12-1-2023

Using file and folder naming and structuring to improve automated detection of child sexual abuse images on the Dark Web

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

Enrique Guerra
Alumni

Recommended Citation

Bryce Westlake and Enrique Guerra. "Using file and folder naming and structuring to improve automated detection of child sexual abuse images on the Dark Web" *Forensic Science International: Digital Investigation* (2023). https://doi.org/10.1016/j.fsidi.2023.301620

Forensic Science International: Digital Investigation 47 (2023) 301620

Contents lists available at ScienceDirect

Forensic Science International: Digital Investigation

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

Using file and folder naming and structuring to improve automated detection of child sexual abuse images on the Dark Web

Bryce Westlake [*], Enrique Guerra

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

ARTICLE INFO

Keywords:
Child sexual abuse material
Hash values
Dark web
Automated detection
Child sexual exploitation

ABSTRACT

Increasing dissemination of child sexual abuse material (CSAM), especially on the Dark Web, has necessitated greater reliance on automated detection tools. These tools typically match images and videos to known CSAM databases, which is an ineffective method for identifying unknown CSAM. To identify potential complimentary methods, we analysed 162 unique known images, displayed 7289 times on 988 Dark Web websites, to determine if patterns in file/folder naming and structuring tendencies existed on websites. Overall, websites prioritised organisation (ease of access) over obfuscation (security) and hosted almost all images they displayed. File/folder names were commonly alphanumeric, however, there was evidence of sequence file naming patterns. Webpages displaying CSAM were explicitly named, often using underage and/or incest-related keywords. Structuring patterns revealed presence of website copies (mirrors) which can impede effective CSAM removal. Recommendations for supplementing automated detection techniques are discussed.

1. Introduction

The prevalence of child sexual abuse material (CSAM), which predominantly encompasses images and videos, being disseminated online is difficult to accurately quantify (Dabrowska, 2021). What we do know is that between 1998 and 2017, the National Center for Missing and Exploited Children (NCMEC) saw a median year-to-year growth in report volume of 51% (Bursztein et al., 2019). In 2021 alone, this resulted in more than 29.3 million reports to NCMEC (2022a). Of these reports, 29.1 million (99.3%) were from electronic service providers (ESPs), such as Facebook and TikTok, who had found CSAM being disseminated on their platforms (NCMEC, 2022b). Additionally, the Internet Watch Foundation's (IWF, 2023) 2022 annual report revealed that while the Surface/Clear Web (e.g., .com, .ru, .me) remains the predominant CSAM distribution channel, there is significant growth in distribution on the Dark Web/Net, through The Onion Router (Tor; .onion).

The amount of CSAM being distributed online, coupled with the traumatic impact on investigators (Mitchell et al., 2022; Redmond et al., 2023; Seigfried-Spellar, 2018; Strickland et al., 2023), has necessitated the utilisation of automated detection tools. Implemented by ESPs (e.g., Apple, 2021; Buckman, 2021), as well as non-governmental organisations (NGOs) such as NetClean and Thorn, and law enforcement (Lee et al., 2020), these tools typically scan all images and videos on a website and match them to existing databases of known CSAM. Although automation speeds up the detection process, considerable manual investigation is still required as the technology is often ineffective at detecting previously unknown CSAM (Dabrowska, 2021). Therefore, additional methods need to be developed that can assist with detecting and identifying CSAM online.

Reports on CSAM dissemination suggest that offenders who operate websites may employ ritualistic behaviours (i.e., patterns) that can be exploited by automated tools to improve detection techniques. First, examining uniform resource locators (URLs), International Association of Internet Hotline (INHOPE, 2021) found that the same image or video is disseminated (i.e., visually displayed) numerous times, often hosted by different websites. Second, Davis (2021) reported that between October and November 2020, six videos account for nearly half of the CSAM reports on Facebook. Third, distributors often trade links to CSAM in plain sight, using coded language to identify content (Solon, 2020). When combined, these suggest that there may be utility in determining if certain patterns can be identified in 1) the file and folder naming practices of websites hosting CSAM, 2) the locations of folders containing CSAM, 3) how websites display and share CSAM, and 4) whether there are structural similarities across hosting websites that may point to a collaboration/link between said websites. If these patterns are

[*] Corresponding author. Department of Justice Studies, San Jose State University, One Washington Square, San Jose, CA, 95192, USA.
E-mail address: Bryce.Westlake@sjsu.edu (B. Westlake).

https://doi.org/10.1016/j.fsidi.2023.301620
Received 8 February 2023; Received in revised form 15 July 2023; Accepted 26 August 2023
Available online 4 September 2023

detected, they could provide an important complementary method to existing techniques to, at the very least, narrow down the files and locations on websites that are most likely to contain previously unknown CSAM and to ensure that all websites containing a given image, or video, are removed simultaneously. To investigate these questions, a secondary analysis of the textual output from an automated detection tool's scan of Dark Web websites for known CSAM, specifically images, was conducted. The textual output consisted of the image's hash value, hosting and displaying websites, and the image location (i.e., server folder where the image was stored, the file name of the image, and the webpage URL where it was displayed).

2. Literature review

2.1. Distribution of CSAM on Dark Web

The Internet is often described as consisting of three layers – the Surface Web, the Deep Web, and the Dark Web. The Deep Web is a component of the Internet that is 'hidden', which is mostly comprised of databases and intranets that are not indexed by search engines, typically do not have static URLs, and often require specific credentials to access (He et al., 2007). The Dark Web, also referred to as Darknet, is the name given to the subsection of the Deep Web that while used for legitimate security and safety reasons (e.g., journalism and activism), is more commonly associated with illicit activity (see Kaur and Randhawa, 2020 for overview). These 'websites' are accessible only through specialised software, with one of the most common being Tor, which utilises encryption and a high degree of anonymity (Spalević and Ilić, 2017). As a result, people use it to openly buy, sell, trade, and freely disseminate drugs, stolen identities, credit card numbers, firearms, malware, hacking information, movies, television shows, music, and CSAM (see Liggett et al., 2020 for overview). The inherent encryption and anonymity of the Dark Web poses considerable challenges for law enforcement and NGOs in identifying and apprehending those disseminating CSAM (Raven et al., 2021; Spitters et al., 2014).

The mediums for distributing CSAM have evolved over time. Throughout the late 1990's and early 2000's, bulletin board systems, UseNet, internet chat relay (IRC), and peer-to-peer (P2P) networks were some of the most popular (Steel et al., 2020). Today, the Dark Web and mobile messaging systems (MMS) are growing in prevalence (Stroebel and Jeleniewski, 2015; Steel et al., 2022), especially during COVID (Europol, 2020; Interpol, 2020). In 2023, NetClean found that nearly seven out of 10 information technology professionals reported the Dark Web making it easier for employees to access CSAM at the workplace. Although there is no concrete way to measure the amount of CSAM being disseminated on the Dark Web, there is evidence that it is substantial and growing. The IWF (2022) reported a 27% increase in CSAM Dark Web services from 2020 to 2021, while Dalins, Wilson et al. (2018) b found that 1.75% of Dark Web websites crawled offered CSAM. Among these websites, only 28% were requiring people to purchase the material, while 33% were openly distributing on forums and 52% on file sharing platforms.

The propensity to freely disseminate CSAM among a community of users rather than commercially profit potentially comes from offenders' desire to garner support and camaraderie with other producers, consumers, and distributors for their interests in CSAM (Holt et al., 2010; Leclerc et al., 2021; Kloess and van der Bruggen, 2021). This process also aids in ensuring they stay relevant with the latest distribution locations (e.g., websites, social media, forums), search terms (e.g., codewords), and content (Fortin et al., 2018). Most importantly, by participating in this process, and making their content easy to access, they garner higher status within the overall virtual community. While this poses challenges for combat, this also presents an opportunity for improving detection techniques. To freely distribute CSAM effectively requires that those seeking it can easily search for and find it. This translates to websites and users likely implementing few/minimal security techniques and countermeasures, along with common language and terminology to describe content. Therefore, it is possible that patterns in the words used in file and folder names, folder locations, and webpage URLs, could be another method for detecting CSAM.

2.2. Methods to identify CSAM

The proliferation of CSAM online has necessitated the development of automated tools and procedures for detection and analysis of suspected media. This is primarily accomplished using databases of known CSAM hash values. A hash value is a hexadecimal code that serves as a digital fingerprint for any file (Ramirez, 2015). When a file is modified in any fashion (e.g., adding a word to a document, cropping an image, cutting the length of a video), a new, unique, hash value is created. In many jurisdictions, these databases are integrated into automated web crawler tools, to permit the active scanning, detection, and eventual takedown of CSAM online. Four well-known NGO web crawlers are I 'See' Child Abuse Material by INHOPE, Project Arachnid by Canadian Centre for Child Protection's CyberTip.ca, ProActive by NetClean, and IntelliGrade by IWF.

Hash values have clear advantages, especially when it comes to processing large amounts of media files in short periods of time. However, 'hard' hashing also comes with some limitations. For example, their effectiveness is reliant on the scanned media file's hash value being present within the used database, making this approach easy to circumvent by CSAM disseminators (Farid, 2021). Fuzzy (perceptual) hashing-based tools, such as PhotoDNA (Microsoft, 2009), have attempted to address these limitations, allowing for media files with slight alterations, such as resizing, cropping, watermarking, to be matched. However, this approach does not account for larger changes, and still relies on the original (i.e., matching media file) to be present in the database. As the majority of CSAM reported each year is previously unknown/unreported (Bursztein et al., 2019; INHOPE, 2021), relying solely on hash value databases for identification will miss a large percentage of CSAM files. These limitations point to the need for other methods, or at the very least methods that can complement hash values.

One potential complimentary method to hash values is using attributes of the files themselves (Epstein et al., 2020). While relying on common keywords alone can lead to issues with false positives, when combined with hash values they have been shown to be effective (Frank et al., 2010). This practice has been implemented successfully for P2P networks (Le Grand et al., 2009; Panchenko et al., 2012; Panchenko et al., 2013). This is because P2P networks rely on people being able to find what they are searching for, in this case CSAM, based on keyword searches. File name and pathway classification has also been demonstrated to have utility when law enforcement agencies seize a computer and want to filter CSAM from non-CSAM (Al-Nabiki et al., 2020a, b; Pereira et al., 2020). One potential reason for its utility in this context is Steel et al. (2021) found that 78% of offenders do not separate their personal collection into different sub-folders (i.e., CSAM was all located in one directory) and 42% attempted to collect all pictures of a victim, or within a series, resulting in repetitive file naming practices (e.g., 'tina1', 'tina2', 'tina3'). Given that keyword searches and file names/pathways have been demonstrated to work in certain contexts – P2P networks and personal hard drives respectively – it is possible that these, and/or similar, techniques can be used to identify CSAM 'out in the wild' (i.e., websites online). While this approach has been tested to some degree on the Surface Web (Guerra and Westlake, 2021), it has yet to be applied to the Dark Web. This is an important distinction as characteristics of Tor (e.g., anonymity and private keys) may lead to different storing and displaying patterns emerging than on personal devices and/or Surface Web. As patterns found in one domain may not translate to the others, this could necessitate a different set of criteria to be used by automated tools when scanning the Dark Web for CSAM.

2.3. Use of countermeasures by CSAM offenders

While the use of countermeasures by CSAM offenders have grown over the past decade, these practices are often rudimentary and used by a minority of offenders (Balfe, et al., 2015). For example, Krone et al. (2017) found that 54% of offenders used no method of concealment for their personal collection (i.e., CSAM found on their computer), with only 26% using inconspicuously named directories and 8% using encryption. Steel et al. (2022) similarly found that 28% mislabelled directories and 18% encrypted individual files. When it came to online activity specifically, Krone et al. learned that less than half used any kind of security during online communications, while Steel et al. reported 68% deleted their web browsing history and only 38% used some type of built-in privacy mode when browsing. In fact, Briggs et al. (2011) found that CSAM offenders often revealed identifying information online such as birthdates, occupations, and name. As a result, it has been suggested that increases in countermeasures, specifically encryption, are the result of built-in software/application settings rather than consciously taken detection avoidance measures by offenders (Steel et al., 2020).

As child abuse and high-risk sex offenders are often creatures of habit (Chamberlain et al., 2020; Elliott et al., 1995; Gies et al., 2016; Grove and Farrell, 2010), we hypothesise that the lack of implementing countermeasures on personal devices is likely to be translated to the websites they use to disseminate CSAM. This hypothesis appears to be supported by the literature, as previous research has found that CSAM distribution platforms do little to hide their content (Latapy et al., 2013; Panchenko et al., 2012; Peersman et al., 2016; Westlake et al., 2017). Specifically pertaining to file and folder structuring patterns, Guerra and Westlake (2021) found that websites hosting CSAM on the Surface Web were likely to locate their files in root directories and focus more on organising rather than hiding CSAM. Moreover, websites displaying CSAM were likely to use explicit file naming practices rather than disguised names. This behaviour could be more pronounced on the Dark Web, where perceived anonymity and prioritisation of free distribution may override any fears of detection and apprehension. If this is indeed true, patterns found in file and folder characteristics could be integrated into automated search tools and combined with hashing techniques to better target and prioritise locations most likely to contain CSAM. In doing so, the efficiency of detection can be improved, which may lead to quicker removal and thus decreased redistribution of new CSAM.

3. Material and methods

3.1. Data collection

Data were collected by researchers external to the authors of this paper, using an automated web crawler, designed specifically for operation on the Dark Web (see Monk et al., 2018). Known as The Dark Crawler (TDC), this tool integrated a law enforcement hash value database, comprising 2.1 million MD5 image hash values. Data collection began with 30 'seed' Dark Web websites identified in previous research as distributing CSAM. On each of these websites, TDC scanned media on each webpage and checked the images against the hash value database. Once the 30 seed websites were analysed, TDC followed hyperlinks to adjoining websites and scanned them. In essence, the hyperlinks served as a form of snowball sampling. If no known hash values were identified, the website was categorised as irrelevant and excluded from any further analysis. This process continued until every hyperlinked website was checked and no new links were identified. At the conclusion of the data collection, the authors of this research were provided only the textual output for analysis.

Institutional research ethics was sought prior to receiving the data file, given the sensitive nature of the topic, however, the ethics board deemed it not required as the data was collected by a third-party (i.e., secondary data) and the analyses were on the textual output only from the data collection. To avoid unnecessary exposure to traumatising material, along with any potential legal issues, the accuracy of the classification of hash values as being, or not being, child sexual abuse related was not verified. As a result, all unknown hash values were excluded from analysis, to be more certain that the data analysed contained only positive identifications of CSAM. This resulted in a sample size of 179 unique image hash values, hosted on 1246 websites, and displayed 14,479 times on 1277 different websites. Given the size of the Dark Web is unknown, and a small number of CSAM images are often displayed hundreds and even thousands of times in multiple locations (e.g., Bursztein et al., 2019; Davis, 2021; INHOPE, 2021), we cannot evaluate whether this sample size is representative of a significant proportion of CSAM dissemination on the Dark Web (see 'Limitations and Future Research' for further discussion).

3.2. The datasets

The sample was divided into four datasets, with each being analysed separately. These four datasets were: main database, non-onion host, hash value #1, and hash value #2 (Table 1). The main database consisted of only websites that had a onion address. Websites from domains such as com were separated into a separate dataset called 'non-onion host', as they are not part of the Dark Web. This dataset comprised 15 hash values and 27 hosting websites. Two hash values were also separated from the main database and analysed as separate datasets called 'hash value #1' and 'hash value #2'. Hash value #1 appeared 4591 times across 197 websites, accounting for nearly one-third of all image hash values displayed, while hash value #2 appeared 1493 times across 13 websites, accounting for just over 10% of all displayed images. No other image was displayed more than 523 times and only eight appeared more than 300 times. Analysing these two hash values separately ensured that they did not adversely skew the results of this research. Excluding the non-onion hosts and the two frequently displayed hashes, the main database was left comprising of 162 unique image hashes hosted by 988 Dark Web websites and displayed 7289 times.

3.3. Data analysis

For each file, qualities/features regarding the location, structure, and naming were analysed. Three manual examinations were conducted for each file identified. First, the physical location on the hosting server was examined to determine the sub-folder 'level'. Therefore, a file located at '/image.jpg' was classified as being at the zero sub-folder level, while a file located at '/folder1/folder2/image.jp' was classified as being at the second sub-folder level. To accurately determine the level of sub-folder, folder names commonly associated with software and server infrastructure were excluded (e.g., htm, wp-content, engine, storage, tag). For example, 'htm/image.jpg' was assigned to the zero sub-folder level and not the one sub-folder level. Folder names such as archives, file, gallery, hthumbs, image, img, photos, pictures, video, *thumbs, and year were included. Second, where an image is physically stored on a server is not necessarily, and often isn't, where a user will view the image. Instead, the location on the server is 'linked' to the web page displaying the image. Therefore, we also analysed the URL (e.g., 'http://website.onion/youngboy/naked) for common attributes, such as references to CSAM in the title. Third, we conducted a qualitative analysis of

Table 1
Characteristics of the four subsets of final database.

Subset	Unique Hashes	# Of Host Websites	Displaying Websites	Times Displayed
Non-onion host	15	27	58	1106
Hash value #1	1	197	197	4591
Hash value #2	1	34	34	1493
Main database	162	988	988	7289

the folder and file names used to host (server) and display (URL) CSAM.

4. Results

4.1. File location practices for hosting websites

At which sub-folder level a website hosts a file, along with the folder and file naming practices (discussed more below), can give an indication of the degree of effort taken to hide CSAM. Table 2 summarises the number of hashes and how often each was displayed for each folder level – zero (main/root directory) to four – and two commonly found folder names, image and date. Image refers to folders found at any sub-folder level that had the name 'image', 'images', or 'img'. Date refers to the use of a year (e.g., 2019) alone or with a month (e.g., 2019-05) as the folder name.

Websites appeared more focused on organisation of images than detection avoidance, using some sub-folders but only a few. While 28 of 988 websites (2.8%) did locate their images in the main directory (zero), the majority used one (63.7%) or two (33.7%) sub-folders. Using multiple sub-folders was not very common, with only three websites using four and none using more than four. Furthering the argument for organisation over obfuscation, 475 (48.1%) website operators appeared to favour default folders used by the website's software, with 109 of 162 (67.3%) hash values being in an 'image' sub-folder, and many of these being at the one sub-folder level. Although less prevalent, date folder names were also common.

To focus website scans, it can be beneficial to know whether CSAM is likely to be hosted in one or multiple folder location. If CSAM is always found in the same sub-folder on a website, then automated tools just need to scan that sub-folder. However, if multiple sub-folders contain CSAM, then tools need to complete more extensive searches. While some Dark Web websites hosted child sexual abuse images at multiple sub-folder levels, the majority were located at one sub-folder level and typically within one sub-folder (Table 2, # of folders containing images). When multiple sub-folders were used, it was common for the same hash values to be found within multiple folders (Table 2, # of images files found).

Finally, where CSAM is physically located (hosted) on a website may be different than where they are displayed. For example, an image in multiple sub-folder locations could be displayed on the same webpage, on multiple webpages on the same hosting website, or webpages on a different website. In this study, Dark Web websites appear to be insular, focusing on displaying the same images in multiple locations on their website (Table 2, # of times displayed). Additionally, it was very rare for an image to be externally displayed (i.e., on another website). Of the 7289 times the 162 unique hash values were displayed, only 65 times (0.9%) were images displayed beyond the hosting website (Table 2, # of externally displayed). Images found at the zero sub-folder level were the most likely to be displayed externally, but it is worth noting that all eight images that were externally displayed, were located on one website and thus likely an outlier rather than a pattern. The next most common was sub-folder level three, with 8% externally displayed. One exception to this finding was the 'date' folder. Here, 36% were externally displayed.

4.2. File naming practices for hosting websites

After examining the sub-folder level, location, and external displaying of images, the file names were qualitatively analysed to determine if any patterns were present. Instances of the same file (i.e., same file name and folder structure) were removed to ensure that each file name only appeared once in the analysis. This resulted in a final dataset of 197 unique file names, which were then placed into one of five categories (Table 3). Files in the alphanumeric category were often just numeric (e.g., 133594739) but some were alphanumeric (e.g., 541nazd5397dxc311). Files in the CSA words category were known terms, such as 'pthc' (preteen hardcore), as well as explicitly named files, such as 'cpporn' or similar. Files in the disguised category included names such as 'triplesomersault', 'apple-touch', and 'simple-smile'. Files in sex words category included terms such as 'sex', 'pussy', 'fuck', and 'slut'. Finally, the other category were names that presented some feature that did not allow them to be accurately categorised in one of the other four groups.

Likely complicating automated detection is that most file names were alphanumeric (44%) or used disguised words (30%). However, keywords related to CSAM were used to some frequency (8%), which highlights that they can still be useful when combined with other detection methods. Finally, the category other comprised nearly 13% of image names. It is possible that this category included files that referenced terminology unknown to the research team, such as 'kidbin', or evidence of a photoset of specific children (e.g., ' … set6 … ', ' … m-20-la', and 'potv03'). If it indeed contained code words, or child identification abbreviations, unfamiliar to the researchers but known by investigators, these could prove useful for automated searches to those with specific knowledge of current lingo.

It was not uncommon for hashes located in the same folder to be labelled, for example, 'v1', 'v2', 'v5', 'v8'. In this example, it is likely that the folder also contained files labelled 'v3', v4′, 'v6', and 'v7'. While we cannot say definitively, as we did not view any of the images nor visit the websites, it is possible that these point to a 'set', or series, of images that were either of the same person or of the same genre.

4.3. URL naming practices of displaying websites

In comparison to file names on hosting servers, the URL (webpage

Table 3
Filing naming patterns for child sexual abuse image hash values.

Category	Total Number	Percentage
Alphanumeric	87	44.2
CSA Words	16	8.1
Disguised	59	29.9
Sex Words	10	5.1
Other	25	12.7

Table 2
Folder location of unique hash values.

Folder Level	Unique Hashes[a]	# Of Host Websites[a]	# Of Folders Containing CSA Images	# Of Image Files Found	# Of Times Displayed	# Of Times Externally Displayed
Zero	9	28	28	30	54	8
One	92	629	670	2836	5733	3
Two	29	333	338	442	803	0
Three	28	14	17	39	647	54
Four	16	3	7	16	52	0
Image[b]	109	475	481	1440	3708	1
Date[b]	25	19	23	37	150	54

[a] Several hash values appeared at multiple sub-folder levels, on multiple websites, hence why totals add up to more than 162 hash values and 988 websites.
[b] Counts from these two folder levels are included in the zero, one, two, three, and four folder levels.

name) where images were displayed did reveal some important characteristics that could prove beneficial for identification. For this analysis, the displaying URL was placed into one of five categories (Table 4). The first category was incest, which included some reference to family, parent/child, and in most cases the word 'incest' explicitly. The second category was home page, which included images displayed on the websites' main page, as identified through locations such as '.index.php' and '.index.html'. The third category was sexual, which was any URL that made a reference to sex and/or pornography but did not make it specifically clear that it was referencing underage material. The terms 'teen' and 'young' were included in this category, as they do not explicitly mean underage. The fourth category was underage, which was any reference to children, such as 'jailbait', 'child', and 'pedo'. For almost all the URLs in this category, the word 'underage' was present. The fifth category was other, which included all URLs that did not fit into one of the other categories. URLs with multiple hash values present on the same webpage were included only once. This resulted in 1384 unique URLs being analysed.

Further demonstrating the priority of website operators to making their content accessible to consumers rather than avoiding detection, the URLs displaying CSAM were more explicit than the file naming practices. Nearly 70% were categorised as underage (40.3%) or incest (29.6%), while 3.5% displayed the image on their home page. Only 13.4% were categorised as other. Finally, many of the websites displayed the same image on multiple webpages. In some cases, more than 100 times.

4.4. Finding copies of websites using folder structures

Instances of the same folder structure and file naming practices across multiple websites were common. For example, Website A would have an image at 'websitea/CSAI/childsex.jpg' while Website B would have the exact same image at 'websiteb/CSAI/childsex.jpg'. Beyond folder and file naming structure, it was common for mirrors to have similar. onion prefixes. For example, Website A would be 'pedohubaaaaaaaa.onion' while Website B would be 'pedohubbbbbbbbb.onion'. The suffixes appear to be autogenerated, but do point to efforts by distributors to ensure their content remains accessible. However, there were situations in which the prefixes were very different. For example, '272qg4i6t6 … ' and 'alhd4z4crn … ' were the prefixes for two onion websites. Put another way, looking at the domain names there was no way to tell that they were copies of each other, but by looking at the folder structuring, it became evident.

4.5. Hash value #1 and #2, and non-onion hosts

In addition to the main database, three separate databases were also analysed – hash value #1, hash value #2, and non-Onion hosts. The frequency of hash value #1 and #2 appeared to be examples of mirrored websites. Both hash values #1 and #2 were hosted on many websites, but only two folder structures were identified for each hash, so four in total. For both hashes, one folder structure corresponded to a series of. onion addresses with the same prefix. However, the other folder structure for each hash corresponded to very different. onion prefixes (e.g., 'oqqsn7y ….onion' and 'ehids45 ….onion'). This reinforces the finding that without identifying the website's structure, it would be difficult to determine that the websites were related (i.e., mirrors).

Table 4
Displaying URL naming patterns for child sexual abuse hash values.

Category	Total Number	Percentage
Incest	409	29.6
Home Page	48	3.5
Sexual	184	13.3
Underage	558	40.3
Other	185	13.4

While known image hash values displayed on the Dark Web are almost exclusively hosted on the Dark Web, there is still evidence of some displayed images being located on the Surface Web. In our analysis, 15 hashes were displayed 1106 times, across 58 Dark Web websites, from 27 websites on the Surface Web. What was most apparent with these hashes was that they were hosted on file-hosting (i.e., cyberlockers) websites, with one being particularly prevalent. This points to the need to improve detection tactics by content-hosting services on the Surface Web. However, with exception of one hosted file, displayed twice, all file names were alphanumeric, meaning that detecting them using anything beyond hash value identification is challenging.

5. Discussion

The proliferation of CSAM online continues to grow, with the Dark Web quickly becoming a platform of preference for dissemination. In this study we sought to determine whether file and folder labelling and structuring behaviours could be identified on websites hosting/displaying CSAM on the Dark Web. The lack of concealment found in this study aligns with existing research of other CSAM distribution platforms (e.g., Panchenko et al., 2012; Latapy et al., 2013; Peersman et al., 2016; Westlake et al., 2017) and of perpetrators' personal devices (e.g., Krone et al., 2017; Steel et al., 2021, 2022). Furthermore, it appears that file name and pathway classification research done on seized computers (Al-Nabiki et al., 2020a, b; Pereira et al., 2020) and on the Surface Web (Guerra and Westlake, 2021) does translate to the Dark Web. This is likely because the architecture of the Internet was developed to inherently focus on data access rather than data security (Fielding and Taylor, 2002). However, there were some ways in which distribution (and storage) on the Dark Web appeared to differ from other domains. These lead to four recommendations for how automated data collection tools used by ESPs, NGOs, and law enforcement can be complemented with the inclusion of file/folder naming and structuring techniques. Additionally, two important challenges that differentiate Dark Web and Surface Web child sexual abuse image practices are highlighted.

5.1. Improving automated detection tools

First, this research demonstrated how websites on the Dark Web reflect some of the structural organisational habits of offenders found by Steel et al. (2021) examining their personal devices and Guerra and Westlake (2021) examining websites they operate on the Surface Web. Images were often found in default directories, such as 'images', 'gallery', and 'uploads' and rarely buried in multiple sub-folders. This means that automated tools could prioritise default image and video storage locations, for the website software being used, and ignore non-default locations, such as system directories, with minimal impact on detection. Additionally, tools could ignore locations that went beyond a certain sub-folder level (e.g., four). Because default directories were the most used by website administrators, multiple image hash values were often located in the same directory. This means that once an automated tool finds a known hash value, it could prioritise scanning all other files in the same location. Additionally, if unknown or non-hashed CSAM were located on a website, they would likely be found in the same folder as the known image hash value. This can help narrow down files to examine, either automatically or manually, for new images and videos. If this were to be implemented, it is worth acknowledging that there is the potential for offenders to modify their behaviours to counter this detection technique long-term. However, the lack of concealment efforts demonstrated by previous research (Guerra and Westlake, 2021; Latapy et al., 2013; Panchenko et al., 2012; Peersman et al., 2016; Westlake et al., 2017), combined with the findings in this study that websites still use long-known CSAM-related keywords, such as 'pthc', suggest that this approach would still be effective on a significant number of websites.

Second, when multiple locations on one website were found to contain child sexual abuse images, there was often overlap in the hash values found in each location. Given that the same hash value(s) is/are likely to be present multiple times on a website, once an automated tool identifies a known hash value, it could first attempt to *hard* match all other image files on the website to that hash value, or *fuzzy* match similar image files using perceptual hashing tools such as PhotoDNA. Put another way, instead of attempting to match any image subsequently detected on the website to all images in a database, it could instead be matched to that singular image hash value using various hashing techniques, as it is more likely to be present elsewhere on the website. This could be a quick way of identifying a second, third, or fourth folder location that may contain additional known, or previously unknown, images, or even videos.

Third, website operators on the Dark Web seem more likely to name CSAM files, and folders, less explicitly, often using alphanumeric names. This makes identifying them automatically more challenging, as there are no 'hints' that a file may be CSAM, simply by the name of the file or folder in which it is located. Some folder and file names were explicit, suggesting that keywords could be integrated into automated tools, to improve detection. While this could lead to a significant number of false positives, and thus unnecessary manual verification, Frank et al. (2010) demonstrated how complementing this approach with other techniques, such as file hashing, can reduce that problem. In comparison to websites hosting images, those displaying often used very explicit keywords in the URL. This is likely because website operators want their content to be easily found by consumers as it would provide them with prestige and notoriety within the online CSAM community (Fortin et al., 2018). Used in conjunction with other detection techniques, scanning for relevant keywords in the displaying location rather than the file (hosting) location may help identify images that are not explicitly named. Once the displaying website was identified, working backwards, the hosting location, likely the same website, could be determined.

Finally, the tendency to use sequential numbering of files (e.g., sally-1 to sally-8) may signify a set, or series, of related images. This finding may be unsurprising given that offenders demonstrate tendencies towards collecting all pictures of specific children, or within a series (Steel et al., 2021). Because of this preference, disseminating a set or series of images likely garners the offender prestige and notoriety within the CSAM community. This points to an important area for further study, as integrating this into automated detection criteria could be beneficial for identifying previously unknown, or non-hashed, CSAM. For example, if an automated tool was programmed so that if sally-1 and sally-5 are known hash values, and files sally-2, sally-3, sally-4, and sally-6 exist, these additional images are flagged for manual review or scanned using other tools, such as nudity detection, age estimation, and so forth, as potentially new, or non-hashed, CSAM. Implementing these strategies into automated tools could expedite detection and subsequent removal of CSAM on websites, thereby possibly reducing dissemination and further victimisation of the child. For investigators, identifying this new CSAM quicker may provide valuable information that could lead to the rescuing of a child currently being victimised.

5.2. Challenges of Dark Web compared to Surface Web

As additional CSAM distribution mediums arise, such as the Dark web, it is important to determine whether previously identified patterns apply to the new medium or whether new patterns need to be identified and exploited. Dalins et al. (2018)a, b argued that automated detection strategies need to be tailored to content and structure of the domain in which they are being mobilised. While offenders employ some similar file and folder naming and structuring patterns, as has been found with personal devices and the Surface Web, we identified two important differences present on the Dark Web, which create new challenges for detection.

First, websites on the Surface Web disseminating CSAM are interconnected, often relying on hosting 'hubs' (Guerra and Westlake, 2021). That is, CSAM on the Surface Web is hosted by a smaller number of websites, typically file hosting services, cyberlockers, and image stores, to which other websites link (IWF, 2022; INHOPE, 2021). This practice allows websites to not physically possess/host media files and disseminate through free websites that have limits on the amount of data they can store. From the website operator's perspective, this potentially eliminates some of the fear of detection, possession, and/or apprehension. In contrast, Tor websites in general are often not visibly linked with each other (Zulkarnine et al., 2016). As more than 99% of the images analysed in this study were hosted and displayed by the same website, it appears that CSAM websites on the Dark Web align closer to dissimilar (i.e., non-CSAM) Dark Web websites than to similar CSAM Surface Web websites. Although offenders on the Dark Web have the same interpersonal goals as offenders on the Surface Web – validation, support, acceptance, and prestige – the insular nature of Dark Web presents a challenge for detecting CSAM. On the Surface Web, additional websites can be identified by examining the hosting information of an image being displayed externally; however, this strategy is unlikely to be effective on the Dark Web. Therefore, public sharing of. onion address and finding rare situations in which a website hyperlinks to another website are likely the most effective strategy for initial detection. This will likely require specific. onion addresses to be input into automated tools rather than allowing a web crawler to freely scan the Dark Web and find interconnected websites.

Second, it is a simple process once a website is shut down for the website owner to start a new website, with all the same content. Further simplifying this process is implementation of mirrors. Website mirroring is when a website owner creates multiple copies of their website, with the exact same file and folder structure. These mirrors, or copies, may be updated manually or automatically synchronised – when a new file is created on Website A, the system copies it to Website B, C, etc. This action can be beneficial as if one of the websites is detected and shutdown, a replica is already operating, with the exact same content, to take its place.

The practice of mirroring has been noted as a challenge for combating illegal streaming and piracy (Ibosiola et al., 2018) and appears to play a central role in CSAM dissemination on the Dark Web. The prevalence of mirrors creates a major challenge for automated identification and effective removal, as well as estimating the overall prevalence of distribution on the Dark Web. Identifying mirrors is imperative to effective detection as unless all mirrors are targeted simultaneously, the removal of one mirror, or image on one mirror, will have no impact on the overall distribution. The current research demonstrated that some mirrors use similar prefixes, making them easier to detect, but others use dissimilar prefixes. However, when. onion prefixes are different, knowing the folder and file structure can be beneficial as it can help automated tools to first identify mirrors and then identify the specific locations of CSAM on the mirrors. Combined, this information can be integrated into automated tools to speed up detection (i.e., telling it to look for a specific folder and find a specific file) and once all mirrors are detected, can be used to more effectively remove CSAM (i.e., all mirrors can be targeted at once to ensure complete removal). The information can also be used by ESPs, NGOs, and law enforcement to prioritise certain websites (e.g., a website with many mirrors), servers hosting many mirrors, and potential key producers and/or disseminators of CSAM, who operate multiple websites. The application of this approach has been demonstrated with illegal gambling (Yang et al., 2019) and CSAM (Westlake and Frank, 2017) websites.

5.3. Limitations and Future Research

The analyses conducted in the current research relied on the textual output of an automated crawl of websites located on the Dark Web found to be hosting and/or displaying known image hash values. Given the potential legal risks and psychological harms, we were unable to complete a visual analysis of the images included and excluded from our analysis. As a result, we are unable to determine whether a) images excluded from the analysis were CSAM, and b) patterns found were unique to known hash values only, to all child sexual abuse images (including unknown), or common across all images (CSAM or non-CSAM) disseminated on each website included in the analysis and the Dark Web more generally. However, we argue that while known image hash values were used to reach conclusions, the patterns identified are also applicable to unknown image hash values. This is because those disseminating images are likely unaware which hash values are known and unknown, and therefore unlikely to implement different structural strategies for distributing either. Nevertheless, additional research needs to be completed with agencies who can visually analyse CSAM, in this case images, to accurately assess them for meeting the definition of CSAM. This may be best accomplished through a thorough analysis of a website seized by law enforcement, where the file structure and all corresponding website posts are able to be processed.

Although the current research appears to align with what is known about CSAM distribution on the Surface Web, extended to the Dark Web, the insular nature of Tor makes it hard to accurately conceptualise how representative results are to the overall problem online. This is especially true given the, relatively, small sample analysed within the current study – 179 unique hash values, displayed 14,479 times and hosted on 1246 websites. Therefore, several large-scale studies of websites hosting and displaying CSAM on the Dark Web, beginning from different 'seed' websites, need to be conducted. Coupled with the utilisation of a larger hash value database, such as Interpol's *International Child Sexual Exploitation* database, results could then be compared to those from this study, to determine if each identify similar patterns. This includes a comparison of websites freely distributing images, such as those examined in this study, and commercial enterprises, as the organisational and displaying attributes may be different. Central to this will be implementing the identified patterns into automated tools and calculating the false positive rate. While the four recommendations provided above could prove beneficial for automated searches, if they lead to high rates of false positives, then their utility could be minimal. However, if combined with other machine learning and artificial intelligence tools, such as Griffeye Brain (Griffeye, 2023), image/video file structure authentication (Epstein et al., 2020), and/or voice and facial recognition (Westlake et al., 2022), false positives and negatives could be reduced. By comparing the patterns found across multiple studies and then testing them, their effectiveness in a practical setting can be better determined.

6. Conclusion

The online proliferation of CSAM, primarily in the format of images and videos, continues to increase unabated. The Dark Web is only one of the latest mediums in which this content is being heavily disseminated. As the number of media continues to increase, it has become unable to be investigated manually. Advancements in automated technology have provided an opportunity to process this media more efficiently and with less psychological harm to investigators. However, even these tools can struggle to keep up with the amount of content appearing daily. Moreover, these tools are only as effective as the rules which govern their operation. That is, if the tools rely on known hash values for detection, they will remain ineffective for finding unknown, or new, CSAM. Therefore, based on an analysis of image hash values on the Dark Web, four additional guidelines to better target automated tools were provided, that if implemented could increase their utility, especially in finding previously unknown image hash values. In this analysis, two key differences from the Surface Web were highlighted – Dark Web websites are insular/disconnected from other websites and operate multiple mirrors of the same website. These point to the need for slightly different strategies for automated scanning of Dark Web and Surface Web CSAM websites. From this research, steps can be taken to combine techniques to improve detection efforts and as a result more quickly identify children currently being abused and rescue them.

Funding

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

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgements

Thanks to Dr. Richard Frank at Simon Fraser University's School of Criminology for providing data for this analysis.

References

Al Nabki, M.W., Fidalgo, E., Alegre, E., Alaíz-Rodríguez, R., 2020a. File name classification approach to identify child sexual abuse. In: Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods. ICPRAM 2020, pp. 228–234. https://doi.org/10.5220/0009154802280234.

Al-Nabki, M.W., Fidalgo, E., Alegre, E., Alaiz-Rodríguez, R., 2020b. Short Text Classification Approach to Identify Child Sexual Exploitation Material. https://doi.org/10.48550/arXiv.2011.01113. *arXiv: 2011.01113*.

Apple, 2021. CSAM Detection Technical Summary. https://www.apple.com/child-safety/pdf/CSAM_Detection_Technical_Summary.pdf.

Balfe, M., Gallagher, B., Masson, H., Balfe, S., Brugha, R., Hackett, S., 2015. Internet child sex offenders' concerns about online security and their use of identity protection technologies: a review. Child Abuse Rev. 24 (6), 427–439.

Briggs, P., Simon, W.T., Simonsen, S., 2011. An exploratory study of internet-initiated sexual offenses and the chat room sex offender: Has the internet enabled a new typology of sex offender? Sexual Abuse 23 (1), 72–91.

Buckman, I., 2021. Hashing it out: how an automated crackdown on child pornography is shaping the fourth amendment. Berkeley Journal of Criminal Law. https://www.bjcl.org/blog/hashing-it-out-how-an-automated-crackdown-on-child-pornography-is-shaping-the-fourth-amendment.

Bursztein, E., Clarke, E., DeLaune, M., Elifff, D.M., Hsu, N., Olson, L., et al., 2019. Rethinking the detection of child sexual abuse imagery on the internet. In: The World Wide Web Conference, pp. 2601–2607. https://doi.org/10.1145/3308558.3313482.

Chamberlain, A.W., Smith, S.M., Turner, S.F., Jannetta, J., 2020. Global positioning system monitoring of high-risk sex offenders: implementation challenges and lessons learned. Crim. Justice Pol. Rev. 31 (9), 1259–1285. https://doi.org/10.1177/0887403419884723.

Dabrowska, M., 2021. The unclear picture of child sexual abuse material (CSAM) online volumes during the COVID-19 pandemic. Bialystok Legal Studies 26 (6), 109–125. https://doi.org/10.15290/bsp.2021.26.06.07.

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. Digit. Invest. 26, 40–54.

Dalins, J., Wilson, C., Carman, M., 2018b. Criminal motivation on the dark web: a categorisation model for law enforcement. Digit. Invest. 24, 62–71. https://doi.org/10.1016/j.diin.2017.12.003.

Davis, A., 2021. Preventing Child Exploitation on Our Apps. *Meta Newsroom*. https://about.fb.com/news/2021/02/preventing-child-exploitation-on-our-apps.

Elliott, M., Browne, K., Kilcoyne, J., 1995. Child sexual abuse prevention: what offenders tell us. Child Abuse Negl. 19 (5), 579–594. https://doi.org/10.1016/0145-2134(95)00017-3.

Epstein, B., Bruehs, W., Lyons, B., Fischer, D., 2020. Digital video source authentication: groundbreaking insights into digital video evidence. Forensic Focus. https://www.forensicfocus.com/articles/digital-video-source-authentication-groundbreaking-insights-into-digital-video-evidence.

Europol, 2020. Exploiting Isolation: Offenders and Victims of Online Child Sexual Abuse during the COVID-19 Pandemic. https://www.europol.europa.eu/cms/sites/default/files/documents/europol_covid_report-cse_jun2020v.3_0.pdf.

Farid, H., 2021. An overview of perceptual hashing. Journal of Online Trust & Safety 1 (1). https://doi.org/10.54501/jots.v1i1.24.

Fielding, R., Taylor, R., 2002. Principled design of the modern web architecture. ACM Trans. Internet Technol. 2 (2), 115–150.

Frank, R., Westlake, B.G., Bouchard, M., 2010. The structure and content of online child exploitation. In: ISI-KDD '10 ACM SIGKDD Workshop on Intelligence and Security Informatics. https://doi.org/10.1145/1938606.1938609. Article 3).

Fortin, F., Paquette, S., Dupont, B., 2018. From online to offline sexual offending: episodes and obstacles. Aggress. Violent Behav. 39, 33–41.

Gies, S., Gainey, R., Healy, E., 2016. Monitoring high-risk sex offenders with GPS. Crim. Justice Stud. Crit. J. Crime Law Soc. 29 (1), 1–20. https://doi.org/10.1080/1478601X.1129088.

Griffeye, 2023. Griffeye Brain: More Brainpower, More Intelligence. https://www.griffeye.com/griffeye-brain/.

Grove, L., Farrell, G., 2010. Repeat victimisation. In: Fisher, B.S., Lab, S.P. (Eds.), Encyclopedia of Victimology and Crime Prevention, vol. 2. Sage Publications, pp. 767–769. https://doi.org/10.4135/9781412979993.n258.

Guerra, E., Westlake, B.G., 2021. Detecting child sexual abuse images: traits of child sexual exploitation hosting and displaying websites. Child Abuse Negl. 122, 105336 https://doi.org/10.1016/j.chiabu.2021.105336.

He, B., Patel, M., Zhang, Z., Chang, K.C.C., 2007. Accessing the deep web. Commun. ACM 50 (5), 94–101. https://doi.org/10.1145/1230819.1241670.

Holt, T.J., Blevins, K.R., Burkert, N., 2010. Considering the pedophile subculture online. Sex. Abuse 22 (1), 3–24.

Ibosiola, D., Steer, B., Garcia-Recuero, A., Stringhini, G., Uhlig, S., Tyson, G., 2018. Movie pirates of the Caribbean: exploring illegal streaming cyberlockers. In: Twelfth International AAAI Conference on Web and Social Media.

INHOPE, 2021. Annual Report 2020. https://inhope.org/media/pages/the-facts/download-our-whitepapers/c16bc4d839-1620144551/inhope-annual-report-2020.pdf.

Internet Watch Foundation, 2022. The Annual Report 2021: Dark Web Reports. https://annualreport2021.iwf.org.uk/trends/darkwebreports.

Internet Watch Foundation, 2023. Annual Report 2022: Domain Analysis. https://annualreport2022.iwf.org.uk/trends-and-data/domain-analysis/.

Interpol, 2020. Threats and Trends Child Sexual Exploitation and Abuse: COVID-19 Impact. https://www.interpol.int/content/download/15611/file/COVID19%20-%20Child%20Sexual%20Exploitation%20and%20Abuse%20threats%20and%20trends.pdf.

Kaur, S., Randhawa, S., 2020. Dark web: a web of crimes. Wireless Pers. Commun. 112 (4), 2131–2158. https://doi.org/10.1007/s11277-020-07143-2.

Kloess, J.A., van der Bruggen, M., 2021. Trust and relationship development among users in dark web child sexual exploitation and abuse networks: a literature review from a psychological and criminological perspective. Trauma Violence Abuse, 15248380211057274. https://doi.org/10.1177/15248380211057274.

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. http://www.crg.aic.gov.au/reports/1617/58-1213-FinalReport.pdf.

Latapy, M., Magnien, C., Fournier, R., 2013. Quantifying paedophile activity in a large P2P system. Inf. Process. Manag. 49, 248–263. https://doi.org/10.1016/j.ipm.2012.02.008.

Le Grand, B., Guillaume, J., Latapy, M., Magnien, C., 2009. Dynamics of Paedophile Keywords in eDonkey Queries: Measurements and Analysis of P2P Activity against Paedophile Content Project. http://antipaedo.lip6.fr/.

Leclerc, B., Drew, J., Holt, T.J., Cale, J., Singh, S., 2021. Child sexual abuse material on the darknet: a script analysis of how offenders operate. Trends and Issues in Crime and Criminal Justice 627, 1–14.

Lee, H., Ermakova, T., Ververis, V., Fabian, B., 2020. Detecting child sexual abuse material: a comprehensive survey. Forensic Sci. Int.: Digit. Invest. 34, 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. The Palgrave Handbook of International Cybercrime and Cyberdeviance 91–116. https://doi.org/10.1007/978-3-319-78440-3_17.

Microsoft, 2009 December 15. New technology fights child porn by tracking its “PhotoDNA”. Retrieved from: https://www.microsoft.com/presspass/features/2009/dec09/12-15photodna.mspx.

Mitchell, K.J., Gewirtz-Meydan, A., O’Brien, J., Finkelhor, D., 2022. Practices and policies around wellness: insights from the internet crimes against children task force network. Front. Psychiatr. 13, 931268.

Monk, B., Mitchell, J., Frank, R., Davies, G., 2018. Uncovering Tor: an Examination of the Network Structure. Security and Communication Networks, 4231326. https://doi.org/10.1155/2018/4231326, 2018.

National Center for Missing and Exploited Children, 2022a. CyberTipline 2021 Report. https://www.missingkids.org/gethelpnow/cybertipline/cybertiplinedata.

National Center for Missing and Exploited Children, 2022b. 2021 *CyberTipline reports by electronic service providers (ESP)*. https://www.missingkids.org/content/dam/missingkids/pdfs/2021-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. Springer, Berlin, Heidelberg, pp. 776–779. https://doi.org/10.1007/978-3-642-36973-5_82.

Peersman, C., Schulze, C., Rashid, A., Brennan, M., Fischer, C., 2016. iCOP: live forensics to reveal previously unknown criminal media on P2P networks. Digit. Invest. 18, 50–64. https://doi.org/10.1016/j.diin.2016.07.002.

Pereira, M., Dodhia, R., Anderson, H., Brown, R., 2020. Metadata-based Detection of Child Sexual Abuse Material. *arXiv preprint arXiv:2010.02387*.

Ramirez, G., 2015. MD5: the Broken Algorithm. Avira. https://www.avira.com/en/blog/md5-the-broken-algorithm.

Raven, A., Akhgar, B., Abdel Samad, Y., 2021. Case studies: child sexual exploitation. In: Dark Web Investigation. Springer, Cham, pp. 249–266. https://doi.org/10.1007/978-3-030-55343-2_12.

Redmond, T., Conway, P., Bailey, S., Lee, P., Lundrigan, S., 2023. How we can protect the protectors: learning from police officers and staff involved in child sexual abuse and exploitation investigations. Front. Psychol. 14, 1152446.

Seigfried-Spellar, K.C., 2018. Assessing the psychological well-being and coping mechanisms of law enforcement investigators vs. digital forensic examiners of child pornography investigations. J. Police Crim. Psychol. 33 (3), 215–226.

Solon, O., 2020. Child Sexual Abuse Images and Online Exploitation Surge during Pandemic. NBC News. https://www.nbcnews.com/tech/tech-news/child-sexual-abuse-images-online-exploitation-surge-duringpandemic-n1190506.

Spalević, Z., Ilić, M., 2017. The use of dark web for the purpose of illegal activity spreading. Ekonomika. J. Econ. Theory Pract. Soc. Issues 63, 73–82.

Spitters, M., Verbruggen, S., Van Staalduinen, M., 2014. Towards a comprehensive insight into the thematic organization of the tor hidden services. In: 2014 IEEE Joint Intelligence and Security Informatics Conference. IEEE, pp. 220–223. https://doi.org/10.1109/JISIC.2014.40.

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 Sci. Int.: Digit. Invest. 33, 300971 https://doi.org/10.1016/j.fsidi.2020.300971.

Steel, C., Newman, E., O’Rourke, S., Quayle, E., 2021. Collecting and viewing behaviors of child sexual exploitation material offenders. Child Abuse Negl. 118, 105133 https://doi.org/10.1016/j.chiabu.2021.105133.

Steel, C., Newman, E., O’Rourke, S., Quayle, E., 2022. Technical behaviours of child sexual exploitation material offenders. Journal of Digital Forensics, Security and Law 17. https://doi.org/10.15394/jdfsl.2022.1794. Article 2.

Strickland, C., Kloess, J.A., Larkin, M., 2023. An exploration of the personal experiences of digital forensics analysts who work with child sexual abuse material on a daily basis:“you cannot unsee the darker side of life”. Front. Psychol. 14, 1142106.

Stroebel, M., Jeleniewski, S., 2015. Global Research Project: A Global Landscape of Hotlines Combating Child Sexual Abuse Material on the Internet and an Assessment of Shared Challenges. National Children’s Advocacy Center. http://hdl.handle.net/11212/4673.

Westlake, B., Brewer, R., Swearingen, T., Ross, A., Patterson, S., Michalski, D., et al., 2022. Developing automated methods to detect and match face and voice biometrics in child sexual abuse videos. Trends and Issues in Crime and Criminal Justice (648), 1–15. https://doi.org/10.52922/ti78566.

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

Westlake, B.G., Frank, R., 2017. Seeing the forest through the trees: identifying key players in online child sexual exploitation distribution networks. In: Holt, T. (Ed.), Cybercrime through an Interdisciplinary Lens. Routledge, New York, pp. 189–209. https://doi.org/10.4324/9781315618456.

Yang, H., Du, K., Zhang, Y., Hao, S., Li, Z., Liu, M., et al., 2019. Casino royale: a deep exploration of illegal online gambling. In: Proceedings of the 35th Annual Computer Security Applications Conference, pp. 500–513.

Zulkarnine, A.T., Frank, R., Monk, B., Mitchell, J., Davies, G., 2016. Surfacing collaborated networks in Dark Web to find illicit and criminal content. In: 2016 IEEE Conference on Intelligence and Security Informatics (ISI), pp. 109–114. https://doi.org/10.1109/ISI.2016.7745452.