Deviant Behavior

ISSN: 0163-9625 (Print) 1521-0456 (Online) Journal homepage: https://www.tandfonline.com/loi/udbh20

How Obvious Is It? The Content of Child Sexual Exploitation Websites

Bryce G. Westlake, Martin Bouchard & Ashleigh Girodat

To cite this article: Bryce G. Westlake, Martin Bouchard & Ashleigh Girodat (2017) How Obvious Is It? The Content of Child Sexual Exploitation Websites, Deviant Behavior, 38:3, 282-293, DOI: 10.1080/01639625.2016.1197001

To link to this article: https://doi.org/10.1080/01639625.2016.1197001

DEVIANT BEHAVIOR
2017, VOL. 38, NO. 3, 282–293
http://dx.doi.org/10.1080/01639625.2016.1197001

Routledge
Taylor & Francis Group

How Obvious Is It? The Content of Child Sexual Exploitation Websites

Bryce G. Westlake[a], Martin Bouchard[b], and Ashleigh Girodat[b]

[a]San Jose State University, San Jose, California, USA; [b]Simon Fraser University, Burnaby, BC, Canada

ABSTRACT

Those who distribute child sexual exploitation (CE) material in the public Internet potentially face greater risks of detection. While public distribution is prevalent, little is known about the structure of these websites. We investigate whether websites take steps to hide their purpose, and, if so, what steps are taken? We analyze 634 websites directly or indirectly, via hyperlinks, connected to websites hosting known CE material, and compare our findings to an automated examination of the same websites. We determine whether the initial visual representation is congruent with the underlying structure and content identified in the automated data collection. Implications for understanding cybercriminal processes are discussed.

ARTICLE HISTORY
Received 28 October 2015
Accepted 26 January 2016

The rise of an accessible global marketplace, via the Internet, and the technological advancements of image and video capturing devices have transformed the crime of child sexual exploitation (CE) from a local issue into an international epidemic. By 2000, 77% of CE cases were Internet related (Hughes 2002). In their annual report, the Internet Watch Foundation (2014) identified "31,266 URL's contain[ing] child sexual abuse imagery." Although some of the Uniform Resource Locators (URLs) were obscure, they found that 77% were located at common domains (.com, .net, .ru, .org, and .info). While the use of private methods have increased, distribution continues to occur via public channels (Fortin and Corriveau 2015; Fournier et al. 2014; O'Halloran and Quayle 2010; Rutgaizer et al. 2012; Tremblay 2006; Westlake, Bouchard, and Frank 2011).

The World Wide Web, which includes public websites, is the second most prominent method for distributing and acquiring child sexual abuse images (Carr 2004); however, little is known about the public websites involved. Although the punishment associated with producing or distributing CE material varies between countries (see Gillespie 2012), those that disseminate through public methods theoretically face greater risks of detection, and thus punishment, than those that opt to use private channels. It follows that public websites would be structured in such a way that the dissemination of CE-related material would be discreet or hidden, at least to the casual observer. However, if these websites are *too* discreet, they risk potential consumers being unaware of the material available and not returning. Therefore, public CE-related websites need to find a balance between overtness, to increase success, and covertness, to maintain survival (Westlake and Bouchard 2015).

Despite the social concern surrounding the exploitation of children, there is a dearth of research examining the content and structure of CE-related websites, using qualitative methods. Doing so is important because a central concern of both parents and social control agencies is the accessibility of illicit images and videos on such websites. More accurately, is the illicit material in plain view, on the homepage, or is it buried on the website so that only the initiated may rapidly access it? This study seeks to fill this gap by analyzing the topology of CE websites, focusing on the ease of accessibility of CE images and videos. We manually examined over 600 websites, initially collected using an

CONTACT Bryce G. Westlake Bryce.Westlake@sjsu.edu Department of Justice Studies, San Jose State University, One Washington Square, San Jose, CA, 95192-0050, USA.

automated webcrawler tool that used specific CE-related criteria to identify websites, to determine how obvious it is that a website's primary purpose is CE-related. In addition, we examine each website's use of advertisements and membership options, and the frequency of updates, to get a general portrait of the content of these types of websites.

Website success

On the surface, legal and illegal websites do not differ greatly. The primary goal for each is to implement an effective business strategy that results in success; typically measured through financial gain, notoriety, and/or prestige. Often, this goal is achieved by attracting and keeping consumers, while maintaining a positive reputation. Although a website can survive without achieving success, the two are closely linked (Heldal, SjøVold, and Heldal 2004; Huizingh 2000; Robbins and Stylianou 2003). For CE-related websites, the ability to move beyond simple survival and into success is influenced by the interaction between two website administrator beliefs. First, the degree to which perceived anonymity allows users to openly discuss sensitive topics (e.g., Daniulaityte et al. 2013). Second, the role of content quantity and quality in success or failure. E-commerce research has pointed to four key factors associated with (legal) website success: accessibility, speed, navigability, and quality of content (Hernandez, Jimenez, and Martin 2009; Miranda Gonzalez and Banegil Palacios 2004; Tarafdar and Zhang 2005). Although both legal and illegal websites need to consider all four factors, the methods and level of importance placed on each factor may differ. While legal website administrators only need to focus on maximizing the four key factors, illegal website administrators, theoretically, need to balance their beliefs, with achieving success and maintaining survival.

The first factor central to website success is accessibility. For legal websites, this is generally the simple process of ensuring that consumers can easily find and identify the website's purpose (Miranda Gonzalez and Banegil Palacios 2004); however, for CE-related websites, the process is more complicated. While common CE code-words such as "pthc" (pre-teen hardcore) will direct potential consumers to a website, these code-words are also known by social control agencies and can be used to target CE-related websites for removal. The second factor is speed, as a website's load speed is directly related to the length of a user's stay (Yen, Hu, and Wang 2007). If a website takes more than ten seconds to load users will usually abandon the website and move on to another. The third factor is navigation. When a user accesses a website, they want to be able to find the content they are interested in quickly and directly (Huizingh 2000). Blogs provide the advantage of often being one-paged or log-based, requiring little navigational beyond the homepage. However, this format also impedes the ability to find specific types of material without manually searching through, potentially, endless webpages (Hernandez et al. 2009). Comparatively, non-blogs often require more complex navigation skills; however, if the website design has well-labeled tabs and permanent menus, this format may be preferred. The fourth factor is content quality. For CE-related websites, this typically means images, videos, and stories involving young boys and girls. Ideally, to promote return visits, content is relevant to a user's interest, of high quality, and updated regularly (Tarafdar and Zhang 2005).

Role of websites in cybercrime and CE distribution

Websites play an integral role in facilitating criminal processes, in cyberspace and offline. Public websites have been shown to be used to promote terrorist activities (Davies et al. 2015; Freiburger and Crane 2008; Weimann 2005), and exchange stolen financial information (Holt 2013; Holt and Lampke 2010), drugs (Dolliver 2015; van Hout and Bingham 2013); and copyrighted material (Burruss, Bossler, and Holt 2012). Although research into the cyber-distribution of CE material has focused on public peer-2-peer networks (Latapy, Magnien, and Fournier 2013; Rutgaizer et al. 2012; Steel 2009), CE material is still regularly distributed on the World Wide Web (e.g., Westlake and Bouchard 2015; Westlake, Bouchard, and Frank 2012). Despite the role of public websites, to our knowledge, no study has described the topology of CE websites and manually/visually assessed their content. Specifically whether they are overt or covert in their behavior.

It appears that website administrators are undeterred by the potential increased risk of detection and punishment operating in the public sphere, possibly believing that the benefits (e.g., easy access for consumers) outweigh the corresponding costs. This is not surprising given that research into offline and online deterrence has shown that the threat of apprehension does not deter criminal behavior (Almutairi, Parish, and Phan 2012; Marín, Muñoz Naranjo, and González Casado 2015; Provos and Holz 2007; Spitzner 2003). For online offenders, the perceived, and often real, anonymity naturally provided by the Internet reduces fears associated with conducting illicit activities publicly; even among sex offenders (Armstrong and Forde 2003; Holt, Blevins, and Burkert 2010; Maimon et al. 2014). In fact, the notoriety and respect acquired using a specific alias lead many offenders to maintain the same pseudonym throughout their online career (Décary-Hétu and Dupont 2012; Décary-Hétu, Morselli, and Leman-Langlois 2012). Likewise, websites potentially operate in the public domain because their administrators believe that (a) the website will not be shut down; (b) if shut down, the website can easily be restored; and (c) they can avoid being apprehended and/or punished. Whether these beliefs are true or not, they beg the question: To what degree do criminal-based, public, website administrators' balance attracting consumers with detection-avoidance?

Current study

In this study, we analyzed the content of 634 websites distributing child sexual exploitation material, or indirectly connected to a website hosting known illegal material via hyperlinks, and compare our findings to an automated examination of the same websites. Through our comparison, we seek to make two important contributions to the existing literature on the role of public websites in cybercrime processes. First, we determine the extent to which website administrators are overt in their illicit activities. This contributes to understanding the ease of accessibility for interested offenders, as well as for unsuspecting visitors who accidentally stumble onto a CE-related website. Second, we determine whether website administrators implement covert tactics to hide illicit content on the website. The understanding of strategies used, or not used, to balance appeal to consumers with maintaining survival, contribute to the refinement of detection strategies among social control agencies.

Methods

The study sample consists in 634 websites connected directly or indirectly, via hyperlinks, to at least one website known to contain CE material. The sample was extracted from four "networks of websites" that were selected from a larger project focused on the evolution of websites involved in the dissemination of CE material. The larger project utilized a custom-designed webcrawler, called the Child Exploitation Network Extractor (CENE), to collect data on the websites within ten networks and their corresponding connections. Each network started from a "seed" website, known to be involved in the dissemination of CE material. The webcrawler imitated a snowball sampling technique, "recruiting" websites via the existing hyperlinks found between websites. Each network consisted of more than 300 websites and approximately 500,000 webpages (the webcrawler was instructed to stop when reaching those numbers). Data were collected on the number of webpages, images, videos, keywords, and connections. Websites included in the data collection were required to contain at least one known CE image, or seven of our 82 CE-related keywords, on the hyperlinked webpage. If the website did not meet either criterion it was excluded and CENE continued to the next hyperlinked website.

The data collection yielded over 1,200 websites (including overlaps). It was necessary to exclude a number of websites, for a variety of reasons. During visual inspection, we excluded websites that were (a) not accessible (i.e., offline); (b) membership-based promotion websites[1]; (c) reported to

[1]Some websites were not focused on providing any type of material. Rather they were a single-webpage website that would redirect the user to a specific membership or pay website. The URL of these websites were most often www.join.xyz.com or www.refer.xyz.com.

contain malicious code, according to our antivirus software; (d) blank; or (e) found in more than one analyzed network. This resulted in a final sample size of 634 websites. Manual observation data collection were conducted over a period of two months, with a 14-month follow-up. Websites were analyzed using a Google© Chrome web-browser with manual entry of each website's URL address. As our primary objective was to determine how obvious it was, at an initial glance, that a website was focused on disseminating CE-related material, we only visually inspected each website's homepage. The visual inspection included determining the type of website (blog/non-blog), victim sex (boy/girl), and medium of distribution (image/video), and whether the website was CE-related, up-to-date, or advertised, and if it had a membership/registration option.

Website inclusion criteria (keywords and CE images)

The CE image database used by CENE was provided by the Royal Canadian Mounted Police. Last updated one month prior to the initial wave of data collection, the database contained 2.25 million hash values[2] classified into three groups (see Table 1), based on Canadian legal definitions of CE material. The first group (*Child Exploitation)* contained 618,632 images that met the Canadian Criminal Code of Canada (1985) definition[3] of CE material and had been used in prosecuting cases involving distribution. The second group (*Child Nudity)* contained 652,223 images that would probably be considered CE material but had not yet been brought before a judge. The third group (*Collateral)* contained 981,232 images that do not meet the definition of CE material, but were important enough to be collected by offenders. Images in this category may have included initial photographs taken by an offender, of a child, prior to the removal of clothing.

The 82 keywords used by CENE were found to be the most prevalent in previous research conducted on the topic of online CE (Le Grand et al. 2009; Steel 2009; Vehovar et al. 2009). The 82 keywords were categorized into three groups (see Table 1). The first group were *Code* keywords (27) commonly used by offenders to alert one another to material, such as "pthc." The second group were *Thematic* keywords (23) not directly linked to CE material but typically present on such websites (e.g., boy, girl, child). The third group were *Sex-Oriented* keywords (32) referencing sexual organs and acts (e.g., pussy, cock, oral).

Observational variables—Automated

Sex of victim (Boy/girl): Using the relative frequency of specific keywords, we classified each website as being *boy* or *girl* focused. The keywords used for these classifications were: boy, son, twink, penis, and cock, or girl, daughter, nimphets/nymphets, Lolita/lola/lolli/lolly, vagina, and pussy.

Medium (Image/video/story): Using the relative frequency of three types of media found on websites, we classified each website as primarily distributing images, videos, or stories. First,

Table 1. Images and keywords inclusion criteria used by CENE to identify CE-related websites.

	Images (Number)	Keywords (Number)
Category 1	Child exploitation (618,632)	Child exploiter-code (27)
Category 2	Child nudity (652,223)	thematic (23)
Category 3	Collateral (981,231)	sex-oriented (32)

[2]A hash value is a 32-hexidecimal code that functions similar to a digital fingerprint. Each computer file is given a hash value based on its binary composition. When a file is edited, even minimally, a new hash value is created. Tretyakov et al. (2013) state that the chances of two files having the same hash value is "negligibly small" ($1/2^{2^{48}}$).

[3]Under section 163.1 (1) of the Canadian Criminal Code (1985), *child pornography* includes any "photographic, film, video, or other visual representation … written material … or audio recording" of a person under the age of eighteen, engaged in an explicit sexual act, or advocating sexual activity. Those depicted can include imaginary people.

for *image* and *video*, we used the average number of instances of each medium found on a webpage. For *story*, we used the average number of our 82 keywords found on a webpage. Second, each website was given a standardized score (0.00 to 1.00) for each medium, relative to the other websites within the network. For example, the website with the most images per webpage was given a score of 1.00 while every other website was standardized against this website, on the same measure. For whichever medium a website received the highest score was its classification.

Observational variables—Manual

While the Child Exploitation Network Extractor collected data on each website's content, there data were not viewed by the (female) research assistant prior to their manual observation. This was done to allow for an unbiased judgment of whether or not the website was perceived to be hosting CE-related material. We acknowledge the process is inherently subjective, which cannot be avoided, but also serves the additional purpose of reproducing similar assessment from real website visitors (even if their experience and tastes could have led to different results than here). All perceptions of a website's characteristics were based on the following criteria.

Child sexual exploitation related: Although it can be difficult to determine if a person depicted is under the age of 18, there are physiological characteristics more commonly found in younger people. For example, those under the age of 18 often have a lack of body hair, a leaner body mass, shorter height, and less sexual organ development. Erring on the side of caution, a website was classified as CE-related if the visual and/or written content depicted or targeted persons appearing to be under the age of 18.

Type of website (Blog/non-blog): A website was classified as being a *blog* if it followed a web-log format typically of the style of a single webpage with a continuous, often chronological, list of user postings. Multiple-page websites, with no sub-headings and only a page counter, were also classified as blogs. A *non-blog* was any multiple-page website not classified as blog. This included discussion forums, photo galleries, and directories to other websites.

Sex of victim (Boy/girl): The sex of victim focus was determined through the genitalia focus. If the visual and/or written descriptions were primarily penis-focused, the website was classified as *boy*.

Medium (Image/video/both): The medium of distribution focus was determined through the visual quantity of images and videos available. If either did not appear to be the overwhelming focus, the website was classified as 'both'.

Up-to-date: A website was classified as being up-to-date if the most recent additions were within the three months prior to the observation. This was determined by the date on the most recent postings. If the posting date was not evident, the "last updated" marker at the bottom of many webpages was used.

Advertisements: A website's advertising techniques, to other websites, were classified as (a) hyperlinks dispersed throughout the webpage or within posts; (b) a specific area for connections to other websites—known as blogrolls; (c) a combination of the two methods; or (d) no methods used. There was no size definition to the advertisement and it could be communicated through pop-ups, hyperlinks, images, or a combination.

Membership: Although all websites were publicly accessible, we noted those that also provided the option to register. Therefore, a website was classified as including membership if there was an option to register or login.

Results

There were two data collections conducted for this study. The first was by the automated webcrawler tool CENE. The second was the manual, visual, investigation of each website's homepage. While the overall purpose of this study was to determine how obvious it was, visually, that a website was

explicitly CE-focused, this question can only be answered in the context of a website's overall content, beyond the homepage, and how well it hides its intentions; information found from the automated data collection process. Through the automated CENE data collection, 47 websites were found to include *Child Exploitation* (Cat.1), *Child Nudity* (Cat.2), and/or *Collateral* (Cat.3) images. Overall, 6,597 CE-related images were identified, with 3,373 being Child Exploitation. Among the 47 websites with CE-related images, 33 had Child Exploitation images; with 11 having at least 10 such images.

Blind to the data collected by CENE, the manual investigation identified 31 of the 33 websites with Child Exploitation images, as well as four without any known CE-related images and one with 106 Child Nudity images. The two websites identified by CENE but not identified in the manual investigation consisted of only two and three Child Exploitation images. The high correlation between the manual identification of explicitly CE-focused websites and the automated identification of websites with Child Exploitation images suggests that CE-related websites do little to hide their purpose. Table 2 summarizes the descriptive results for the 634 websites visited. Results are compared between the sample of websites CENE identified as having Child Exploitation images (n = 33), websites *only* manually identified as being CE-focused (n = 5), and all websites not identified as being explicitly CE-focused by either the automated or manual data collection (n = 596). Statistically significant results are noted, and have taken into consideration unequal variance. Statistical significance, though provided, should be interpreted with caution given the small size of the manually identified CE website category (n = 5).[4]

Keywords are important for identifying CE-related websites. Some keywords, such as *pthc*, are well-known to be related to CE material. As a result, these types of *Code* keywords can used by distributors to attract consumers; however, they can also be used by those seeking to detect and remove content. This knowledge, by offenders, may deter the use of Code keywords and result in a greater reliance on *Thematic* and/or *Sex-Oriented* keywords. Table 2 shows that websites with known Child Exploitation images averaged six Code keywords, per webpage, while non-identified websites averaged 180. Among explicitly CE-focused websites, identified automatically or manually, Thematic and Sex-Oriented keywords were *more* prevalent than among non-identified websites. These results suggest that while websites may be visually obvious in their CE purpose, content specific/descriptive keywords appear to be favored over Code keywords.

Other website characteristics

CE material is broad and covers a wide assortment of content foci and distribution methods. In our manual observation we identified each website's primary victim focus and medium of distribution, the average number of outgoing and incoming hyperlinks, and whether it (a) was a blog/non-blog; (b) was up-to-date; (c) used advertisements; and (d) had an option for membership. We also compared the correlation between the automatic and manual investigations, characterizing each website's victim focus and distribution method.

Recall that a boy-focused website was determined by the relative frequency of boy-related keywords in the automated data collection, and by the more prominent visual focus in the manual data collection. Between the two methods 94% of websites were categorized the same, with only 37 (6%) being labeled differently. The majority of the discrepancy stemmed from websites being labeled as girl-focused in the manual data collection but boy-focused in the automated data collection. Of these websites, 15 were described as "girls performing sexual acts on men"; typically, oral sex. Given the automated data collection used the presence of keywords, it is likely that the descriptions were focused on what was happening to the men, pointing to it being boy-focused, while our manual observation

[4]This category may be small, but is very important to the validation of automated webcrawlers such as the one used in this study. These five websites would have been missed, for example, in a study where the child pornography label is only attributed to websites where a known illegal image is found (i.e., via its hash value).

Table 2. Website characteristics for all websites, CENE identified CE websites, and manually identified CE websites.

		CENE identified CE websites (n = 33)	Manual only identified CE websites (n = 5)	Non-identified websites (n = 596)	All websites (n = 634)
No. CE-related		33	5	0	38
No. Child exploitation	Cat. 1	3,373[ac]	0	0	3,373
	Cat. 2	4	106	76	186
Avg. CE code keywords		6[a]	168	180	171
Avg. thematic keywords		879,333	1,649,803	180,625	228,580
Avg. sex-oriented keywords		300,014	648,626	207,236	215,547
Pct. boy		100.0[b]	80.0	88.0	84.3
Pct. image		27.3	0.0	25.8	25.7
Pct. video		42.4[b]	40.0[b]	48.5	48.1
Avg. outgoing links		85.5[b]	47.4	21.5	25.1
Avg. incoming links		28.0[bc]	14.4	20.9	21.2
Pct. blog format		78.8[d]	0.0[a]	70.0	70.7
Pct. up-to-date		57.1	—	54.3	54.4
Pct. advertisements		60.6[c]	20.0[a]	67.4	66.6
Pct. membership		3.0[b]	0.0	21.1	20.0
Pct. active at 14-month follow-up		84.8	80.0	84.9	84.9

[a]Statistically significantly (0.1) different than non-identified websites.
[b]Statistically significantly (0.01) different than non-identified websites.
[c]Statistically significantly (0.1) different than manual only websites.
[d]Statistically significantly (0.01) different than manual only websites.

focused on the person performing the sexual acts. Comparing websites with known Child Exploitation images to non-identified websites (see Table 2), boy-focused websites were more prevalent among the former ($p < .01$).

The two primary mediums of CE distributed are images and videos. While images have been traditionally more common, advancements in video recording devices and Internet speeds and storage options suggest that CE video distribution may be on the rise. There are two common ways that videos are displayed on webpages. The first is through directly providing (hosting) the video. The second is indirectly, through embedding a video on the webpage, hosted by another website. Through the first method, higher storage capabilities are required. Through the second method, storage requirements are reduced; however, the hosting website keeps a web-log of other websites accessing their video. Therefore, if the hosting website is detected, and removed, social control agencies can view the web-log to identify other CE-related websites. Explicitly CE-focused websites, identified in *both* the manual and automatic data collections, were less likely to be video-focused than websites not identified as being CE-focused.

The agreement between the automated and manual data collection identification of images or videos distribution was 41%; however, this low agreement is a bit misleading. Where a website was described as being image-focused in the manual observation agreement was high. In less than 5% was a website described as image-focused manually but video-focused automatically. In comparison, of the 282 websites identified as video-focused manually, 88% were described as image-focused in the automatic data collection. If the results of the manual data collection are deemed more representative (i.e., accurate), there is support for the increase in video distribution methods and the need for strategies and criteria for better identification.

The embeddedness of any website within a larger community can impact the ability to inform others of the website's focus and attract new consumers. Outgoing hyperlinks can be a means for a website to identify its overall focus (e.g., connecting to other CE-related websites), while incoming hyperlinks can be used to subtly advertise the intent of a website, and potentially give instructions for finding hidden content. Part of the automated data collection process included identifying the connections made between websites within the larger network. Our analyses, summarized in Table 2, revealed that websites with known Child Exploitation images averaged more outgoing (85.5) and incoming (28.0) hyperlinks than non-identified websites (21.5 and 20.9) and more outgoing hyperlinks than manually (only) identified websites (14.4).

How obvious a CE website is with its purpose can be influenced by multiple factors. First, a non-blog with multiple webpages can option to be more discreet as they can easily disseminate content on "hidden" webpages. Second, a website can appear not to be up-to-date, again only posting new content on hidden webpages. Third, they can use advertisements to connect to like-minded websites. Fourth, they can hint toward their purpose but require registration for access to specific content. In Table 2 we compare our three samples across each of these variables.

Websites with or without known Child Exploitation images were equally as likely to (a) be a blog; (b) be up-to-date; and (c) post advertisements to other websites. Where websites with Child Exploitation images differed from non-identified websites was on options for membership/registration. To specifically address our overall research question of how obvious is the CE-focus of a website, those with known Child Exploitation images gave consumers no options to register and view more content. In other words, those that had verified Child Exploitation images made no efforts to disguise their purpose or to provide additional content beyond what was freely available. Of course, those that were not identified as explicitly CE-focused, but did have an option for membership, may have been better at disguising their purpose. However, the manual, visual, and data collection did not provide any evidence to suggest this was the case. Therefore, even in this scenario, only consumers that were privy to "insider information" would know to register to obtain private content. A potential explanation for the lack of secrecy is that website operators know consumers will not register to view images or related material, for fear of identification, and therefore do not provide that option.

14-month follow-up

Do illegal websites have to balance success and survival by operating more covertly, or does the benefit of attracting consumers, through being explicit in their purpose, outweigh any risk of failure? We conducted a 14-month follow-up on the 634 websites visited to determine which remained active and which failed. Our findings suggest that the presence of Child Exploitation images and/or the explicit CE focus of a website did not impact survival. Across the full sample, 84.9% of websites were active at the follow-up, compared to 84.8% among websites with Child Exploitation images and 80% among manually (only) identified CE-focused websites. Combining the CENE and manually identified groups, 32 of the 38 (84%) remained active. Of the six websites not active at the 14-month follow-up, five had Child Exploitation images. However, only two of those failed websites had more than 2 images, while all websites with 40 or more—including one with 1,452 and another with 1,447—remained active. Of note is the sixth failure, which contained 106 Child Nudity images. Although five of the failures were classified as blogs, none were 'sub-hosted' by a third-party service (e.g., Blogger©). Finally, all six failures were manually described as being video-focused. While the failure size does not allow for generalizable conclusions to be drawn, it might be an indication of the medium priority for removal efforts.

Discussion

How obvious is it that a public website is disseminating CE material? Within this study we sought to answer this question by visually inspecting 634 websites collected automatically, using CE-related criteria from previous research. We compared the findings of our manual inspection to the data collected by our custom-designed webcrawler, to determine whether websites with known CE images were covert or overt in their CE intentions. Overall, we found that websites with Child Exploitation images did not try to hide their intention. And despite their overt purpose, a 14-month follow-up revealed that explicitly CE-focused websites were no more likely to fail than other websites. We discuss three important implications of our findings to the study of online child sexual exploitation distribution.

Websites disseminating known Child Exploitation images and/or explicitly CE-focused appeared to do little to hide their intended purpose. The general structure of these websites was to provide direct access to the CE-related material that consumers were seeking. This finding reaffirms research that illicit activities are conducted overtly, in public spaces, on the World Wide Web (Armstrong

and Forde 2003; Holt et al. 2010; Maimon et al. 2014; O'Halloran and Quayle 2010; Tremblay 2006). This finding, coupled with that of previous researchers, points to an important aspect of combating cybercrime. As many illicit activities occur in public spaces, their detection is not difficult. While easy to detect, the vast number of activities occurring, combined with jurisdictional, privacy, and identification issues, make combating difficult (see Gillespie 2012). The complexity of social control for cybercrime was evident in our 14-month follow-up findings. Websites with Child Exploitation and CE-related images were equally likely to be active as those without said images. This finding reinforces offender assumptions that conducting illicit activities in the public realm of the Internet does not, at a general level, increase risk of failure. That is, specific to tactics used by website administrators to balance success and survival, offenders do not appear to take great strides to hide their intended purpose, and failing to do so does not appear to negatively impact survival.

The abundance of cybercrime activities occurring globally impact the ability of social control agencies to keep pace. One potential option is to automate some social control efforts. The similarities in our automatic and visual CE-related detection, and keyword-based focus identification, point to the ability to use autonomous tools and techniques, which follow topic-specific criteria, to target cybercrime operations. In general, this speeds up the process and allows agents to focus on processing cases rather than identifying them. For those addressing cybercrimes of a more graphic nature, such as child sexual exploitation, the use of automated detection can assist in reducing the high emotional and psychological trauma associated (Bourke and Craun 2014; Burns et al. 2008; Craun, Bourke, and Coulson 2015; Krause 2009; Perez et al. 2010).

The ability for autonomous tools and techniques to be successful is contingent on their validity and reliability. Our findings showed that *Thematic* keywords, such as "boy," "girl," and "child," were more frequent on websites with Child Exploitation images and those manually identified as explicitly CE-focused, while *Code* keywords, such as preteen hardcore (pthc) and preteen softcore (ptsc), were not. For autonomous data collection tools to be effective, they need to be provided with proper guidance. The lack of prevalence of Code keywords point to another difficulty in combating distribution: language is continually evolving. The evolution of language occurs naturally, but also in response to social control efforts. As offenders learn that certain keywords have become commonplace, they may modify their language—a tactic to avoid detection and maintain survival—and develop new "code" keywords. This means that autonomous tools need to include keywords that extend beyond those typically used to specifically identify CE content, to keywords that describe the children and activities being disseminated. While this language will change over time too, it is less likely to be as quick. In some cases, the language used will never change. For instance, terms like "boy," "smooth," "soft," and "young" will remain ever present. Therefore, failure to include these types of keywords can result in websites without known CE images (i.e., websites with newly produced content) being missed. The challenge then becomes finding the correct balance, to minimize false positives, given the generalness of the non-code keywords.

Our manual observation of the general structure of websites also pointed to a growth in the number of websites focused on distributing videos. This is further supported by evidence of the quickly growing market of live webcam sexual exploitation (Bryce 2010; Hillman, Hooper, and Raymond Choo 2014; Kloess, Beech, and Harkins 2014). Efforts to combat video distribution appears to lag behind that of image distribution as there are no CE video databases currently being used by social control agencies. However, for autonomous detection efforts to be successful, specifically for CE-related material, there is a need for researchers and law enforcement agencies to hasten development on a video database, similar to hash value databases used for identifying images.

Conclusions

As the Internet continues to grow and expand, so too does the sexual exploitation of children and the dissemination of child sexual exploitation (CE) material. While steps have been made in recent times to combat this disturbing trend, the dynamic landscape of the Internet is as such that researchers need to be continuously molding their understanding of the situation to suit the

changing environment. Child exploitation is no longer a hidden, underground, realm only accessed by sophisticated, technological masters. Material is readily available to anyone of simple to moderate technological skill—essentially, if an individual can access and use a search engine with a modicum of skill, they can assuredly find CE material.

The current study examined different elements of Web platforms hosting known CE images. The purpose was to understand the layout and the initial user experience to determine how obvious CE-related material is on the World Wide Web. This current research should be viewed as a preliminary study that examines what aspects of a Web-based platform are present on illegal websites and how accessible that illegal content is to arbitrary viewers on the Internet. Future research needs to go beyond the homepage of the websites and delve further into what can be housed openly. Through understanding the layout, the accessibility, and the readily available content, social control agencies can more effectively use current tools to remove the content from the Internet.

Further research should also be conducted, in a more in-depth manner, comparing completely legal websites to those hosting illegal content. While the current study compared the factors needed for a successful legal website and the factors present on illegal websites, no actual legal websites were analyzed. False positives found by a webcrawler program could potentially be avoided if the elements of both kinds of Web platforms were cross-examined.

Acknowledgment

The authors thank Dr. Richard Frank for his work on creating the webcrawler used for this study.

Funding

We thank the Social Sciences and Humanities Research Council for funding this project (#435-2012-0336).

Notes on contributors

BRYCE G. WESTLAKE is an Assistant Professor of justice studies at San Jose State University. His work focuses on digital evidence, online sexual offending, and illegal social networks, melding criminology with computer science. His current research focuses on child sexual exploitation material in cyberspace and the longitudinal evolution of these online distribution networks. Findings from this research have been published in *Justice Quarterly, Sexual Abuse*, and *Social Science Research.*

MARTIN BOUCHARD is an Associate Professor of criminology at Simon Fraser University. His work focuses on the organization and dynamics of illicit markets and on examining the impact of social networks in various criminal career outcomes. He also published extensively on street gangs, organized crime, online illicit networks, and methodologies to estimate the size of illicit markets. His current projects focus on the social structure of serious crime in British Columbia, and its implications for understanding the dynamics of violence and illicit markets in the area.

ASHLEIGH GIRODAT completed a Bachelor of Arts, with a Certificate in Police Studies, in Simon Fraser University's School of Criminology. She is a Research Assistant at the International Cybercrime Research Centre and examines the evolution of online child sexual exploitation networks. In 2014, her work in this area led to the receipt of the Social Science and Humanities Research Council's annual Storytellers Award.

References

Almutairi, Abdulrazzaq, David Parish, and Raphael Phan. 2012. "Survey of High Interaction Honeypot Tools: Merits and Short-comings." Retrieved January 14, 2016 (http://www.cms.livjm.ac.uk/pgnet2012/Proceedings/Papers/1569604821.pdf).

Armstrong, Helen L. and Patrick J. Forde. 2003. "Internet Anonymity Practices in Computer Crime." *Information Management & Computer Security* 11:209–215.

Bourke, Michael L. and Sarah W. Craun. 2014. "Secondary Traumatic Stress among Internet Crimes Against Children Task Force Personnel Impact, Risk Factors, and Coping Strategies." *Sexual Abuse: A Journal of Research and Treatment* 26:586–609.

Bryce, Jo. 2010. "Online Sexual Exploitation of Children and Young People." Pp. 320–342 in *Handbook of Internet Crime*, edited by Y. Jewkes and M. Yar. New York: Routledge.

Burns, Carolyn M., Jeff Morley, Richard Bradshaw, and José Domene. 2008. "The Emotional Impact on and Coping Strategies Employed by Police Teams Investigating Internet Child Exploitation." *Traumatology* 14:20–31.

Burruss, George W., Adam M. Bossler, and Thomas J. Holt. 2012. "Assessing the Mediation of a Fuller Social Learning Model on Low Self-Control's Influence on Software Piracy." *Crime & Delinquency* 59:1157–1184.

Canadian Criminal Code, *RSC*. 1985. c C–46 s163.

Carr, Angela. 2004. *Internet Traders of Child Pornography and Other Censorship Offenders in New Zealand*. Wellington, NZ: Department of Internal Affairs. Retrieved November 10, 2015 (http://www.dia.govt.nz/Pubforms.nsf/URL/entirereport.pdf).

Craun, Sarah W., Michael L. Bourke, and Frances N. Coulson. 2015. "The Impact of Internet Crimes against Children Work on Relationships with Families and Friends: An Exploratory Study." *Journal of Family Violence* 30:393–402.

Daniulaityte, Raminta, Robert Carlson, Russel Falck, Delroy Cameron, Sujan Perera, Lu Chen, and Amit Sheth. 2013. ""I Just Wanted to Tell You that Loperamide WILL WORK": A Web-Based Study of Extra-Medical Use of Loperamide." *Drug and Alcohol Dependence* 130:241–244.

Davies, Garth, Martin Bouchard, Edith Wu, Kila Joffres, and Richard Frank. 2015. "Terrorist and Extremist Organizations' Use of the Internet for Recruitment." Pp. 105–127 in *Social Networks, Terrorism and Counter-Terrorism*, edited by M. Bouchard. New York: Routledge.

Décary-Hétu, David and Benoit Dupont. 2012. "The Social Network of Hackers." *Global Crime*, 13:160–175.

Décary-Hétu, David, Carlo Morselli, and Stéphane Leman-Langlois. 2012. "Welcome to the Scene: A Study of Social Organization and Recognition among Warez Hackers." *Journal of Research in Crime and Delinquency* 49:359–382.

Dolliver, Diana S. 2015. "Evaluating Drug Trafficking on the TOR Network: Silk Road 2, the Sequel." *The International Journal of Drug Policy*. doi:10.1016/j.drugpo.2015.01.008.

Fortin, Francis and Patrice Corriveau. 2015. *Who is Bob_34?* Vancouver BC: UBC Press.

Fournier, Raphaël, Thibault Cholez, Matthieu Latapy, Isabelle Chrisment, Clémence Magnien, Olivier Festor, and Ivan Daniloff. 2014. "Comparing Pedophile Activity in Different P2P Systems." *Social Sciences* 3:314–325.

Freiburger, Tina and Jeffrey S. Crane. 2008. "A Systematic Examination of Terrorist Use of the Internet." *International Journal of Cyber Criminology* 2:309–319.

Gillespie, Alisdair A. 2012. *Child Pornography: Law and Policy*. New York: Routledge.

Heldal, Frode, Endre SjøVold, and Anders Foyn Heldal. 2004. "Success on the Internet—Optimizing Relationships through the Corporate Site." *International Journal of Information Management* 24:115–129.

Hernandez, Bianca, Julio Jimenez, and Jose Martin. 2009. "Key Website Factors in E-business Strategy." *International Journal of Information Management* 29:362–371.

Hillman, Henry, Christopher Hooper, and Kim-Kwang Raymond Choo. 2014. "Online Child Exploitation: Challenges and Future Research Directions." *Computer Law & Security Review* 30:687–698.

Holt, Thomas J. 2013. "Examining the Forces Shaping Cybercrime Markets Online." *Social Science Computer Review* 31:165–177.

Holt, Thomas J. and Eric Lampke. 2010. "Exploring Stolen Data Markets On-Line: Products and Market Forces." *Criminal Justice Studies* 23:33–50.

Holt, Thomas J., Kristie R. Blevins, and Natasha Burkert. 2010. "Considering the Pedophile Subculture On-Line." *Sexual Abuse: Journal of Research and Treatment* 22:3–24.

Hughes, Donna M. 2002. "The Use of New Communications and Information Technologies for Sexual Exploitation of Women and Children." *Hastings Women's LJ* 13:127–325.

Huizingh, Eelko K. R. E. 2000. "The Content and Design of Web Sites: An Empirical Study." *Information and Management* 37:123–134.

Internet Watch Foundation. 2014. *IWF Operational Trends 2014*. Retrieved July 7, 2016 (https://www.iwf.org.uk/accountability/annual-reports/2014-annual-report).

Kloess, Juliane A., Anthony R. Beech, and Leigh Harkins. 2014. "Online Child Sexual Exploitation: Prevalence, Process, and Offender Characteristics." *Trauma, Violence and Abuse* 15:126–139. doi:10.1177/1524838013511543.

Krause, Meredith. 2009. "Identifying and Managing Stress in Child Pornography and Child Exploitation Investigators." *Journal of Police and Criminal Psychology* 24:22–29.

Latapy, Matthieu, Clémence Magnien, and Raphaël Fournier. 2013. "Quantifying Paedophile Activity in a Large P2P System." *Information Processing & Management* 49:248–263.

Le Grand, Bénédicte, Jean-loup Guillaume, Matthieu Latapy, and Clémence Magnien. 2009. "Dynamics of Paedophile Keywords in eDonkey Queries: Measurement and Analysis of P2P Activity against Paedophile Content Project." Retrieved March 15, 2015 (http://antipaedo.lib6.fr/).

Maimon, David, Mariel Alper, Bertrand Sobesto, and Michel Cukier. 2014. "Restrictive Deterrent Effects of a Warning Banner in an Attacked Computer System." *Criminology* 52:33–59.

Marín, José Manuel Fernández, Juan Álvaro Muñoz Naranjo, and Leocadio González Casado. 2015. "Honeypots and Honeynets: Analysis and Case Study." Pp. 452–482 in *Handbook of Research on Digital Crime, Cyberspace Security, and Information Assurance*, edited by M. M. Cuz-Cunha and R. M. Portela. Hershey: IGI Global.

Miranda Gonzalez, F. J. and T. M. Banegil Palacios. 2004. "Quantitative Evaluation of Commercial Web Sites." *International Journal of Information Management* 24:313–328.

O'Halloran, Elaine and Ethel Quayle. 2010. "A Content Analysis of a 'Boy Love' Support Forum: Revisiting Durkin and Bryant." *Journal of Sexual Aggression* 16:71–85.

Perez, Lisa M., Jeremy Jones, David R. Englert, and Daniel Sachau. 2010. "Secondary Traumatic Stress and Burnout among Law Enforcement Investigators Exposed to Disturbing Media Images." *Journal of Police and Criminal Psychology* 25:113–124.

Provos, Niels and Thorsten Holz. 2007. *Virtual Honeypots: from Botnet Tracking to Intrusion Detection*. Boston: Pearson Education.

Robbins, Stephanie S. and Antonis C. Stylianou. 2003. "Global Corporate Web Sites: An Empirical Investigation of Content and Design." *Information & Management* 40:205–212.

Rutgaizer, Moshe, Yuval Shavitt, Omer Vertman, and Noa Zilberman. 2012. "Detecting Pedophile Activity in BitTorrent Networks." *Lecture Notes in Computer Science* 7192:106–115.

Spitzner, Lance. 2003. *Honeypots: Tracking Hackers* (Vol. 1). Reading: Addison-Wesley.

Steel, Chad M. S. 2009. "Child Pornography in Peer-to-Peer Networks." *Child Abuse & Neglect* 33:560–568.

Tarafdar, Monideepa and Jie Zhang. 2005. "Analysis of Critical Website Characteristics: A Cross-Category Study of Successful Websites." *Journal of Computer Information Systems* 46:14–24.

Tremblay, Pierre. 2006. "Convergence Settings for Nonpredatory 'Boy Lovers.'" Pp. 145–168 in *Situational Prevention of Child Sexual Abuse*, edited by R. Wortley and S. Smallbone. Monsey: Criminal Justice Press.

Tretyakov, Konstantin, Sven Laur, Geert Smant, Jaak Vilo, and Pjotr Prins. 2013. "Fast Probabilistic File Fingerprinting for Big Data." *BMC Genomics* 14:S2–S8. doi:10.1186/1471-2164-14-S2-S8

van Hout, Marie Claire and Tim Bingham. 2013. "'Silk Road', the Virtual Drug Marketplace: A Single Case Study of User Experiences." *International Journal of Drug Policy* 24:385–391.

Vehovar, Vasja, A. Ziberna, M. Kovacic, Andrej Mrvar, and M. Dousak. 2009. "An Empirical Investigation of Paedophile Keywords in eDonkey P2P Network: Measurement and Analysis of P2P Activity against Paedophile Content Project." Retrieved March 15, 2015 (http://antipaedo.lib6.fr/).

Weimann, Gabriel. 2005. "How Modern Terrorism Uses the Internet." *The Journal of International Security Affairs* 8:91–105.

Westlake, Bryce G. and Martin Bouchard. 2015. "Criminal Careers in Cyberspace: Examining Website Failure within Child Exploitation Networks." *Justice Quarterly* doi:10.1080/07418825.2015.1046393

Westlake, Bryce G., Martin Bouchard, and Richard Frank. 2011. "Finding the Key Players in Child Exploitation Networks." *Policy & Internet* 3(2). doi:10.2202/1944-2866.1126

———. 2012. "Comparing Methods for Detecting Child Exploitation Content Online." Paper presented at the *European Intelligence and Security Informatics Conference 2012*, Odense, Denmark.

Yen, Benjamin, Paul Jen-Hwa Hu, and May Wang. 2007. "Toward an Analytical Approach for Effective Web Site Design: A Framework for Modeling, Evaluation and Enhancement." *Electronic Commerce Research and Applications* 6:159–170.