9-21-2021

Detecting child sexual abuse images: Traits of child sexual exploitation hosting and displaying websites

Enrique Guerra
San Jose State University

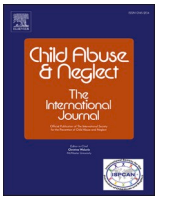
Bryce G. Westlake
San Jose State University, bryce.westlake@sjsu.edu

Child Abuse & Neglect 122 (2021) 105336

Contents lists available at ScienceDirect

Child Abuse & Neglect

journal homepage: www.elsevier.com/locate/chiabuneg



Detecting child sexual abuse images: Traits of child sexual exploitation hosting and displaying websites

Enrique Guerra, Bryce G. Westlake *

Department of Justice Studies, San Jose State University, San Jose, CA, USA

ARTICLE INFO

Keywords:
Child sexual exploitation images
Child sexual abuse images
Child pornography
Limitation of hash values
Websites
Automated data collection

ABSTRACT

Background: Automated detection of child sexual abuse images (CSAI) often relies on image attributes, such as hash values. However, electronic service providers and others without access to hash value databases are limited in their ability to detect CSAI. Additionally, the increasing amount of CSA content being distributed means that a large percentage of images are not yet cataloged in hash value databases. Therefore, additional detection criteria need to be determined to improve identification of non-hashed CSAI.
Objective: We aim to identify patterns in the locations and folder/file naming practices of websites hosting and displaying CSAI, to use as additional detection criteria for non-hashed CSAI.
Methods: Using a custom-designed web crawler and snowball sampling, we analyzed the locations and naming practices of 103 Surface Web websites hosting and/or displaying 8108 known CSAI hash values.
Results: Websites specialize in either hosting or displaying CSAI with only 20% doing both. Neither hosting nor displaying websites fear repercussions. Over 27% of CSAI were displayed in the home directory (i.e., main page) with only 6% located in at least 4th-level sub-folder. Websites focused more on organizing images than hiding them with 68% of hosted and 54% of displayed CSAI being found in folders formatted year/month. Qualitatively, hosting websites were likely to use alphanumeric or disguised folder and file names to conceal images, while displaying websites were more explicit.
Conclusion: File and folder naming patterns can be combined with existing criteria to improve automated detection of websites and website locations likely hosting and/or displaying CSAI.

1. Introduction

Videos and images depicting children in sexual activities are distributed heavily throughout the Surface and Deep Web. This child sexual abuse media (CSAM) can be found on websites, forums, social media, chat applications, messenger services, file sharing programs, and most other places connected to the internet (Steel et al., 2020; Westlake, 2020). Because of the large volume of CSAM being distributed, law enforcement agencies and researchers have increasingly turned to automated tools to detect media (see Lee et al., 2020 for summary). Although these tools have proven useful, they are often limited by their reliance on matching detected media to previously known media, via different hashing techniques. This is a major limitation as an analysis of more than 23 million reports to the National Center for Missing and Exploited Children (NCMEC) found that 84% of images and 91% of videos were flagged only once

* Corresponding author at: San Jose State University, One Washington Square, San Jose, CA 95192-5000, USA.
E-mail address: Bryce.Westlake@sjsu.edu (B.G. Westlake).

https://doi.org/10.1016/j.chiabu.2021.105336
Received 29 December 2020; Received in revised form 25 June 2021; Accepted 15 September 2021
Available online 21 September 2021

(Bursztein et al., 2019). This means that in most cases, knowing the hash value of CSAM does not increase detection as so many files reported are previously unknown. Therefore, additional detection methods need to be considered that do not rely on hash values.

Additional methods for identifying CSAM have begun to be explored, including using visual and audio characteristics of images and videos (e.g., Schulze et al., 2014) or using the naming attributes of files being shared within peer-2-peer networks (Panchenko et al., 2012; Panchenko et al., 2013). We propose that similar methods can be applied to identify patterns in file and folder structuring (e.g., location and naming) on websites hosting and displaying child sexual abuse images (CSAI). Westlake et al. (2017) found websites that display CSAM on the Surface Web do little to hide their intended purpose, visually. If websites also do little to hide CSAM at the structural level, this could be a useful avenue for narrowing down places to search for media on websites. In the current study we examine the underlying file structure of 103 websites displaying and/or hosting 8108 known (hashed) CSAI on the Surface Web. The aim is to identify patterns quantitatively and qualitatively in sub-folder structuring, folder and file naming, and image hosting locations, to be integrated into automated detection tools.

1.1. Combating child sexual abuse media distribution

In 2019, the NCMEC (2020) received nearly 16.9 million reports from 153 electronic service providers (ESP) of suspected CSAM. These reports totaled 69.1 million videos, images, and files. In many of these reports it is likely that the CSAM was detected through automated tools, with some reports being the result of users 'flagging' the media as inappropriate. Historically, detection tools have relied on keywords commonly associated with CSAM and hash value databases. A hash value is a hexadecimal code which acts like a digital fingerprint for any file. The probability of two files having the same MD5 hash value is, at least, 1.47×10^{-29} (Ramirez, 2015). However, these methods have clear limitations. First, language is constantly changing and as soon as a keyword is identified as being relevant to CSAM, it has likely become irrelevant or outdated. Second, websites may highlight keywords not allowed to be used on their platform. Automated tools will identify these as being related to CSAM distribution, when they are actually false positives. Third, the hash value of a file changes as soon as the file is modified. This may be done by the end user, such as through cropping, but could also be done automatically by the website during the upload process, such as reducing the size or putting it in a specific format. Fourth, while steps have been made to develop video-based hash value databases, most are still image-based. As internet access and speed continues to increase, videos are becoming more popular, with 2019 marking the first time that the reporting of CSA videos has outpaced the reporting of CSAI (Dance & Keller, 2020). Fifth, because the use of hash value databases relies on the CSAM being previously identified and cataloged, they are ineffective for detecting new CSAM. The limitations of keywords and hash values has led to additional techniques being developed by law enforcement, researchers, and technology companies.

Effectively combating the distribution of CSAM requires the collective efforts of law enforcement, national and international agencies, non-profit organizations, technology companies, and researchers. The first most notable example of this collaboration was the development of PhotoDNA by Microsoft Research and Dr. Hany Farid at Dartmouth College in 2009. PhotoDNA relies on hash values, but can compensate for slight modifications, including resizing and color alterations (Microsoft, 2009). Subsequently, search engines, such as Google and Bing, have implemented CSAM blockades and filtering algorithms to curb keyword searches (Steel, 2015). Facebook, in conjunction with the NCMEC, has applied artificial intelligence (AI) and machine learning (ML) to detect child nudity and new CSAM (Davis, 2018), as well as recognize child grooming techniques (Dave, 2018). Meanwhile, convolutional neural networks, a type of ML, has also been used to analyze CSAI and classify the presence of nudity and children, and categories of CSAM (Dalins, Tyshetskiy, et al., 2018). While the implementation of AI and ML has been beneficial, they are not without their limitations and mistakes. These have included incorrectly blocking artwork (Trotter, 2016) and Holocaust education (Ehrenkranz, 2018) on Facebook because of nudity, and recommending graphic and/or child abuse content on Instagram (Hamilton, 2018). As a result, it is important that additional techniques and identification criteria be developed that can be used in conjunction with existing tools to improve accuracy and decrease error.

1.2. Efforts to disguise illicit online activity

It is logical to assume that people who engage in illicit or illegal activities online would go to great lengths to avoid detection. However, this is not actually the case, especially when it comes to those disseminating CSAM. Early analysis of those arrested for possessing CSAM revealed that only 20% used any type of method to hide their content on their computer (Wolak et al., 2005). More recently, Krone et al. (2017) found that 54% of Australians arrested for CSAM possession also did not use any concealment efforts, with 26% using inconspicuous directory names and 25% hiding the media after viewing. There has been an increase in the use of encryption by offenders, but the majority of this is the result of built-in encryption from the software/application (e.g., Whatsapp or Android) rather than something consciously done by offenders (see Steel et al., 2020 for summary). Despite increases to the use of subterfuge tactics by offenders, a large percentage still operate with minimal protection. As a result, there is little reason to believe that those operating websites on the Surface Web employ detection avoidance tactics at a greater rate than individual offenders (who are apprehended). However, this is not certain as there has been few studies on how CSAM websites operate.

The research conducted on CSAM websites has demonstrated that most do very little to visually hide their intention (Westlake et al., 2017). Further work on this topic by Westlake and Bouchard (2016a, 2016b) revealed that CSAM websites with a lot of content, up to a ceiling, and connections to other CSAM websites were more likely to persist than those with little content and few connections. That is, websites that attempted to minimize the amount of CSAM and did not integrate into the larger CSAM community were more likely to be a) detected or b) inactive/abandoned/closed. In addition, websites used keywords known to be associated with CSAM, such as PTHC (preteen hardcore), extensively. These findings point to website operators' motivations being more focused on promotion than

detection avoidance, and that doing so appears to increase, rather than decrease, longevity.

There does appear to be an exception to CSAM websites not employing avoidance tactics. One tactic that has been evidenced is the use of digital pathways (Internet Watch Foundation, 2018). This is the process whereby users with specific links can access illicit content. Those who attempt to access it without the link are brought to legal content. While this is important it is worth noting that this countermeasure was found on commercial websites and not publicly accessible websites, and that its use was on the decline. If the data provided by the Internet Watch Foundation (IWF) is the exception and not the norm, it might be possible to use structural and naming conventions as additional criteria for detecting CSAM on the Surface Web.

1.3. Current study

This study explores the structure of websites on the Surface Web that are hosting and/or displaying CSAI to determine if there are specific patterns that can be used to improve the identification of suspected child sexual exploitation media. While this methodology has been applied to peer-2-peer networks (e.g., Al Nabki et al., 2020; Panchenko et al., 2012, 2013; Peersman et al., 2016), it has yet to be applied to websites nor to the naming of folders containing CSAI. By analyzing the structural patterns of websites disseminating CSAI, and how, if at all, websites disguise the locations of images, we can potentially identify new detection criteria. These criteria can then be added to existing criterion that automated tools can incorporate into their searches for CSAM, which can result in a greater outcome of positive matches and potentially locating non-hashed CSAM. Lastly, understanding how hosting hubs and high contributors structure their content on disseminating websites is critical to being able to target the most threatening websites.

2. Method

This research utilized a custom-designed web crawler to collect and categorize CSAI found on public websites on the Surface Web. This web crawler operated like many others in that it scanned a website's underlying code and extracted information about the content type and location. Data is collected based on user-defined criteria, such as keywords and specific media types, as well as limits placed on the tool (e.g., scanning or excluding specific websites). The criterion for this research was the presence of at least one CSAI hash value. The hash database used for this research was provided by the Royal Canadian Mounted Police and contained more than 2.1 million MD5 CSAI hashes.

2.1. Sample

The web crawler began on a set of 10 ‘seed’ websites known to distribute CSAM, based on previous research and law enforcement information. Ten seed websites were chosen to diversify the websites visited. That is, by starting on 10 different seeds we attempted to control for websites that were connected to each other and displaying images from the same hosting sources. The web crawler then followed hyperlinks to seed-adjoining websites and scanned for CSAI hash values. The 10 networks were combined into one dataset, and overlapping websites were collapsed. This left a final dataset of 103 websites and 8108 CSA hash values. From these 77 unique CSA hash values were identified, as many of the hash values were repeated across websites. All 103 websites analyzed in this study were freely accessible, with none requiring registration to view content. Websites were a combination of blog-based and generic website-based, with all but a few being primary domains. That is, the websites were not sub-domains (e.g., www.sub-domain.domain.com). Finally, websites were predominantly boy-focused and used explicit domain names such as ‘allcuteboys’ and ‘freeteenboys’.

2.2. Procedure

To determine if websites distributing CSAI employed structural detection avoidance tactics or countermeasures, we collected details on the physical location for each hash value. This included server folder locations and URLs for both the website displaying the image and the website hosting the image, which may or may not be the same. The number of folder sub-levels, along with the names of folders and images, were used to demonstrate efforts to hide or organize CSAI. The sub-folder level is significant because it signals the ease of access to people viewing the images. If a website stores its CSAI in the root directory that means that it is accessible on the homepage (i.e., www.website.com). As the number of subfolders increase, the ease of access can also decrease. For example, if a CSAI is located at the second sub-folder level, that means the URL for that image is ‘www.website.com/folder1/folder2/image.jpg’. This potentially means that when someone goes to the website, they need to click on another hyperlink/icon to get to the CSAI (i.e., it is not readily available upon visiting the website).

To portray the sub-folder level of images more accurately, folder names commonly associated with software and server infrastructure were excluded (htm, content, engine, img, uploads, crtr, dtr, r, at3, atx, rt, cgi, rcj, scj, tag, tp). Folder names such as pictures, photos, year, channels, category, archives, video, gallery, blog, and pages were included. How displaying and hosting websites utilized sub-folders and named their folders and files were then compared.

3. Results

3.1. Traits of displaying websites

Many websites identified made little effort to conceal the nature of their content (Table 1). Of the 82 websites displaying CSAI, 50

(61%) had them located in their root folder (i.e., no sub-folder), accounting for 2219 (27%) CSAI hash values. Thirteen (16%) websites did demonstrate efforts to possibly conceal their purpose by displaying images at the fourth sub-folder level, or lower, but this accounted for only 515 (6%) hashes. It was more likely that sub-folders were used to organize content rather than conceal it. Eighteen (22%) websites used some type of date (typically year/month) folder structure, accounting for 4345 (54%) hashes. More importantly, 39 different, unique, hash values were stored in this sub-folder group, more than any other. These findings suggest that many CSA-related websites displaying images do not fear detection, but that a minority do take steps to hide their purpose.

3.2. Specialization

Not all websites displaying CSAI were the websites hosting them. Of the 103 websites that either hosted or displayed images only 21 (20%) did both. This accounted for just 1876 (23%) of the 8108 hash values. On the other hand, 21 websites were involved in *only* hosting CSAI, while 61 websites were involved in *only* displaying images. Displaying websites often distributed the same CSAI on multiple web pages across their website (see Table 1, 'Unique Display Locations of Hash Values'). Moreover, it was common for multiple displaying websites to obtain their CSAI from the same hosting website. That is, there was clear evidence of hosting 'hubs'. Therefore, there appears to be a degree of specialization within online CSAM displaying and hosting, with only slight overlap between the two.

3.3. Traits of hosting websites

It is easy to assume that websites hosting CSAI would take measures to remain inconspicuous, such as using multiple sub-folders. However, we found that most did not use this tactic (Table 2). Instead, 24 of the 42 hosting websites (57%) were focused on organizing their content, with over two-thirds (51 of 77) of unique hash values being sorted into a year/month structure. While we cannot say definitively whether the year/month structure was created by the website user or software, neither answer points to concealment efforts.

Although 21 of our 103 websites hosted and displayed CSAI, and another 21 only hosted CSAI, neither group appeared focused on promoting their media through minimal sub-folders. Only one website hosted images in their root folder, while another four hosted in the first sub-folder. Together, these accounted for 17% of unique hash values. Of these five websites, only one was also displaying images. Yet like displaying websites, there appeared to be a small minority of hosting websites that did make efforts to conceal their media. Of the 77 unique hash value, six (8%) were fourth-level sub-folder or lower, with none of the four websites hosting these hashes also displaying them. Our findings on hosting websites reveals that most are concentrated on organizing their images more than they are concerned about concealing it.

3.4. Concealment through file and folder naming

Concealment can take two forms. The first is through hiding information in multiple layers of folders. The second is through the naming of folders and files. A manual examination of image names was used to draw conclusions on how well, or if at all, they tried to conceal their material (Table 3). Because websites hosting CSAI carry more risk than displaying websites, it is unsurprising that hosting websites went to greater lengths to mask their content through innocuous file naming. Hosting websites were most likely to use alphanumeric file names, with 48 of 77 (62%) unique hash values fitting this pattern. They also disguised their intention for 19 hash values (25%) with irrelevant names, such as "friendforyouflowerblank.gif", "maybe.jpg", and "miley-cyrus-pink-bikini-pretty.jpg", or unclear names, such as "VICTOR-01-006" and "AARONB12". When looking closer at the 50 unique folder locations on hosting websites, none were explicitly titled, demonstrating that hosting websites are very keen on avoiding the use of explicit keywords to indicate their content. The most common folder names were data formatted (34), with few (5) being alphanumeric. Most importantly, 17 were found in image-named folders, such as "pictures", "photos", "gallery", and "banner". Finally, it is worth noting that seven of these 17 (41%) were organized *within* the image-named folders by year/month sub-folder formats. To summarize, hosting websites are likely to use alphanumeric file names in directories that are then organized into folders by year/month data format naming practices and/or image-based names.

Comparatively, displaying websites less often attempted to falsely label their CSAI folders, using explicit names such as "a-boy-wanking-his-nice-uncut-cock" and "hot-boy-with-a-rock-hard-big-cock-shows-how-he-comes-for-you". Only 2% of hashes were disguised using an irrelevant folder titles (e.g., "muscles") and roughly 4% used an alphanumeric title. Like hosting websites,

Table 1
Location of folders containing CSAI on displaying websites.

	Number of websites	Unique hash values	Unique display locations of hash values
No sub-folder	50	21	2219
One sub-folder	24	24	161
Two sub-folders	17	23	649
Date format	18	39	4345
Three sub-folders	13	22	219
Four (or more) sub-folders	13	22	515

Table 2
Location of folders containing CSAI on hosting websites.

	Number of websites	Unique hash values	Unique hash value locations
No sub-folder	1	4	1
One sub-folder	4	9	4
Two sub-folders	9	5	4
Date format	24	51	34
Three sub-folders	1	2	1
Four (or more) sub-folders	4	6	6
Total	42	77	50

Table 3
File and folder naming practices of websites hosting and displaying CSAI.

	Alphanumeric	Disguised	Explicit	Data format	Root folder	Image folder
File naming (hosting website)	48	19	10	–	–	–
Folder naming (hosting website)	5	–	–	34	1	17
Folder naming (displaying website)	293	161	960	4382	1814	498

displaying websites favored organizing their content by a year/month format, with 54% following this pattern. While less prevalent on displaying websites, image-based folder names were more frequent (6%) than alphanumeric and disguised. These findings indicate that folder naming is not complex and that most displayed CSAI will often be indicated by a year/month structure and easily available for viewing.

3.5. Quantity versus quality

There appeared to be two different types of hosting websites. The first were websites who hosted a small number of CSAI but were hyperlinked by many third-party websites. For example, Website A had three hash values that were displayed by others 2647 times, while Website B had two hash values that were displayed 2688. The second were websites who hosted a (relatively) large number of images but were minimally linked. For example, Website C had nine hash values that were displayed only 28 times and Website D had eight hash values hyperlinked 92 times. This may be an artificial finding that has been created by the parameters of our initial web crawl (i.e., we do not have *all* websites displaying a website's image), but it may also point to the importance of networking, accessibility, and navigability.

There were four websites that hosted more than seven unique CSAI hash values. For three, the primary naming pattern was alphanumeric. For the fourth, 'miley-cyrus' was used as a cover. While all four websites used some method of naming concealment, each still organized their images by a type of date format, with three being Year/Month. For websites that hosted few images, but were displayed many times, no specific naming pattern was recognized.

4. Discussion

The abundance of CSAM available on the Internet has necessitated the use of automated analysis techniques for detection. These have traditionally relied on keywords and hash value databases. The limitations of these two methods have led to the development of tools that integrate other methods, including machine learning. However, most of these focus on using the attributes of images. In this study, we argued that the structural attributes of the distributor (i.e., the website) should also be used. Specifically, we found patterns in the locations on a website where CSAI were most likely to appear and patterns in the folder and file naming practices. These findings provide the opportunity for expanding and modifying detection techniques to meet the desired outcome, the ability for ESP to implement criteria to improve CSAM detection on their own websites, and the potential to increase the identification of non-hashed CSAM.

The primary aim of websites displaying CSAI is to garner viewers and distribute the media. Therefore, they make few attempts, visually, to disguise their intention (Westlake et al., 2017). Like individuals arrested for possession of CSAI (Krone et al., 2017; Steel et al., 2020; Wolak et al., 2005), our study revealed that few websites use structural strategies (e.g., multiple sub-folders or innocuous names) to hide their images. Because CSA-related websites employ few detection countermeasures, it means that techniques developed to detect CSAM do not need to be sophisticated. When it comes to the locations of images, prioritizing the root -used to maximize visibility to visit- and date format -used to organize images- directories would improve search efficiency. For displaying websites, scanning for folder and file names associated with children (e.g., boy and girl), describing the body (e.g., naked and cock), and/or sexual actions (e.g., wanking and cums) is most likely to target CSAI. For hosting websites, scanning for certain length (e.g., 10 characters) non-descript alphanumeric folder and file names (e.g., -h6F_9F-PxO) is the best strategy. Each of these strategies could be further refined and improved through machine learning training and evolved over time as needed.

Our key finding of the prevalence of year/month folder structure may be partly attributed to a website's default settings. However, we argue that this provides valuable information for detection tools as it signals two things. First, that distributors do not modify the

default settings of the website. Again, this provides support that detection tools do not need to be overly sophisticated to improve CSAM detection. Second, that detection tools can be tailored towards specific software constructions. For example, if WordPress is used, the default location for images is likely the best place to focus initial detection efforts. This has important implications for the role of ESP in combating CSAM distribution. Combined with folder and file naming patterns, these reinforce Dalins, Tyshetskiy, et al. (2018) assertion that detection strategies (i.e., training of automated tools) need to be tailored to the target's content and structure.

As it is impossible to completely eradicate CSAM, ESP need to play a critical role in minimizing distribution on their own platforms. Providing all websites with sophisticated tools or databases of known CSAI is not possible and can create security problems along with access to a list of known CSAI hashes. However, many ESP conduct regular scans of their platforms to identify content that violates their terms of service or some other policy. Therefore, it *is* possible to provide ESP with structural patterns to seek out, that can be integrated into their existing automated tools, to search for CSAM. These can be tailored to the software configuration and other common attributes of the ESP through natural language processing and machine learning. Used with other detection techniques, the potential locations of CSAM on a website can be narrowed down for more intense or sophisticated searches or flagged as needing to be scanned by tools that match the media to hash value databases. This functions as an effective data reduction technique that could hasten the analysis process of law enforcement and/or tools such as Project Arachnid, improving the detection and removal time frame.

When combating the distribution of CSAM there are two potential targets, the website displaying the media and the website hosting the media. Current combat strategies appear to be indiscriminate in targeting or focus on websites displaying CSAI. As we found that only 20% of websites both hosted and displayed CSAI, this strategy is likely to have little impact on overall distribution. While scanning for displaying websites is easier, as they are more likely to display images prominently and use descriptive, accurate, names for their files, Westlake and Frank (2016) and da Cunha et al. (2020) found that strategic targeting of hosts can have a greater impact on the distribution network. This is because a hosting website's CSAI can be disseminated by numerous displaying websites. This is echoed by Joffres et al. (2011) and Krone (2005) who emphasized the importance of hub disruption in combating CSAM distribution. By removing a hosting website, CSAI are, effectively, removed from all displaying websites. This can be accomplished more effectively by directly targeting hosts or by targeting the more overt displaying websites and structurally linking the displayed image back to the host.

Our study revealed two types of hosting websites. The first were hosts with a lot of images but fewer connections. The second were hosts with fewer images but a lot of connections. It is important to acknowledge here that the two types may be an artifact of our data collection. That is, we do not know every website an image is displayed. However, assuming that there are two types of hosts, a decision needs to be made about which is more important, the most images or the most connections (for a more in-depth discussion of this see Westlake et al., 2011; Westlake & Frank, 2016). Regardless of the priority, it is likely that the known hash values identified are not the only CSAI located on the hub. Using the folder prioritization and folder/file naming techniques outlined above on hubs could be useful for potentially identifying unknown hash values.

4.1. Recommendations

ESP (e.g., websites, service providers, and hosting companies) play a critical role in curbing the accessibility of CSAM. While large ESP such as Facebook and Microsoft have access to hash databases, a greater percentage of ESP do not. While efforts such as Project Arachnid, which is operated by the Canadian Centre for Child Protection (CCCP, 2018) and crawls reported URLs and compares the media displayed to a database, are instrumental, they cannot be relied upon to analyze everything. Providing ESP with structural criteria they can search for to detect potential CSAM on their own services, without having to visually inspect it or having access to hash databases, would be beneficial.

The resources necessary to implement a structural analysis detection system would not be overly complex, especially on smaller websites. Many publicly accessible web crawling tools exist, while partnerships could be made with law enforcement agencies or organizations such as NCMEC and CCCP, to provide these tools and continually update them with the latest structural trends they are witnessing. While this might seem like a substantial endeavor for law enforcement and non-profit organizations, it is likely not. If they make the tools easily accessible (to the 'correct' people of course), and provide a tutorial on implementing them, they would, likely, only have to do this once and each website could take it upon themselves to monitor their own websites. Moreover, it would reduce the amount of work they would be doing to monitor all these websites themselves. That is, they transfer the initial searching workload to the individual ESP.

There are three structural criteria that ESP and others can implement into automated crawling tools to improve detection of CSAI (Table 4). If the intention is to identify websites hosting CSAI, we recommend first targeting folders that have a year/month structure. While some put both in one folder name, it was far more common for the year to be one folder and the month to be a sub-folder within.

Table 4
Recommendations for targeting websites hosting or displaying CSAI.

Hosting websites	Displaying websites
• Year/month (e.g., 2011/11/file.jpg) folder names • Common image folders, such as, pictures, gallery, photos, camera • Folders with known hash values • Alphanumeric file names, with a focus on numbers over letter	• Root and year/month (e.g., 2018/08/file.jpg) folder names • Common image folders, such as, pictures, gallery, photos, camera • Explicit folder names used in combination (e.g., girl, sex, cock)

This should be easy to implement as the tool would only have to search for twelve numbers (1 through 12) and "20XX". If the focus was on new material only, this search criteria could be narrowed down even further to the current (and previous) year. While it wasn't the case for all CSAI, searching for years and months within common image folder names, such as pictures, photos, uploads, etc. would be beneficial. Within each of these folders, alphanumeric image names should be targeted, with an emphasis on numbers over letters. Finally, it was common for multiple known hash values to be present in the same folder. Therefore, if the automated tool does incorporate known hash value databases, it would be worthwhile to prioritize folders where these hashes were found as it is likely other CSAI will be present.

If the intention is to identify displaying websites, a different set of criteria need to be implemented into automated crawling tools. In addition to the year/month folder format present on hosting websites, displaying websites are likely to put their content in the root directory, likely to attract viewers to the website. Different from hosting websites is that displaying websites are more overt with their folder naming conventions. Therefore, we recommend that explicit keywords be used in the search for CSAI. This additional criterion is important because it will help eliminate a lot of false positives in searches and help with data reduction. It also points to the lack of effort made to conceal the activities and, again, a focus on attracting viewers to the website.

4.2. Limitations and future research

The websites analyzed in this study were found on the Surface Web and were publicly accessible (i.e., did not require registration, payment, etc. to access). Given IWF's (2018) discovery of digital pathways, it is important that subsequent research determine whether our findings are unique to publicly accessible websites or if they also apply to commercial and closed/private CSA websites on the Surface Web. Additionally, we have seen a significant transition of CSAM distribution to the Dark Web in recent years (Guitton, 2013; Kaur & Randhawa, 2020; Liggett et al., 2020). While the attempts by websites in our study to disguise the nature of their images was minimal, it is unknown whether websites on the more secure Dark Web use more or fewer countermeasures. The assumption is fewer, however, law enforcement activity on the Dark Web has increased, the images are often more severe (Hennessey, 2017), and the motives of distributors are different than those on the Surface Web (Dalins, Wilson, & Carman, 2018). Therefore, replication of this study on the Dark Web would be beneficial for customizing detection criteria to the Surface Web and Dark Web.

Along with websites on the Surface Web, this study focused on known CSAI hash values. We are unable to conclude if our structural and naming findings would extend to unknown CSAI hash values. Mind you, website operators are unlikely to know which hash values are known and which are unknown. Therefore, it is unlikely that they would take different measures to hide unknown hash values. Then again, some hashes might remain unknown *because* of effective countermeasures taken by website operators. As such, it is important that future research determine if our patterns hold for unknown CSAI hash values. Likewise, with the growth of CSA videos, it is important to determine if different detection avoidance tactics are used for the different types of media. Of course, these two needs are impacted by the limits placed on researchers, often by laws, in accessing and identifying this media.

5. Conclusion

The combat of CSAM distribution online is a never-ending battle that cannot be won but can be minimized. The growing abundance of media proliferation has led to an expansion of the combat ownness beyond law enforcement, to researchers, non-profit organizations, technology companies, and ESP. While this has improved the tools being developed to detect CSAM, and increased reporting, existing techniques still have many limitations. In this research, we found patterns in how publicly accessible, Surface Web, websites disseminating CSAI organized their content. This provides another criterion that can be used by automated detection software to increase CSAM identification and address some of the existing limitations. This is especially true for ESP that want to scan for potential CSAM but do not have known-media hash value databases. One of the long-time limitations of CSAM detection has been the inability to identify previously unknown hashes (i.e., new media that enters the online marketplace). Our findings could point to a criterion that when combined with other criteria may increase the possibility to identify new, previously un-hashed, CSAI and videos. Regardless to the criteria used to identify CSAM, our findings reinforce the need to customize strategies to the end goal (e.g., maximum media detection or connections) and the type of website (e.g., hosting or displaying website, and type of software used). If combating the distribution of CSAM continues to be seen as a problem that needs to be addressed by all, and we continue to develop software, techniques, and criteria that can be used by all, then we may begin to decrease the amount of CSAM being disseminated online.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Al Nabki, M. W., Fidalgo, E., Alegre, E., & Alaíz-Rodríguez, R. (2020). File name classification approach to identify child sexual abuse. In *Proceedings of the 9th international conference on pattern recognition applications and methods – Volume 1* (pp. 228–234).

Bursztein, E., Bright, T., Clarke, E., DeLaune, M., Elifff, D. M., Hsu, N., … Thomas, K. (2019, May). Rethinking the detection of child sexual abuse imagery on the internet. *The World Wide Web Conference*, 2601–2607. https://doi.org/10.1145/3308558.3313482

Canadian Center for Child Protection. (2018). Project Arachnid. https://projectarachnid.ca/en/.

da Cunha, B. R., MacCarron, P., Passold, J. F., dos Santos, L. W., Oliveira, K. A., & Gleeson, J. P. (2020). Assessing police topological efficiency in a major sting operation on the dark web. *Scientific Reports, 10*, Article 73. https://doi.org/10.1038/s41598-019-56704-4

Dalins, J., Tyshetskiy, Y., Wilson, C., Carman, M. J., & Boudry, D. (2018a). Laying foundations for effective machine learning in law enforcement. Majura–A labelling schema for child exploitation materials. *Digital Investigation, 26*, 40–54.

Dalins, J., Wilson, C., & Carman, M. (2018b). Criminal motivation on the dark web: A categorization model for law enforcement. *Digital Investigation, 24*, 62–71. https://doi.org/10.1016/j.diin.2017.12.003

Dance, G. J. X., & Keller, M. H. (2020 February 20). Tech companies detect a surge in online videos of child sexual abuse. New York Times. https://www.nytimes.com/2020/02/07/us/online-child-sexual-abuse.html.

Dave, P. (2018 October 24). Facebook removes 8.7 million sexual photos of kids in last three months. Reuters. https://www.reuters.com/article/us-facebook-child-safety/facebook-unveils-systems-for-catching-child-nudity-grooming-of-children-idUSKCN1MY1SE.

Davis, A. (2018 October 24). New technology to fight child exploitation. Facebook. https://about.fb.com/news/2018/10/fighting-child-exploitation.

Ehrenkranz, M. (2018 August 30). Facebook sorry for deleting Holocaust education post over photo of nude, starving Jewish children. Gizmodo. https://gizmodo.com/facebook-sorry-for-deleting-holocaust-education-post-ov-1828714854.

Guitton, C. (2013). A review of the available content on Tor hidden services: The case against further development. *Computers in Human Behavior, 29*(6), 2805–2815. https://doi.org/10.1016/j.chb.2013.07.031

Hamilton, I. A. (2018 September 20). Instagram's new TV service recommended videos of potential child abuse. *Business Insider*. https://www.businessinsider.com/instagram-tv-service-igtv-recommended-potential-child-abuse-2018-9.

Hennessey, S. (2017). *The elephant in the room: Addressing child exploitation and going dark*. Hoover Inst.

Internet Watch Foundation. (2018). 2018 annual report. IWF. https://www.iwf.org.uk/report/2018-annual-report.

Joffres, K., Bouchard, M., Frank, R., & Westlake, B. G. (2011). In *Strategies to disrupt online child pornography networks* (pp. 163–170). IEEE. https://doi.org/10.1109/EISIC.2011.32.

Kaur, S., & Randhawa, S. (2020). Dark web: A web of crimes. *Wireless Personal Communications, 112*, 2131–2158. https://doi.org/10.1007/s11277-020-07143-2

Krone, T. (2005). International police operations against online child pornography. *Trends and Issues in Crime and Criminal Justice, 296*(April), 1–6.

Krone, T., Smith, R. G., Cartwright, J., Hutchings, A., Tomison, A., & Napier, S. (2017). *Online child sexual exploitation offenders: A study of Australian law enforcement data Criminology Research Grants (CRG 58/12-13)*. Australian Institute of Criminology. https://www.aic.gov.au/crg/reports/crg-5812-13.

Lee, H., Ermakova, T., Ververis, V., & Fabian, B. (2020). Detecting child sexual abuse material: A comprehensive survey. *Forensic Science International: Digital Investigation, 34*, Article 301022. https://doi.org/10.1016/j.fsidi.2020.301022

Liggett, R., Lee, J. R., Roddy, A. L., & Wallin, M. A. (2020). The dark web as a platform for crime: An exploration of illicit drug, firearm, CSAM, and cybercrime markets. In T. Holt, & A. Bossler (Eds.), *The Palgrave handbook of international cybercrime and cyberdeviance* (pp. 91–116). Palgrave Macmillian. https://doi.org/10.1007/978-3-319-78440-3_17.

Microsoft. (2009 December 15). New technology fights child porn by tracking its "PhotoDNA". *Microsoft*. https://news.microsoft.com/2009/12/15/new-technology-fights-child-porn-by-tracking-its-photodna.

National Center for Missing and Exploited Children. (2020). 2019 reports by electronic service providers (ESP). *Missing Kids*. https://www.missingkids.org/content/dam/missingkids/gethelp/2019-reports-by-esp.pdf.

Panchenko, A., Beaufort, R., & Fairon, C. (2012). Detection of child sexual abuse media on p2p networks: Normalization and classification of associated filenames. In *Proceedings of the LREC workshop on language resources for public security applications* (p. 27e31).

Panchenko, A., Beaufort, R., Naets, H., & Fairon, C. (2013). Towards detection of child sexual abuse media: Categorization of the associated filenames. In *European conference on information retrieval* (pp. 776–779). Berlin, Heidelberg: Springer. , March.

Peersman, C., Schulze, C., Rashid, A., Brennan, M., & Fischer, C. (2016). iCOP: Live forensics to reveal previously unknown criminal media on P2P networks. *Digital Investigation, 18*, 50–64.

Ramirez, G. (2015 July 28). MD5: The broken algorithm. Avira. https://www.avira.com/en/blog/md5-the-broken-algorithm.

Schulze, C., Henter, D., Borth, D., & Dengel, A. (April, 2014). Automatic detection of CSA media by multi-modal feature fusion for law enforcement support. In *Proceedings of international conference on multimedia retrieval e ICMR '14, Glasgow, United Kingdom*, Article 353e360. https://doi.org/10.1145/2578726.2578772

Steel, C. M. (2015). Web-based child pornography: The global impact of deterrence efforts and its consumption on mobile platforms. *Child Abuse & Neglect, 44*, 150–158. https://doi.org/10.1016/j.chiabu.2014.12.009

Steel, C. M., Newman, E., O'Rourke, S., & Quayle, E. (2020). An integrative review of historical technology and countermeasure usage trends in online child sexual exploitation material offenders. *Forensic Science International: Digital Investigation, 33*, Article 300971.

Trotter, J. K. (2016 September 9). Facebook admits Pulitzer-winning photograph is not child pornography. Gizmodo. https://gizmodo.com/facebook-admits-pulitzer-winning-photograph-is-not-chil-1786441732.

Westlake, B. G. (2020). The past, present, and future of online child sexual exploitation: Summarizing the evolution of production, distribution, and detection. In T. Holt, & A. Bossler (Eds.), *The Palgrave handbook of international cybercrime and cyberdeviance* (pp. 1225–1253). Palgrave Macmillian. https://doi.org/10.1007/978-3-319-78440-3_52.

Westlake, B. G., & Bouchard, M. (2016a). Criminal careers in cyberspace: Examining website failure within child exploitation networks. *Justice Quarterly, 33*(7), 1154–1181. https://doi.org/10.1080/07418825.2015.1046393

Westlake, B. G., & Bouchard, M. (2016b). Liking and hyperlinking: Examining reciprocity and diversity in online child exploitation network communities. *Social Science Research, 59*(September), 23–36. https://doi.org/10.1016/j.ssresearch.2016.04.010

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

Westlake, B. G., Bouchard, M., & Girodat, A. (2017). How obvious is it: The content of child sexual exploitation websites. *Deviant Behavior, 38*(3), 282–293. https://doi.org/10.1080/01639625.2016.1197001

Westlake, B. G., & Frank, R. (2016). Seeing the forest through the trees: Identifying key players in online child sexual exploitation distribution networks. In T. Holt (Ed.), *Cybercrime through an interdisciplinary lens* (pp. 189–209). Routledge. https://doi.org/10.4324/9781315618456.

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