Comparing Methods for Detecting Child Exploitation Content Online

Bryce Westlake, Martin Bouchard, Richard Frank
School of Criminology
And
International Cybercrime Research Centre
Simon Fraser University
Burnaby, Canada
{ bwestlak, mbouchard, rfrank }@sfu.ca

Abstract— **The sexual exploitation of children online is seen as a global issue and has been addressed by both governments and private organizations. Efforts thus far have focused primarily on the use of image hash value databases to find content. However, recently researchers have begun to use keywords as a way to detect child exploitation content. Within the current study we explore both of these methodologies. Using a custom designed web-crawler, we create three networks using the hash value method, keywords method, and a hybrid method combining the first two. Results first show that the three million images found in our hash value database were not common enough on public websites for the hash value method to produce meaningful result. Second, the small sample of websites that were found to contain those images had little to no videos posted, suggesting a need for different criteria for finding each type of material. Third, websites with code words commonly known to be used by child pornographers to identify or discuss exploitative content, were found to be much larger than others, with extensive visual and textual content. Finally, boy-centered keywords were more commonly found on child exploitation websites than girl-centered keywords; though not at a statistically significant level. Applications for law enforcement and areas for future research are discussed.**

Keywords- Cybercrime; online child exploitation; image hash database; web-crawler

I. INTRODUCTION

The growth of the Internet has been met with a paralleled growth in cybercrime. Although some forms of cybercrime, such as fraud, result in a greater impact on the global financial market, the crime of child exploitation is, perhaps, one of the most impactful forms of cybercrime that currently exist. This point is supported by the United Nations, who in 2009 estimated that there were over four million websites containing child exploitation content [1]. This, of course, precludes Internet chat relay, newsgroups and peer-2-peer networks, such as eDonkey, where the distribution of child exploitation content is also rampant [2].

The fight against online child exploitation is a strongly advocated issue with plenty of global support. Government agencies such as The United Nation's International Criminal Police Organization (INTERPOL) have spearheaded the fight with the creation of the International Child Sexual Exploitation image database. This database contains all known sexually explicit photos of children and is used by various police organizations throughout the world.

Efforts to combat online child exploitation have also been aided by key private organizations. Two well-known Internet companies that have provided several software programs for use by police organizations are Google© and Microsoft©. Working with the National Center for Missing and Exploited Children (NCMEC), Google© has adapted their pattern recognition program, used on Youtube© to detect copyrighted material, to aid in searching through files uploaded to the Internet, for child pornography [3]. This program has aided the NCMEC "in organizing and indexing…information so that analysts can both deal with new images and videos more efficiently and also reference historical material more effectively" [4] para. 6. Microsoft, while also working with NCMEC, has created a similar program known as PhotoDNA [5]. PhotoDNA has been beneficial because of its ability to analyze a large number of potential images in a short period of time as well as its ability to detect modified versions of known child pornography images.

Research organizations have also been helpful in developing tools for combating online child exploitation. A conglomerate of universities in Europe, headed by Matthieu Latapy has examined peer-2-peer child exploitation and has examined the use of keywords in detecting child exploitation content (see http://antipaedo.lip6.fr/results.htm for a list of publications). Latapy and colleagues have found that peer-2-peer networks are a hotbed for the proliferation of child exploitation videos and images. In fact, they found that more than 0.25% of *all* queries were pedophilic [6]. Similarly, at the International Cybercrime Research Centre at Simon Fraser University, research has been conducted examining online child exploitation, focusing on the structure of [7], the ability to disrupt [8], and the identification of key players within [9] large networks of child exploitation blogs and sites.

In this study, we draw on a custom-made web-crawler, specifically designed to extract publicly accessible child exploitation website networks, to compare the networks created using 1) only image hash values, 2) only keywords,

and 3) a hybrid model that uses both image hash values and keywords. Our attention is focused on finding whether one method is more appropriate than others in finding actual child exploitation content. Indicators such as the number of hash value images found in the networks created, as well as the prevalence of specific keywords used by child pornographers will be used to assess the nature of the websites extracted by the web-crawler.

Current approaches to combating child exploitation online have focused primarily on the use of image hash value databases to identify pictures that meet the criteria of being classified as child pornography. Although this approach has proven to be successful thus far -especially for peer-2-peer networks- there are two key limitations that can be improved upon, with more integrated methods. The first key limitation is that searches using the database allow for the discovery of only images. Given the expansion of broadband Internet connections, which has increased the average user's ability to transfer large files, there is a need to integrate other types of content (i.e., videos) into the child exploitation searches. Second, image hash databases are ineffective at finding new content as they rely on pre-existing content. In other words, a website with only new content would never be found using this technique alone.

As global police organizations have focused on the use of image hash databases, researchers have begun to stretch their attention to using keywords to detect child exploitation content. The use of keywords has the advantage of finding content that may be new, which can aid in finding children who are currently being abused, but also has the limitation of possibly identifying a lot of false positives. That is, identifying websites that actually contain legal, adult pornography, instead of illegal child pornography. This process can be minimized through an effective choice in keywords to search for.

II. METHODS

A. Web-crawler

We used an updated version of a custom-written web-crawler used in previous research (see [7]) called the Child Exploitation Network Extractor (CENE) to map out our child exploitation networks [1]. CENE is a computer program designed to automatically browse websites and collect data about each website based on pre-defined criteria. Within the current study, the web-crawler was asked to collect data on the number of images, videos, keywords, and linkages between child exploitation websites. The CENE starts at a user-specified website, looks for a specific criterion, and then recursively follows the links out from the webpages that meet that criterion. We make this criterion vary in three ways to mirror our research question. The criterion was set at 1) presence of one known child exploitation hash value (i.e., an image found in the hash value database), 2) presence of seven child exploitation keywords (to be specified below), or 3) either seven keywords or one hash value. The web-crawler continues to examine websites until it reaches its pre-set size criteria. Within the current study, that criteria was set to 500 websites and 300,000 webpages – a large enough number of webpages to detect patterns, though not large enough to slow down network extraction or complicate subsequent network analyses. Another criterion used in the crawling process was the exclusion of websites known to not contain child exploitation content. These websites were based off of a list of the most popular websites (e.g., Google©) and a list of websites collected during previous data collections that were verified to not contain child exploitation content [2]. The resulting network contains information about the number of images (overall and known child exploitation), videos, keywords, and linkages, which is aggregated up to the root website level (i.e., www.website.com).

B. Image Hash Database

A hash value refers to the mathematical process whereby a large piece of data is reduced into a single, 24-hexidecimal code [10] which can act as a form of encryption and used to authenticate the contents of an image [11]. The hash values are calculated from any binary file, and are represented as a 128bit value, such as "79054025755fb1326e46c822af954eb4". This serves as a fingerprint to the file it was produced from. A change, even a bit, will result in a completely different hash-value. The potential for two different images having the same hash value is extremely small to non-existent.

The image hash database used for this project was provided by the Royal Canadian Mounted Police's (RCMP) This database was used with the permission and support of the RCMP and is regularly updated with new images as they are found in Canadian-based investigations. The version used for this study was last updated on January 5th, 2011. The images within the database are categorized into one of six categories. The first category (Category 1) was images that are considered to be child pornography as defined by the Canadian Criminal Code of Canada (CCC). This category consisted of 173,291 hash values. The second category (Category 2) was images of children that could be classified as child pornography (Category 1) but are not as clear. In other words, images placed into Category 2 were images that would be classified as 'grey-area' images. This category contained 140,809 hash values. The third category (Category 3) was images that were seized from offenders that would not be considered child pornography under the CCC, but were important enough to the offender that they were collected. An example of this would be an image that portrayed a child as (partially) clothed (e.g., in a costume) and was a pre-cursor in a photo shoot, to an image of child pornography. Category 3 contained 2,691,793 hash values. The fourth category (Category 4) were known adult pornography images and consisted of 4,472,717 hash values. Category 5 was obscene child exploitation images that include content such as bestiality, torture, or de-humanization. In total, there were 70,119 hash values in this category.

[1] A series of websites connected between them via hyperlinks..

[2] During each subsequent web-crawl, the database of known 'good' websites is added to, when a website collected during a web-crawl is verified to be a false positive.

Finally, Category 6 was images known to be irrelevant to sexuality. In total, there were 36,188,568 hash values. Although data was collected on the number of images found that were contained within Category 4 and 6, neither was used as a criterion for inclusion within our networks. In addition to the images found within the database, images that were unclassified were also collected. It is within this category that 'new' images would be categorized.

C. Keywords

Keywords were matched using a root word methodology[3]. The keywords used within this study were broken down into three categories. In total, eighty-two keywords were used. Category 1 was words that were known to be commonly used by child pornographers. These were words that were provided by the RCMP as well as those used in previous research [12] [13] [14] [15]. These words included *qwerty, qqaazz, ptsc,* and *pthc*. This category of keywords consisted of twenty-seven words. Given that these words are synonymous with child exploitation content, we will pay special attention to the websites containing at least one of them, as the most likely to contain illegal content. Category 2 was terms often used in child exploitation searches, which could also be found in non-child exploitation contexts. These terms included *boy, girl, bath,* and *child*. These words help provide context on the content of the website (e.g. boy-centered vs. girl centered). This category consisted of sixteen words. Category 3 was keywords commonly used to find sexual content, but not necessarily child exploitation content. These terms included *sex, anal, lolita,* and *twink*. In total, there were thirty-nine words contained within this category.

All three categories of keywords were included within the present study as previous research by [16] found that non-explicit terms such as *child, teen, kid,* and *sex*, were commonly associated with child exploitation material. Therefore, we wanted to use a selection of all three types of keywords to ensure that we minimized false negatives: websites identified as not having child exploitation content but actually containing the content. Countering this, the inclusion of so many keywords, not directly associated with child exploitation content, increased the possibility of false positives. Preliminary tests of the crawler helped determine that using a seven keyword floor limit was appropriate to maximize the search for child exploitation content, while minimizing false positives. However more research is needed on validation strategies for proper classification of websites, such as the use of hash value database- our current strategy.

In addition to comparing different categories of keywords, we also compared whether a website was boy-focused or girl-focused. Boy-focused keywords included words such as boy, penis, and twink, while girl-focused keywords included words such as girl, vagina, and lolita. In total, there were five boy-centered words and ten girl-centered words[4].

[3] For example, one of our keywords was bath. This meant that variations on this word, such as bathing, were also picked up.

[4] Although there were twice as many girl-centered words, this was the result of variants of the same word: lolli, lola, lolita, lolly and nymphet, nimphet

D. Starting Website

One website (blog) was selected as the starting point for our three searches. We selected a blog as a starting point as previous research [13] showed that blog networks were more reliable and easier to construct than non-blog networks through a web-crawler strategy focusing on publically accessible websites. The website chosen as a starting point for this study was selected based off a list of websites already verified as containing child exploitation content. This list of websites came from the RCMP as well as a database of websites verified to contain child exploitation content from our previous research. The three networks were created simultaneously using each of the three criterions: 1) presence of one hash value, 2) presence of a minimum of seven keywords, or 3) either. Performing the search simultaneously avoided the problem of one website being available in time 1, but unavailable in time 2. As previously noted, each of these networks was allowed to range up to 500 websites and 300,000 webpages. Only one of our networks reached this threshold however.

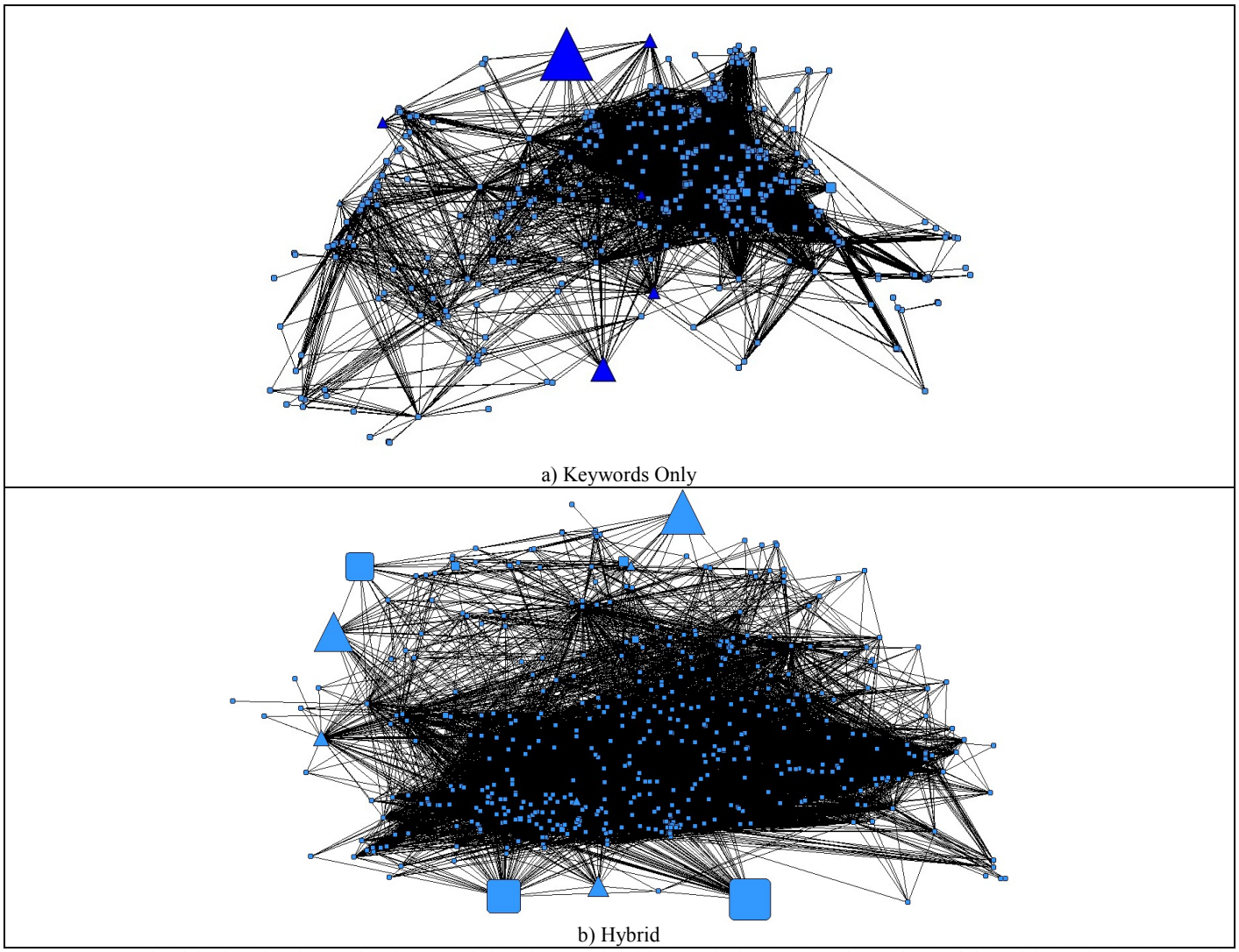
III. Results

A. Network Characteristics

The first result is a negative one: although we were able to obtain full networks using the keywords only and hybrid methods (see Fig. 1), we were unable to create a full network using the hash value only method. While our starting website contained known hash values, it was only able to connect to one other website; at which point the network could not be expanded further. The reasons for this will be discussed in further detail below. Websites that contained verified child exploitation hash values and appeared in both networks are highlighted in Fig.1.

Since we used the same starting websites for the creation of all three networks, it is interesting to look at the level of overlap between them. First, both websites found in the hash value only network were also present in the other two networks. Second, the keywords only network and the hybrid network showed an overlap of 160 websites of a total of 507, or 31.6%. This shows that adding the criterion of hash value images – even if few images were eventually found (look ahead to table 1) – does lead the web-crawler to explore the websites differently than would be expected by chance[5]. As

[5] Given the random component built into the web-crawler, even if the same network criterion is used on multiple crawls, the networks is likely to be slightly different. This is because the web-crawler does not search links in the order they appear on the website. Instead, it selects one at random to follow. This randomness is important as it more accurately mirrors the possible path a user might take when searching websites (e.g., looking for a specific topic). Preliminary tests comparing the overlap between two test networks, starting at the same website, and our findings, suggest that in a random situation, the number of websites that would be present in both networks would be substantially higher than what we obtained within this analysis. Nevertheless, more systematic tests of the behavior of the web-crawler using similar and different sets of criteria are needed before we can have a more definite conclusion regarding this.



a) Keywords Only

b) Hybrid

Figure 1– A graphical display of the keywords only (a) and hybrid (b) networks with the node size reflecting the number of child exploitation hash values found and the triangle representing websites found in both networks that had child exploitation content.

such, this may help find more potential websites containing child pornography, or simply more false positives.

Table 1 provides the descriptive details for each network. The keywords only network consisted of 507 websites totaling 290,455 webpages. Each website consisted of an average of 572.9 webpages, 4,165.3 images, and 6.2 videos. The network that incorporated a hybrid method was very similar, but searched more webpages overall (311,849) and analyzed websites that contained, on average, more videos (22.5) and more total keywords (143,860.4). As noted earlier, the hash value only network consisted of only two websites, totaling twenty-eight webpages.

B. Hash Values

Analogous to our result regarding the difficulty of mapping a network through the hash value method, few confirmed child pornography images (Category 1) were found across the three networks. Only eighty-nine Category 1 hash values were discovered in total. No Category 2 and Category 5 hash values were found, while 322 Category 3 hash values were obtained. It is not surprising that Category 3 hash values would be the most predominant as these hash values are not *technically* against the law. However, it is important to recall that although they are not illegal, they are parts of photo shoots that *do* contain illegal content. As such they may be more likely to lead to content that *is* illegal. Based on the criteria of there being at least one hash value from Category 1, 2, 3, or 5, only 2.96% of the websites in the keywords only network had verified hash values. Within the hybrid network, the number of websites that met the criteria was slightly higher at 3.55%. By definition, the two websites analyzed with the hash value only method had to contain hash values.

Category 1 images are important for our purposes because they consist of confirmed child pornography images. While the two websites analyzed for the hash value only method did not contain any Category 1 images, the keywords only network found four of those images, all located on the same website. The hybrid strategy was the most successful at identifying confirmed child pornography images, with a total of eighty-four across the entire network. However, seventy-four of these were found on one website while the rest were spread out over an additional three websites (including the

TABLE 1: DESCRIPTION OF THREE NETWORKS

Network Type		Hash Value Only	Keywords Only	Hybrid
Number of Websites		2	507	507
Number of Pages		28	290,455	311,849
Per Website	**Pages**	14.0	572.9	615.1
	Images	504.0	4,165.3	4,215.5
	Videos	0.0	6.2	22.5
	Keywords	2,466.5	91,486.9	143,860.4
Hash Value Category (Total # of Images)	**Category 1**	0	4 (n = 1 website)	85 (n = 4 websites)
	Category 2	0	0	0
	Category 3	70 (n = 2 websites)	80 (n = 14 websites)	172 (n = 15 websites)
	Category 5	0	0	0
Percentage C.P. Verified		100.00	2.96	3.55
Category 1 (Code Words) Keywords Per Website		15.0 (n = 2 websites)	31.5 (n = 93 websites)	54.5 (n = 110 websites)
Category 2 (Generic Words) Keywords Per Website		1,678.5	43,513.5	89,473.4
Category 3 (Sexual Words) Keywords Per Website		773.0	47,941.9	54,332.6
Boy-Centered Keywords		1,317.5	43,218.0	54,651.7
Girl-Centered Keywords		14.0	3,732.9	2,620.8

single website identified with the keywords only method as containing those images).

In comparison to Category 1 hash values, Category 3 hash values were more spread out over the various websites within each network. Within the keywords only network, eighty were found, spanning fourteen websites ranging from one to thirty-four hash values. For the hybrid network, there were 172. These hash values were spread out over fifteen websites and ranged from one to thirty-two. Within this network, there was an overlap between websites that had Category 1 and Category 3 hash values. The website that contained seventy-four Category 1 hash values also had twenty-four Category 3 hash values. No other such overlap was found in any website in the network, suggesting some degree of 'specialization' for the limited number of websites containing images classified by law enforcement agencies in Canada.

Although the total number of Category 1 hash values was low overall and tended to cluster in a small number of websites, this should not be taken as evidence that the ability to use hash values to verify child pornography websites is poor nor that the majority of the websites within our networks did not contain child exploitation content. As will be discussed next, the presence of a large number of child exploitation keywords found on each website suggest that many of the websites within the networks simply contained hash values that were not included in our version of the image hash database. However, this stresses the importance of continually updating this database and ensuring that the newest content is available to be searched and verified.

C. *Keywords*

As we outlined previously, there were three categories of keywords used in this study: 1) code words, and 2) generic terms used by child pornographers, as well as 3) sexuality-related words. Within the keyword only network, the average number of Category 2 and 3 keywords on a website were equally common (43,513.5 to 47,941.9). However, there was a slight difference within the hybrid network. Here, Category 2 keywords were more common than Category 3 keywords (89,473.4 to 54,332.6). Both Category 2 and 3 keywords were less common in the hash value only network.

More importantly for our purposes, is the prevalence of Category 1 keywords: those known to be used by child pornographers to identify content. On average, each website within the keywords only network had 31.5 Category 1 keywords, while websites within the hybrid network had 54.5 of those keywords. Note that those scores were above the two websites found using the hash value method (15.0 Category 1 keywords).

Overall, 110 (or 21.7%) of the websites within our hybrid model contained at least one Category 1 keyword. Only one of those was also found to have a confirmed child exploitation image (Category 1). However, nine of our fifteen websites that contained at least one Category 3 hash value had at least one Category 1 keyword. In the keywords only network, 93 (or 18.3%) of our websites contained at least one Category 1 keyword. The one website within this network that contained Category 1 hash values did not contain any Category 1 keywords. In addition, only five out of the fourteen websites that contained Category 1 keywords also contained Category 3 hash values.

D. *COMPARING (POTENTIAL) CHILD EXPLOITATION WEBSITES TO OTHERS*

The design of an automated tool to identify child exploitation websites requires a good understanding of the

characteristics that make them different from other types of websites. In table 2 we compare the network characteristics of websites with known hash values and websites without known hash values. Because we lacked a proper sample of websites with Category 1 images, we used all three networks and merged all individual websites where a hash value image was found (whether Category 1 or 3) together as a single category (n=21). In situations where a website was present in more than one network, the version of the website with the most information was kept and the other versions were removed. Comparing the two groups (see table 2) we found that they differed on the number of webpages per website as well as the total number of keywords per page; however, these differences were not statistically significant.

Are the websites with known hash value images more likely to focus on boy or girl content? Knowing that information may help us learn more about the nature of the images known to law enforcement that were found for the purpose of this study. Results indicate that overall the majority of websites within each network were boy-centered (see table 2). However, when we break these down into websites with hash values and websites without hash values, we find that all websites with at least one hash value was boy-centered, while 90.0% of websites without a hash value were boy-centered. This difference was statistically significant and is worth investigating in future research endeavors.

Also important for our purposes is whether the websites found to contain a code word used by child pornographers differed from others, as they, too, may contain illegal content. The larger number of websites with Category 1 keywords allows us to analyze networks derived from both methods separately. Table 3 compares these two groups across both network types. Within the keyword only network, websites with Category 1 keywords were significantly larger (more webpages per website), however, they did not differ in regards to any of the other characteristics. On the other hand, in the hybrid model, websites differed with regards to how many keywords (overall) they contained and how many images were present per webpage. For both of these differences, websites *without* Category 1 keywords were significantly higher. These findings suggest that child exploitation websites may be larger (more webpages) but have fewer images (specialization in content). In addition Category 1 keywords may not be useful as a marker for distinguishing between different types of websites. Note that these websites are generally much larger than others, which may increase the likelihood of detecting any keywords – including Category 1 keywords.

TABLE 2: COMPARING NETWORK CHARACTERISTICS FOR WEBSITES WITH CATEGORY 1 KEYWORDS AND WITHOUT

Content Per Webpage	Contains Hash Value (n=21)	Does Not Contain Hash Value (n=802)	t-value
# of Pages (Per Website)	30,698.7	1,538.4	1.13
# of All Keywords	1,196.7	278.4	1.38
# of Images	14.2	13.3	0.11
# of Videos	0.00	0.05	-0.28
Category 1 Keywords	0.04	0.08	-0.26
Category 2 Keywords	745.8	125.0	1.55
Category 3 Keywords	450.9	153.4	0.89
% Boy-Centered (Network)	100.0	90.0	9.62*

*=p<0.001

TABLE 3: COMPARING NETWORK CHARACTERISTICS FOR WEBSITES WITH CATEGORY 1 KEYWORDS AND WITHOUT

	Keywords Only		Hybrid	
Content Per Webpage	**Cat.1 Keyword (n=93)**	**No Cat.1 Keyword (n=414)**	**Cat.1 Keyword (n=110)**	**No Cat.1 Keyword (n=397)**
# of Pages	2,679.0**	99.8	1,712.7*	311.0
# of All Keyword	323.64	324.20	304.24	423.66*
# of Images	16.91	20.97	15.96	26.94**
# of Videos	0.04	0.01	0.05	0.05
Cat. 1 Keyword	1.06*	---	0.56*	---
Cat. 2 Keyword	175.35	181.47	152.88	235.55
Cat. 3 Keyword	147.23	142.73	150.79	188.10
% Boy-Centered (Network)	80.6	87.2	93.6	96.0

*=p<0.05
**=p<0.001

Unlike the hash value only network, not all websites containing a Category 1 keyword were boy-centered. In fact, in the keywords only network, only seventy-five out of ninety-three (or 80.6%) of the websites with a Category 1 keyword were boy-centered, while 361 out of 414 (or 87.2%) of the websites without a Category 1 keyword were boy-centered. Similarly, within the hybrid model, 103 out of 110 (or 93.6%) of the websites with a Category 1 keyword were boy-centered, compared to 381 out of 397 (or 96.0%) of the other websites. However, in both situations, the difference was not statistically significant.

IV. DISCUSSION

The objective of this study was to investigate the utility of several methods that are used to find child exploitation content on the Internet. Although this should be looked at as a preliminary study and that more investigation is needed before any firm conclusions can be drawn, two key trends appeared. First, an integrated approach using both hash values and keywords can be more helpful in finding child exploitation websites than either method alone. Second, that websites containing verified child pornography images differ from those that do not contain verified child exploitation content.

The first criterion that we used for our web-crawler was that employed by many law enforcement agencies today: the presence of image hash values. Although this criterion removes the possibility of false positives inherent within the

keyword criterion, as evident by our web-crawler's inability to create a full network, relying on this method alone created a large false negatives issue. Several reasons may explain this result. First, it is possible that the hash value database used for this analysis was outdated and less likely to find matches on the webpages analyzed by the web-crawler. Second, because there is an element of chance involved inherently in the method, it is possible that the starting website chosen for analysis was simply not surrounded by websites containing those images. However, it is equally possible that the website chosen was indeed surrounded by websites of similar content, but that the web link followed by the web-crawler did not lead to a webpage where such content could be found. The reason for this is that the web links posted on a given website will traditionally point towards the home page of another website. Often, the home page does not display any images. Therefore, the web-crawler fails to find a known child exploitation hash value on the webpage, resulting in the website being deemed non-child pornography related and the web-crawler moves on to the next website. The bottom line, for the purpose of this study, is that using hash values alone contain an element of uncertainty that may reduce the likelihood of finding child exploitation content.

The second criterion that we used was the presence of at least seven child exploitation keywords. Although on the surface, this criterion seemed to be better because it could construct an entire network, there were also problems with this methodology; namely false positives. Within the created network, there were websites that were identified that did not contain child exploitation content. An example from this search was the inclusion of the popular image website Flickr©[6]. Again, this does not invalidate this search process. Rather it suggests that more efforts are needed at refining the search process to eliminate false positives and that, alone, this may not be the best methodology.

The third criterion that we used was the presence of seven keywords *or* one known hash value. Our research suggests that this methodology performs at least equally if not better than the keywords only method. The differences found in the two networks and their websites are worth noting. First, the website with the most (seventy-four) Category 1 hash values was found within this network. Using only hash values did not result in it being found nor did the use of just keywords. Instead an integrated approach was successful at finding the website. Second, the keywords only network contained only one website with Category 1 hash value content. In comparison, the hybrid search model found four websites with Category 1 hash values. Third, the website found in the keywords only network was also in the hybrid network, plus several additional websites. Fourth, the hybrid criterion also resulted in the highest number of Category 3 hash values.

[6] Although Flickr© may need to be classified as a 'safe' websites for subsequent web-crawls, it is important to note that two Category 3 images were collected from Flickr©. Given the nature of the website, which gives people the ability to upload any photos they wish, it may be that Flickr© is a website where people commonly post child exploitation content.

On the surface the third, hybrid, criterion seemed to perform the best out of the three. It was a more flexible method, in that it allowed the web-crawler to adapt its search to the nature of the websites involved. For example, some websites have very few, if any, words (e.g., image galleries). In this scenario, the website is best captured by the hash value only criterion. While, for websites with a lot of words (e.g., blog), the keyword criterion is the most suitable. Further refinements to the criteria are needed and will be the focus of our future research, most pressingly in finding whether the use of certain keywords on our list lead to more false positives than others.

The second key finding was that websites verified to contain child exploitation hash values had different characteristics than websites that did not have confirmed hash values. First, they had more Category 2 and 3 keywords per webpage. The presence of high levels of Category 2 and Category 3 keywords support the research by [16] who argues that generic and sexual keywords are equally important in finding child exploitation content online. Second, Category 1 keywords were not present at higher rates on websites with known hash values Apart from the fact that few of those keywords were found per webpage overall, this suggests that there were no more or, no less false positives in the vast majority of websites where no known images were found. In other words, child pornography is just as likely to be found in either sample., Our third key finding was that websites with child exploitation hash values contained a smaller number of videos per webpage suggesting that websites may specialize in their content and that different criteria are needed to find websites that distribute child exploitation images and those that distribution videos.

Finally, as our study focused on publicly accessible websites, the question needs to be raised as to whether our methodology would translate over to peer-2-peer networks and Usenet groups. We believe that it would. One of the reasons for this is that the hash value database has been used by law enforcement agencies to target peer-2-peer networks. Additionally, the keywords used in this study have been used by previous researchers [2], [12], [13], [14], [15], to analyze peer-2-peer networks. Of course, our overall method might change slightly as within other Internet domains we are analyzing networks at the individual level rather than the website level. Therefore, we would be identifying specific individuals who are sharing specific content rather than groups of people (i.e., a website). Also, peer-2-peer networks operate on the basis of file names and do not contain excessive amounts of keywords. Therefore, the search could be different, resulting in the hash value method being more effective. Within Usenet groups, we suspect that our findings would be comparable as they would operate similarly to public websites in so much as there would be some posts that would be text driven while others would be image focused. Regardless, with minimal modifications, it can hypothesized that our methods for finding child exploitation content can be transferable to other Internet domains.

V. CONCLUSIONS

The anonymous nature of the Internet suggests that the problem of online child exploitation will continue to grow as time passes. As a result, it is imperative that we continue to conduct research examining the best practices for finding content and apprehending offenders. Traditionally, efforts to combat child exploitation have relied almost exclusively on image hash databases of known child exploitation content. Although this has proved to be a successful method for finding content, it may not be the most efficient. As time progresses and offenders become more sophisticated in their practices, the ability for hash value databases to be effective decreases. This is because any slight altercation to the image results in a different hash value. Therefore, it is vital that we continue to find new methods that can aid in finding child exploitation content; both old and new. If new methods, such as the use of keywords, are found to be successful, this can have important implications for the overall practice of law enforcement agencies.

Within the current study, we have begun to explore the use of keywords in Internet searches and have found that they may be a positive alternative or compliment to the pre-existing hash value method. The current research should be viewed as a preliminary study that investigates the ability for keywords to be integrated into search criteria. Future research needs to compare the existing hash value database method to the keywords method more extensively, through simulations, to understand the behavior of the web-crawler in a variety of contexts; using different combinations of keywords. It is possible that our current keyword net is too wide and catches too many false positives in the hopes of minimizing false negatives.

In addition to reducing false positives and false negatives subsequent research needs to devise a way to ascertain which images are new content; arguably the most important content to find. For example, the thirty-three websites that were shown to host known hash values contained many more images that potentially contain illegal content. A database of potential child exploitation images could be constructed around those hash values, and a second web-crawl could be launched to analyze whether those images can be found elsewhere in the network. Future research also needs to find a way to better integrate video searches, given that they did not appear on the image websites. This may be the most challenging limitation to address as it may require an entirely new database built around videos.

REFERENCES

[1] Engeler, E. (2009, September 16). UN expert: Child porn on internet increases. *The Associated Press.* Retrieved from: http://abcnews.go.com/Technology/wireStory?id=8591118.

[2] Latapy, M., Magnien, C., & Fournier, R. (2009). Technical report on the Quantification of paedophile activity in a large p2p system. Measurement and Analysis of P2P Activity Against Paedophile Content Project. Retrieved from: http://antipaedo.lip6.fr.

[3] Shiels, M. (2008 April 14). Google tackles child pornography. *BBC News.* Retrieved from: http://news.bbc.co.uk/2/hi/7347476.stm.

[4] Baluja, S. (2008). Building software tools to find child victims. Retrieved from http://googleblog.blogspot.com/2008/04/building-software-tools-to-find-child.html.

[5] Microsoft. (2009 December 15). New technology fights child porn by tracking its "PhotoDNA". Retrieved from: https://www.microsoft.com/presspass/features/2009/dec09/12-15photodna.mspx.

[6] Latapy, M., Magnien, C., Fournier, R. (in press). Quantifying paedophile activity in a large p2p system. *Information Processing &Management.*

[7] Frank, R., Westlake, B.G., & Bouchard, M. (2010). The structure and content of online child exploitation. Proceedings of the 16th ACM SIGKDD Workshop on Intelligence and Security Informatics (ISI-KDD 2010).

[8] Joffres, K., Bouchard, M., Frank, R., & Westlake, B. (2011). Strategies to disrupt online child pornography networks. Paper presented at the European Intelligence and Security Informatics Conference 2011, Athens, Greece.

[9] Westlake, B.G., Bouchard, M., & Frank, R. (2011). Finding the key players in online child exploitation networks. *Policy and Internet, 3(2) Article 6.* doi:10.2202/1944-2866.1126.

[10] Howard, T.E. (2004). Don't cache out your case: Prosecuting child pornography possession laws based on images located in temporary Internet files. *Berkely Technology Law Journaly, 19,* 1227-1273.

[11] Hoffman, S. (2010). An illustration of hashing and its effect on illegal file content in the digital age. *Intellectual Property and Technology Journal, 22,* 6-14.

[12] Magnien, C., Latapy, M., Guillaume, J.L., & Le Grand, B. (2008). Technical report on Paedophile Keywords Observed in eDonkey. Measurement and analysis of P2P activity against paedophile content project. Retrieved from: http://antipaedo.lib6.fr/.

[13] Latapy, M., Magnien, C., & Fournier, R. (2009). Technical report on the Automatic identification of paedophile keywords. Measurement and Analysis of P2P Activity Against Paedophile Content Project. Retrieved from: http://antipaedo.lip6.fr/.

[14] Latapy, M., Magnien, C., & Fournier, R. (2009). Technical report on Quantification of Paedophile Activity in a Large P2P system. Measurement and analysis of P2P activity against paedophile content project. Retrieved from: http://antipaedo.lip6.fr/.

[15] Vehovar, V., Ziberna, A., Kovacic, M., Mrvar, A., & Dousak, M. (2009). Technical report on An Empirical Investigation of Paedophile Keywords in eDonkey P2P Network. Measurement and analysis of P2P activity against paedophile content project. Retrieved from: http://antipaedo.lip6.fr/.

[16] Steel, C.M.S. (2009). Child pornography in peer-to-peer networks. *Child Abuse & Neglect, 33,* 560-568.

.