "Young" contains too much ambiguity. For filtering and monitoring software, this knowledge is likewise important. Filtering may be appropriate (for example, on a corporate network) for the word "Underage," whereas monitoring would be appropriate for "Teen" and neither for "Young," depending on the acceptable false positive rates.

Child pornography and violence. Another common assertion is that the need to view child pornography has a high coincidence with traditional violent pornography (Botting, 2005). To determine the prevalence of searching for violent child pornography, a list of violent words (such as rape, bound, and forced) was looked at in conjunction with the words used to search for child pornography.

Despite common assertions, there appeared to be little overlap between searches for child pornography and searches for violent pornography. Of the searches for violent pornography were conducted ($N = 527$), only 14 of those, or 2.7%, also contained terms associated with child pornography. Additionally, only .7% of the searches for child pornography contained terms associated with traditional violent pornography.

The availability, however, indicates the opposite and shows a strong overlap between child pornography and traditional violent pornography. One hundred and twenty four, or 40%, of the files associated with traditional violent pornography contained child pornography language. Likewise, 36% of the files purporting to contain child pornography also contained language associated with traditional violent adult pornography.

By definition, all child pornography is forced in that it represents sex between at least one party who cannot legally consent to it. The search analysis provides an indication, however, that those searching for child pornography do not necessarily associate it with violent action. On the other side, it appears that individuals who name the files that are available for sharing strongly associate it with violent imagery. This may reflect the common assertion is valid on the part of the producers and distributors, and it appears to indicate a difference in mindset between them and the consumers, the psychological analysis of which is beyond the scope of this paper.

Table 3
Associations with child pornography words.

Word	Confidence	Support
Zoofila	1.00	.02
Adolescent	.97	.01
Boquete	.94	.02
Prima	.94	.02
JHO	.92	.01
Estrupo	.92	.02
Nuas	.92	.02
Foda	.91	.02
Mulheres	.88	.02
Carla	.84	.02
Matrix	.83	.02
Babes	.78	.02

Gender and child pornography. Child pornography and other forms of child abuse can affect both male and female victims, but little has been done to quantify any gender bias in the prevalence and demand for child pornography on a gender-basis. The demand is quantified below by looking for gender-specific words in the queries. Similarly, the supply is quantified by looking for those same gender-specific words in the available files. Gender terms specifically referencing the abuser (father, mother, uncle, aunt, etc.) as opposed to the abused were not included.

On the demand side, the majority of queries contained no gender-specific language. This could indicate individuals looking for either gender, or an assumption that the results will match the envisioned but not explicitly stated gender. Of those queries with a gender bias, 79% of the requests were for pornography containing underage females.

For the files available, there is a more clear gender bias. Ninety five percent of the available files advertise underage females, indicating a higher availability of child pornography showing abused girls. Additionally, only 21 of the files were not tagged as to gender, indicating those naming the shared files were explicit in the gender-specific contents.

Both the supply and demand for content containing young females is higher based on the results. The results, though, cannot be directly correlated with the gender of the requestor, only that of the abused. Additionally, the large amount of gender-tagging in the file names indicates the producers are more concise than the requestors.

Additional associations. In addition to those words which are directly related to child pornography, a series of additional terms were identified using a simple association analysis to determine confidence and support. First, the list of words associated with child pornography queries but not exclusive to child pornography was generated. Then, the support and confidence of these words was calculated with *any* child pornography query. Words with at least a 10% support were ranked, and the top confidence words were generated.

Of the top confidence words (shown in Table 3), there are a few interesting findings. First, several seemingly innocuous words are strongly associated with child pornography. Taking a closer look, adolescent (sp) could be part of innocuous searching, but in fact it has an extremely high likelihood of being associated with child pornography. Prima (cousin), boquete (blowjob), estrupo (rape), nuas (nudes), foda (fuck), and mulheres (women) show the influence of Brazilian traffic on the search terms. Most disturbing is "Zoofila," showing a strong link between bestiality and child pornography. JHO shows an influence from the texting world, where it's short for Just Hanging Out.

Another notable association feature—individuals searching for child pornography tended to limit their searches to child and adult pornography, with little crossover to other popular searches (e.g., for movies and/or music). Less than 5% of the queries associated with IP addresses searching for child pornography had non-pornographic search terms in other queries.

Discussion

Future work

The initial investigation into peer-to-peer child pornography shows distinct behavior patterns associated with child pornographers. One direction of future work is confirming the same prevalence and techniques on other peer-to-peer networks such as e-Donkey and BitTorrent. The centralized network architecture e-Donkey uses relies on large, semi-private servers which are not available to researchers. BitTorrent lacks any central service, and relies heavily on websites to maintain lists of clients sharing a particular file and for searching amongst files as well. Additionally, filtering such as certain Kazaa clients implement and geographic preference for certain clients (due to language localization and other features) may show differing usage patterns which may be of interest in comparing communities between the different peer-to-peer networks.

A second area of investigation would require legal approval—confirmation of the files advertised as child pornography and a characterization of that content. This is not likely to be performed in the USA, but could be approximated through a hash analysis.

Third, the behavior of those seeking child pornography as a function of time would be of interest. To collect this, a study would need to look at a temporal analysis of the hard drives recovered from those arrested for child pornography possession.

Finally, automated detection mechanisms to identify child pornography, both from a supply and demand side, could be tested and implemented on peer-to-peer networks to identify and track child pornographers in real-time.

Conclusion

This study shows a significant community of individuals using peer-to-peer networks to traffic in child pornography. A series of differentiating behaviors between normal queries and those of child pornographers was identified, for use in later classification, and the preferences of online child pornographers were noted.

As an additional outcome, lists of keywords representing high likelihood child pornography activity that can be used on intrusion detection systems, for filtering software, and in forensic analysis were created. These lists will be made available to law enforcement on an as-requested basis.

The study itself was limited to a popular peer-to-peer network. Because of the nature of queries on peer-to-peer networks, generalizations about the presence of child pornography on the web or even other peer-to-peer networks with different topologies (e.g., BitTorrent) are not substantiated at this point. The presence of child pornography on the Gnutella network, however, is both extensive and behaviorally telling.

References

Botting, A. (2005). *Actor Chris Langham tells of 'compassionate' viewing of child porn*. Retrieved December 2, 2007, from http://www.timesonline.co.uk/tol/news/uk/crime/article2906052.ece?token=null&offset=12.

CBS. (2007). *Operation predator busts child porn violators*. Retrieved November 17, 2007, from http://www.action3news.com/Global/story.asp?S=6497873.

GAO. (2003a). *Combating child pornography: Federal agencies coordinate law enforcement efforts, but an opportunity exists for further enhancement*. Washington, DC: Government Accountability Office.

GAO. (2003b). *File-sharing programs: Child pornography is readily accessible over peer-to-peer networks*. Washington, DC: Government Accountability Office.

GAO. (2003c). *File-sharing programs: Peer-to-peer networks provide ready access to child pornography*. Washington, DC: Government Accountability Office.

Hoffman, D., & Novak, T. (1999). *A detailed critique of the TIME article: On a screen near you: Cyberporn* (*DeWitt, 7/3/95*). Retrieved December 2, 2007, from http://sloan.ucr.edu/cyberporn/time.dewitt.htm.

Hughes, D., Walkerdine, J., Coulson, G., & Gibson, S. (2006). Peer-to-peer: Is deviant behavior the norm on P2P file-sharing networks? *Distributed Systems Online*, 7.

Hughes, D., Rayson, P., Walkerdine, J., Lee, K., Greenwood, P., Rashid, A., May-Chahal, C., & Brennan, M. (2008). Supporting law enforcement in digital communities through natural language analysis. In *Proceedings of IWCF'08* Washington, DC, USA, August, 2008,

Initiative, O. N. (2008). *Thailand: Open network initiative*. Retrieved August 8, 2007, from http://opennet.net/research/profiles/thailand.

Jacob, A. (21, 2007). *Recording industry knocks out eDonkey servers in new actions against Internet piracy: Legal steps in France, Germany and the Netherlands cut off more than one million users of one of the largest P2P networks*. International Federation of the Phonographic Industry Press Release.

Karagiannis, T., Broido, A., Brownlee, N., & Faloutsos, M. (2004). Is P2P dying or just hiding? In *Proceedings of Globecom 2004* Dallas, TX,

Klain, E., Davies, H., & Hicks, M. (2001). *Child pornography: The criminal–justice-system response*. Alexandria, VA: National Center for Missing and Exploited Children.

List of Common Stopwords. (n.d.). Retrieved August 9, 2007, from http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/stop_words.

Makosiej, P., Sakaryan, G., & Unger, H. (2004). Measurement study of shared content and user request structure in peer-to-peer {Gnutella} network. *Design, analysis, and simulation of distributed systems* (pp. 115–124). Arlington, VA.

Matei, R., Iamnitchi, A., & Foster, P. (2002). Mapping the Gnutella network. *Internet Computing*, *6*, 50–57.

Mehta, M. (2001). Pornography usage in usenet: A study of 9.800 randomly selected images. *Cyberpsychology and Behavior*, *4*, 695–703.

Resnikoff, P. (2007). *Digital media desktop report, Q4 2007*. Digital Music News.

Rimm, M. (1995). Marketing pornography on the information superhighway: A survey of 917,410 images, description, short stories and animations downloaded 8.5 million times by consumers in over 2000 cities in forty countries, provinces and territories. *Georgetown Law Journal*, *83*, 1849–1915.

Sigel, L., & Sauer, G. (1999). *Critique of rimm article on online pornography*. Retrieved November 9, 2007, from http://sloan.ucr.edu/cyberporn/sigel.sauer.critique.htm.

Wolak, J., Finkelhor, D., & Mitchell, K. (2005). *Child pornography possessors: Arrested in Internet-related crimes: Findings from the National Juvenile Online Victimization Study*. Alexandria, VA: National Center for Missing and Exploited Children.