Article

Sexual Abuse
2017, Vol. 29(7) 685–708

DOI: 10.1177/1079063215616818
journals.sagepub.com/home/sax
SAGE

Assessing the Validity of Automated Webcrawlers as Data Collection Tools to Investigate Online Child Sexual Exploitation

Bryce Westlake[1,2], Martin Bouchard[2], and Richard Frank[2]

Abstract

The distribution of child sexual exploitation (CE) material has been aided by the growth of the Internet. The graphic nature and prevalence of the material has made researching and combating difficult. Although used to study online CE distribution, automated data collection tools (e.g., webcrawlers) have yet to be shown effective at targeting only relevant data. Using CE-related image and keyword criteria, we compare networks starting from CE websites to those from similar non-CE sexuality websites and dissimilar sports websites. Our results provide evidence that (a) webcrawlers have the potential to provide valid CE data, if the appropriate criterion is selected; (b) CE distribution is still heavily image-based suggesting images as an effective criterion; (c) CE-seeded networks are more hub-based and differ from non-CE-seeded networks on several website characteristics. Recommendations for improvements to reliable criteria selection are discussed.

Keywords

webcrawler, automated data collection, child sexual exploitation, child pornography, cybercrime

Introduction

The increasing prevalence of cybercrime over the last two decades has resulted in a re-examination of existing criminological paradigms and theories, to incorporate cyber aspects or to transition entirely to cyberspace (Bossler & Burruss, 2011; Higgins

[1]San Jose State University, CA, USA
[2]Simon Fraser University, Burnaby, British Columbia, Canada

Corresponding Author:
Bryce Westlake, San Jose State University, 1 Washington Square, San Jose, CA 95192, USA.
Email: Bryce.Westlake@sjsu.edu

& Marcum, 2011; Patchin & Hinduja, 2011). One of the challenges to this transition has been that the change in environment often requires the application of different analysis methods and/or data collection techniques (e.g., Burris, Smith, & Strahm, 2000; Grabosky, Smith, & Dempsey, 2001; Holt, Blevins, & Burkert, 2010; Karpf, 2012; Layton, Watters, & Dazeley, 2011). Cyberspace provides a unique environment for data collection, but one that is vast and complex, and thus yet to be fully understood. To rectify this, social scientists have formed interdisciplinary partnerships with computer scientists to design innovative methods and tools to collect online data. Because of the abundance of data available, interdisciplinary partnerships have focused on methods for simplifying the collection process of relevant data. In most cases, this simplification has involved partly or entirely automating data gathering, through the building of custom-designed webcrawlers (e.g., Ball, 2016; Bouchard, Joffres, & Frank, 2014; Chen, 2012; Kontostathis, Edwards, & Leatherman, 2010). Although each webcrawler differs slightly, depending on the intended purpose, in general, webcrawlers scan text on webpages, or in forums or databases, and compile the data for further analysis (Kanich et al., 2011).[1] To guide and target the data collection, webcrawlers often follow user-defined criteria placed on the crawling process. Despite the growing abundance of cybercrime researchers using automated data collection tools, and developing them for police investigations (e.g., Dykstra & Sherman, 2013; Saari & Jantan, 2013), no studies have been undertaken, within criminology, with the purpose of validating their ability to discriminate between relevant and irrelevant data. Instead, it is assumed that, through the user-specified rules and use in other fields, the data collected is accurate and on-topic. However, the criminal nature of the material being collected, and the tendency for people to hide their activities, suggests that a close assessment of the data produced by webcrawlers is needed. This is especially important in the context of online child sexual exploitation (CE) where illegal content is often both accessible to the public and hidden within otherwise legitimate content. Webcrawlers can separate the wheat from the chaff, but with the benefits of automation come the potential for these tools to drift away from their intended target.

The Internet has helped create new types of crimes such as malicious software production and hacking (McGuire & Dowling, 2013). However, it has also aided in modifying existing offline crimes such as fraud and sexual exploitation. As a result of the Internet, the distribution of CE material has boomed in prevalence to a level never before seen offline (Beech, Elliott, Birgden, & Findlater, 2008). Given the graphic nature of the material, and its abundance, researchers have been strong proponents of using webcrawlers for data collection. Although the data collected are often assumed, and described as "child exploitation" in nature, because of the presence of specific keywords or known images, the ability for a webcrawler to discern between child exploitation material and other, legal, material is unclear.

In the current study, we design a webcrawler tool to collect data on websites associated with the distribution of CE material. Using the Child Exploitation Network Extractor, we construct hyperlinked networks surrounding 10 CE-seed websites from our inclusion criteria: presence of CE-related keywords selected from previous research in the field, and/or known CE images from a database provided by law

enforcement. We then compare the CE-seeded networks with those constructed starting from a similar, but legal, genre (non-CE sexuality) and a dissimilar genre (sports). We compare how frequently our CE inclusion criteria appear in non-CE genres' networks, especially sexuality, and what differentiates CE-seeded networks from non-CE-seeded networks. Through this comparison, we can provide support, or recommendations, to the currently used standard for CE identification criteria. We can then determine if the currently used criteria can allow for autonomous searching and detection of CE-related websites on the public world wide web (WWW). We begin with an overview of Internet-mediated research (IMR) methods, paying specific attention to their impact for online sexual offending research. We follow with a systematic review of the advances in methods used to investigate online sexual offending with an emphasis on CE material distribution and automated data collection tools.

Literature

The use of, and reliance on, IMR methods for analyzing various phenomena continues to grow in both primary and secondary research (Hewson, Yule, Laurent, & Vogel, 2003). Within criminology, IMR has primarily been utilized for victimization surveys, conducted among high school and college students (e.g., Bossler & Holt, 2010; Choi, 2008) and content analyses of criminal networks (e.g., Décary-Hétu & Dupont, 2012).

Hesitancy by some to fully embrace IMR have centered on concerns regarding the quality, representativeness, reliability, and validity of Internet data (Schonlau, van Soest, Kapteyn, & Couper, 2009; Shropshire, Hawdon, & Witte, 2009). More accurately, because of the clear advantages of IMR, there are fears that IMR techniques and methodologies have yet to undergo rigorous validation. To address some of these concerns, research examining the limitations of Internet data collection has grown exponentially. These studies have concluded that the quality and representativeness of Internet-mediated data are comparable with offline survey data (e.g., Chang & Krosnick, 2009; Dillman, 2007). Although there is evidence that IMR data are valid, in general, there are specific situations where the evidence is less clear. Examinations of online and offline sexual offenders has shown that the two differ on key demographics (Babchishin, Hanson, & van Zuylen, 2015; Elliott, Beech, Mandeville-Norden, & Hayes, 2009; Seto, Hanson, & Babchishin, 2011). Given the possible differences between online and offline sexual offenders, it is unclear whether the assumptions about offline sexual offending practices can be directly applied to online practices and vice versa. Therefore, the collection of online sexual offending data needs to be validated independently, and differ from other type of material found online, including legal, sex-related material. A key part of that validation process is identifying the best criteria to differentiate between websites that may potentially contain online CE and others.

Unique to the cyber-environment is the opportunity to safely observe, control, and possibly even induce, the criminal event, thus providing important information regarding the objectives and criminal processes unable to be effectively studied offline. This opportunity has facilitated the creation of fictionally vulnerable computer systems,

known as honeypots, to observe the motives and techniques used by hackers (Almutairi, Parish, & Phan, 2012; Marín, Naranjo, & Casado, 2015; Provos & Holz, 2007; Spitzner, 2003). This ability to observe and collect data unobtrusively has also led to the study of strategies, tactics, motives, and rules of cyber-communities involved in the distribution of live webcam recordings of sex (Roberts & Hunt, 2012).

Offline, non-incarcerated sexual offenders, or those in the early stages of their criminal career, are often difficult to access in part to real, or perceived, risks to safety. Online, perceptions of anonymity, by offenders, have allowed researchers access to those sexually interested in children, to survey their interests (Wurtele, Simons, & Moreno, 2014), viewing habits (Seto et al., 2015), and the personality and demographic differences between consumers and non-consumers (Ray, Kimonis, & Seto, 2014; Seigfried, Lovely, & Rogers, 2008). Anonymity has also provided the opportunity to obtain background characteristics, sexual preferences, attitudes, and motives for men seeking sexual services online (Milrod & Monto, 2012) and compare them with offline customers (Monto & Milrod, 2014), to investigate sex tourism networks (Chow-White, 2006; Evans, Forsyth, & Wooddell, 2000), and to understand the interplay between online negotiations and offline sexual risk reduction (Rice & Ross, 2014). The abundance of data available for collection on the Internet, and the desire of researchers to have as much information as possible, has led to the development of automation of data collection techniques.

While sex offenders take considerable risks conducting their illicit affairs in the public domain of the WWW, we know from offline deterrence research that the threat of a long prison sentence does not deter crime. On the Internet, the perceived, and often real, anonymity provided lead many offenders to conduct criminal activities in the public domain (Armstrong & Forde, 2003; Holt et al., 2010; Maimon, Alper, Sobesto, & Cukier, 2014). For offenders operating in the public domain, one of the easiest ways to avoid detection would be to continually modify one's moniker. However, research reveals that offenders maintain the same pseudonym throughout their "online criminal career" as it becomes associated with notoriety and respect, which outweigh the corresponding costs (Décary-Hétu & Dupont, 2012; Décary-Hétu, Morselli, & Leman-Langlois, 2012). Specific to CE, Wolak, Finkelhor, and Mitchell (2005) found that very few CE-related offenders used any security measures to hide their activities. The argument could be made that those who avoided detection were the ones who used security measures and that the knowledge of detection-avoidance tactics were significantly less in 2005. However, the current availability of CE-related material on the WWW suggests that the findings of Wolak, Finkelhor, and Mitchell still hold today. As a result, access to CE material, through public websites, remains an important area for research.

Examining the distribution of CE images, Carr (2004) identified the WWW (e.g., public websites) as the second most prominent method of acquisition. More recently, O'Halloran and Quayle (2010) conducted a qualitative study of boy love support forums and found that despite public forums being "old technology," they were still being used prominently. Similarly, Tremblay's (2006) study of boy love forums highlighted the use of public spaces for the discussion of illicit activities, such

as the distribution of CE material. Within peer-to-peer networks, researchers have highlighted the excessive distribution of CE-related material on public and semi-public networks (Fournier et al., 2014; Rutgaizer, Shavitt, Vertman, & Zilberman, 2012; Steel, 2009). Finally, an annual analysis by the Internet Watch Foundation (2014) identified "31,266 URLs contain[ing] child sexual abuse imagery" and that 77% of these URLs were located at .com, net, ru, org, and info domains. Combining these findings, not only is the public WWW a suitable domain for conducting research on illicit sexual activities, it appears to be reflective of an important population among CE offenders, and subsequent distribution techniques.

The Current Study

The distribution of CE material is conducted using a variety of online and offline methods (Callanan, Gercke, De Marco, & Dries-Ziekenheiner, 2009; Fortin & Corriveau, 2015; van Wijk, Nieuwenhuis, & Smeltink, 2009). For those researching CE, the graphic nature of material, combined with the abundance of data available, has made automated data collection tools appealing, especially for longitudinal studies (Latapy, Magnien, & Fournier, 2013; Wolak, Liberatore, & Levine, 2014). Automated data collection tools (e.g., webcrawlers) have incorporated CE-related keywords (Frank, Westlake, & Bouchard, 2010), to guide the process, and have examined peer-to-peer networks such as Gnutella (Steel, 2009), eDonkey (Fournier et al., 2014), and BitTorrent (Rutgaizer et al., 2012), and publicly accessible websites (Westlake, Bouchard, & Frank, 2011). More recently, researchers have begun to incorporate CE image databases into existing search criteria, after determining that keywords alone did not provide a full picture of the distribution network (Westlake, Bouchard, & Frank, 2012). Incorporating social network analyses, researchers have used webcrawlers to identify key players (Westlake et al., 2011) and cliques (Iqbal, Fung, & Debbabi, 2012) within CE networks and the most optimal strategies for network fragmentation (Joffres, Bouchard, Frank, & Westlake, 2011). Despite the increasing use of CE-related keywords and, more recently, CE image databases, no research has determined whether these criteria adequately distinguish between CE and non-CE-related data. To investigate the effectiveness of commonly used CE identifying criteria we (a) determine the prevalence of inclusion criteria in CE, non-CE sexuality, and sports-seeded networks; and (b) identify other website and network characteristics that distinguish CE-seeded networks from comparisons, across multiple waves, including more than a year after the start of data collection. By identifying the differences between CE-seeded networks and comparisons, we can refine and improve on the existing criteria used within the field.

Method

Webcrawler (Child Exploitation Network Extractor [CENE])

The webcrawler, referred to as CENE, designed for this study follows a similar structural and functionality design as automated data collection tools used by

search engines and researchers to index websites (Burris et al., 2000; Chau, Shiu, Chan, & Chen, 2007; Chau & Xu, 2008; Frank et al., 2010). CENE was designed to follow a method similar to that of a person browsing the Internet looking for illegal material. CENE begins at a user-specified website, analyzes the hypertext markup language (HTML)—the standard coding language used to create webpages—and collects information about the website's structure and pertinent characteristics. It then recursively follows hyperlinked text that, when clicked, transports a user to a separate webpage or website, and continues the analysis process. Like if a person was browsing, the webcrawler scans the linked webpage to determine if the website is relevant to the topic being searched. If the website does not meet the pre-defined criterion, it is discarded; as it would be by a user who views the webpage, sees that it is not what they are looking for, and closes the website.

As the Internet is substantial in size, first, data collection size-limits were placed on CENE with regard to the number of websites (approximately 300) and the total number of webpages comprising those websites (500,000). Second, a set of 14 "safe" domains (Google, Yahoo, Microsoft, Facebook, Twitter, Wikipedia, Doubleclick, Reddit, Gizmodo, Adobe, YouTube, Macromedia, w3schools, and adlog) and 1,259 websites were programmed into the webcrawler as being "off-topic." The list of off-topic websites was derived from our previous research into online CE (Frank et al., 2010; Westlake et al., 2011, 2012). Third, and most importantly, criteria associated with CE material were implemented to guide the webcrawler's search and decision-making process about the inclusion of each website scanned. For a website to be included, the hyperlinked webpage had to contain at least seven of our 82 keywords and/or one known CE image, as identified by the integrated database provided by the Royal Canadian Mounted Police (RCMP). Again, this simulated the search process of a user in requiring specific material to be present for them to remain on the website. The seven keyword requirement was determined through a manual verification process conducted during previous research (see Westlake et al., 2011), while the requirement of keywords *or* image has been identified as a better method for identifying CE-related websites than either method alone (Westlake et al., 2012). Given the limitations placed on CENE, the network of websites collected should not be viewed as exhaustive but rather a representation of what a typical user might do during their search. This also means that the data collection on each website may not be exhaustive; depending on the size of the website. However, the size is large enough to draw conclusions regarding the validity of automated data collection methods and for mapping illegal networks online.

Once the defined limitations of the data collection were met, CENE compiled the data and aggregated it to the domain level. In other words, the data on www.website.com/webpage1 and www.website.com/webpage2 were summed and listed in www.website.com. A list of all websites and webpages scanned by CENE were stored and reused during subsequent crawls. This ensured that the same webpages, if they were still online, were analyzed at each time point.

Table 1. Description of the Categories of Keywords and Hash Values Used By the Webcrawler.

	Keywords (number)	Hash values (number)
Category 1	Child Exploiter-Code (27)	Child Exploitation (618,632)
Category 2	Thematic (23)	Child Nudity (652,223)
Category 3	Sex-Oriented (32)	Collateral (981,231)

Webcrawler Criteria

To identify CE-related websites, CENE integrated a database of 2.25 million hash values, collected during RCMP investigations and used during prosecution. A hash value is a 32-hexidecimal code which functions similar to a digital fingerprint. Each computer file is given a hash value based on its binary composition. When a file is edited, even minimally, a new hash value is created. Tretyakov, Laur, Smant, Vilo, and Prins (2013) stated that the chances of two distinct files having the same hash value is "negligibly small" ($1/2^{2048}$). Last updated on June 1, 2012, for our purposes (CENE was launched July 2012), the database was divided into three categories (see Table 1). These three categories were based on Canadian legal definitions of CE material. As important differences exist internationally regarding legal definitions (see Gillespie, 2011), we quickly summarize Canadian law. Under section 163.1 (1) of the Canadian Criminal Code (1985), *child pornography* includes any "photographic, film, video, or other visual representation . . . written material . . . or audio recording" of a person under the age of 18, engaged in an explicit sexual act, or advocating sexual activity. Those depicted can include imaginary people.

The first database category (*Child Exploitation*) contained 618,632 images all of which were classified, under the Canadian Criminal Code, as being CE. The second (*Child Nudity*) contained 652,223 images that would probably be considered CE by a judge. However, images in this category were not blatant and, thus, the risk averse nature of law enforcement resulted in these images being placed into a separate category. The third (*Collateral*) contained 981,231 images that were important enough to be collected by offenders but would not be defined as CE under the Canadian Criminal Code. For example, the initial images in a photo-shoot whereby a child was still clothed and not being directly sexually exploited would be in this category. For an image to be included in a website's count, it had to be at least 150 pixels (approximately 2 in., or 4 cm) by 150 pixels.

CENE used a set of 82 keywords, selected from previous research conducted on the topic of online CE (LeGrand, Guillaume, Latapy, & Magnien, 2009; Steel, 2009; Vehovar, Ziberna, Kovacic, Mrvar, & Dousak, 2009). The 82 keywords were classified into three broad categories (see Table 1). The first category included *code* keywords (27), commonly used by offenders to alert one another to material. These included keywords such as pthc (pre-teen hardcore). The second included *thematic* keywords (23) not directly linked to CE but typically present (e.g., boy, girl, child). The third included *sex-oriented* keywords (32) that referenced sexual organs or acts

(e.g., pussy, cock, oral). Although three keyword categories were created, for analysis purposes, none were weighted as more or less important during the CENE search process. For example, the presence of three *code* keywords, but no *thematic* or *sex-oriented*, did not result in the inclusion of a website in the data. Research by Steel (2009) revealed that non-code keywords were just as likely to be linked to CE material as code keywords (see also Fournier et al., 2014). As the webcrawler only searched for the presence of keywords, the context of the keywords' use was not able to be determined. This is a limitation of keywords as a criterion; however, the general patterns found using keywords commonly linked to CE material are important. To aid in addressing this limitation, we selected roots of words (e.g., bath instead of bathing) and included multiple spellings (e.g., paedo and pedo). However, the use of short keywords, examined outside of their context, means that the webcrawler could, for example, identify a website as containing the word "anal," when it actually contained "analyze" or "analogy." We address this limitation in more detail in the following.

Data

CENE collected data on 30 networks surrounding *seed* websites. Ten of the networks began from a CE-seed while the remaining 20 comparison networks began from a sexuality-seed (10) or sports-seed (10) website. Using a repeated measures design, data were collected in 10 waves, at an interval of 42 days. Network composition followed a snowball sampling method via hyperlinks among websites (Burris et al., 2000; Westlake et al., 2011). As the nature of the seed can bias the sample derived from snowball sampling (Heckathorn, 2007; Salganik & Heckathorn, 2004), half of the networks began with a *blog* seed while the other half began with a *site* seed. This aided in maximizing network diversity, allowing us to determine whether the starting point influenced the created network. A *blog* was defined as a website with user-generated posts, in a traditional web-log setup. A *site* was defined as a website with interlocking webpages that did not meet the criteria of a blog. This included discussion forums and photo galleries.

Seed Website Selection

Each of the initial 30 CENE crawls began with a seed website selected by the researchers. For the 10 CE networks, seed websites were selected from two sources. The first source was a list, provided by the RCMP, of websites known to be involved in the distribution of CE material. This list accounted for four (two blogs and two sites) of our 10 seeds. These four were chosen because they did not require registration to view website content. We excluded websites that required registration for three reasons: (a) Websites use various methods for registering users. Therefore, the additional coding required to address each method was beyond the capabilities of CENE; (b) even if multiple registration methods were included in CENE coding, websites use different tools, such as CAPTCHA images, sounds, and unique questions, to minimize bot registration; (c) there were potential legal and ethical issues with accessing private websites. The

second source was a list of websites identified, and inspected in previous research, to be involved in the distribution of CE material. This list accounted for the remaining six seed websites. Each CE-seeded network initially included an average of 305.10 (*SD* = 2.33) websites and was recrawled every 42.14 (*SD* = 4.45) days.

For the 10 non-CE sexuality networks, seed websites were chosen using Google© search engine. Several search terms were used, to include a broad spectrum of websites related to sexuality. Four (two blogs and two sites) seeds were sex education (SE) websites selected using the terms *sexuality* and *education*. The remaining six seeds were adult pornography websites. Three (one blog and two sites) were selected using the term *BDSM*, which stands for bondage/discipline, sadomasochism, and dominance/submission (Wiseman, 1996), whereas the remaining three were selected using the term *sex*. For each search, the most popular websites (i.e., the websites that were first in the search results) that met our criteria were chosen. Although the data collected on the non-CE sexuality and sports websites were the same as for the CE websites, no inclusion/exclusion criterion was specified. Each non-CE sexuality network began with an average of 306.30 (*SD* = 6.00) websites and was recrawled every 41.62 (*SD* = 7.71) days.

For the 10 sports networks, blog seeds were selected using a sports blog popularity ranking website (http://labs.ebuzzing.com/top-blogs/sports), while the site seeds were selected using a sports marketing website (http://www.marketingcharts.com). Websites that tailored to specific teams were excluded whereas those covering an array of sports were preferred. Each sports network began with an average of 301.40 (*SD* = 1.65) websites and was recrawled every 41.69 (*SD* = 4.44) days.

Composite Measures

Network and website characteristics were compared across the 30 networks, between seed-type (blog/site) and genre-type (CE, non-CE sexuality, and sports). Two additional composite measures were created using subsets of keywords.

Sex focus. Websites were classified as being either boy or girl oriented based on the relative frequency of specific keywords. *Boy* keywords were boy, son, twink, penis, and cock while the *Girl* keywords were girl, daughter, nimphets/nymphets, Lolita/lola/lolli/lolly, vagina, and pussy.

Content focus. Websites were classified as being either explicit or non-explicit oriented based on the relative frequency of specific keywords. *Explicit* comprised 21 keywords related to severe sexual abuse (e.g., cries, torture, and rape) whereas *Non-explicit* comprised 15 keywords related to personal characteristics (e.g., innocent, lover, and smooth).

Network Measures

The Internet is an ideal medium for extracting and analyzing social networks, given the inherent nature of online interactions and digital information (Hogan, 2008).

Online interactions are directed, in that there is usually a sender and a receiver, whereas the encoding process of digital information makes identifying network connections straightforward. For websites, this is evidenced by the directing from one website to another through encoded text, known as hyperlinks. Thus, website networks can be conceptualized as evidence-based hyperlinked networks (De Maeyer, 2013; Park, 2003; Rodriguez, Leskovec, & Scholkopf, 2013). Using measures of network cohesion—density, clustering coefficient, and reciprocity—provides important information regarding how networks containing websites involved in the distribution of CE material compare with non-CE networks. Cohesion measures were compared across network genre and seed-type at Wave 1 and Wave 10.

Density. It measures the proportion of direct connections present between websites in relation to all possible network connections (Garton, Haythornthwaite, & Wellman, 1997). Density is used to determine how cohesive is a network and how effectively websites communicate with one another.

Clustering coefficient. This examines the likelihood that two connected websites are connected to a same third website (Medina, Matta, & Byers, 2000). Compared with density, the clustering coefficient informs whether network ties are evenly distributed or whether websites are clustered in sub-groups.

Reciprocity. Reciprocity identifies the proportion of nodes that directly reference one another (Wasserman & Faust, 1994). That is, if Website A references Website B, does Website B also reference Website A? Like density, reciprocity measures global camaraderie or whether websites operate in isolation.

Results

Validating Selection Criteria

While CE-related keywords have been used prominently to identify material, the only true valid measure is a hash value, as it confirms the presence of a known illegal image. Yet, there are trade-offs to both. The problem with hash values is that they are dependent on the images having been unmodified and detected in prior police investigations. A keyword strategy brings more false positives but avoids the narrow funnel of a pure hash value strategy. The total frequency of known CE images criterion and per webpage frequency of keywords criterion, along with subsets of keywords (per webpage) are presented in Table 2. As an initial within-genre comparison showed no significant differences between seed-type, across any set of keywords, Table 2 displays the combined blog-seed and site-seed networks. Measures where websites within CE-seeded networks differed from non-CE sexuality or sports websites are noted.

The use of hash value databases and keywords appear to be effective criteria for delineating between CE websites and both similar (non-CE sexuality) and dissimilar (sports) websites. First, regardless of classification category, all but two of the 22,729

Table 2. Presence of Webcrawler Criteria Across Three Network Genres at Wave 1 and Wave 10.

	CE networks	Non-CE sexuality networks	Sports networks
Total images			
Child exploitation			
Wave 1	12,966	0	0
Wave 10	239	7	0
Child nudity			
Wave 1	481	2	0
Wave 10	6	0	0
Collateral			
Wave 1	9,180	0	0
Wave 10	2,782	1,240	5
Keywords per webpage			
Code	0.11	0.13	0.16
Thematic	195[a,b]	64	68
Sex-oriented	124[a,b]	77	39
Boy	146[a,b]	25	39
Girl	16[b]	16	3
Explicit	46[a,b]	17	8
Non-explicit	191[a,b]	31	21

Note. CE = child sexual exploitation.
[a]Statistically different compared with non-CE sexuality networks ($p < .01$).
[b]Statistically different compared with sports networks ($p < .01$).

hash values in our database identified at Wave 1 were located in CE-seeded networks. This included all 12,966 *Child Exploitation* images identified. By Wave 10, there were 81% fewer identified hash values, suggesting those that were present a year back (Wave 1) had been removed. Of the 4,274 CE images identified in Wave 10, 71% were located in CE-seeded networks. Although 1,240 *Collateral* images were identified within non-CE sexuality networks, 94% were located on three websites.

Overall, keywords were a valid criterion to discriminate between the websites connected directly and indirectly to a CE-seed, compared with others (Table 2); however, that comes with a caveat. *Code* keywords, the expected most reliable subset, was found equally across network genres. Moreover, their presence was carried by several outliers. For example, in one sports network, a website had 140,047 references to *pthc*, whereas in a non-CE sexuality network, one website referenced *paedo* 35,630 times. When these outliers were removed, each network's average fell from 475 to 10 and 131 to 14, respectively. The impact of outliers was not limited to comparison networks. Within one CE-seeded network, a website referenced *paedo/pedo* 351,445 times, whereas in another network, a different website referenced it 260,307 times. For each website, these accounted for 99% of all code keyword references. Likewise, their removal resulted in similar drops in network averages. Most important to our findings

was, of the 27 code keywords, almost 60% (16) were present less than 0.1 times *per website* in CE-seeded networks, with six never being present. This stresses the importance of selecting current criteria, rather than criteria that has been useful in previous research, and is discussed in greater detail below.

Beyond code keywords, *thematic, sex-oriented, explicit*, and *non-explicit* subsets of keywords were significantly ($p < .01$) less frequent within the sports networks and non-CE sexuality networks (Table 2). While the lower frequencies in sports networks are expected, the lower frequencies within non-CE sexuality networks is important to highlight. The nature of adult pornography suggests that sex-oriented (e.g., sex and naked) and explicit (e.g., anal and fuck) keywords would feature prominently within non-CE sexuality networks. The higher frequency in non-CE sexuality networks, compared with sports networks, supports this hypothesis. However, the even higher frequency in CE-seeded networks suggests that while the mere presence of explicit and sex-oriented keywords does not distinguish CE-related websites from non-CE sexuality websites, the quantity of the keywords can. As a result, the keywords function as an important quantity criterion rather than a presence criterion. Combined, our hash values and keywords findings suggest each distinguishes one genre (e.g., child exploitation) from both similar (non-CE sexuality) and dissimilar (sports) genres. However, the frequency and current relevance (i.e., popularity) of keywords have mediating effects.

Comparing CE Networks and Non-CE Networks

Hash value databases and code keywords have been the primary indicators used by researchers to distinguish CE-related websites from non-CE related websites. However, a key question is whether websites within networks beginning with a CE-seed differ in any other significant way. Preliminary within-network-genre comparisons revealed that the seed-type did not modify the website characteristics within a network. Table 3 summarizes the general website characteristics of each network genre—average number of websites, webpages per website, and images and videos per webpage. The median value for each characteristic is also displayed while characteristics where CE-seeded networks differed ($p < .01$) from non-CE sexuality and/or sports networks are noted.

General website characteristics, summarized in Table 3, show that CE-seeded networks were easily distinguishable from sport networks, having more webpages per website, and images and videos per webpage. At the same time, they were less distinguishable from non-CE sexuality networks, having more images per webpage, and outgoing and incoming hyperlinks. The limit placed on the size of each network (approximately 300 websites and 500,000 webpages) means that we would expect the average number of websites and webpages to be comparable across genres. However, websites within CE-seeded networks, and non-CE sexuality networks, were significantly ($p < .02$) larger. The larger size of sex genre websites (CE, pornography, or education) may be the result of volume of material or, in the case of CE, a function of the detection-avoidance strategy of hiding content on hard-to-reach webpages.

Table 3. General Website Characteristics Across Three Network Genres at Wave 1.

	CE networks	Non-CE sexuality networks	Sports networks
Average number of websites	306	306	302
Average number of webpages (median)	1,583[a] (77)	1,837 (71)	1,084 (20)
Number of images per webpage (median)	20.7[b,c] (7.0)	5.9 (1.9)	4.9 (2.0)
Number of videos per webpage (median)	0.16[c] (0)	0.21 (0)	0.04 (0)
Outgoing hyperlinks (median)	23[b,c] (10)	15 (8)	31 (12)
Incoming hyperlinks (median)	23[b,c] (20)	15 (8)	31 (13)

Note. CE = child sexual exploitation.
[a]Statistically different compared with sports networks ($p < .02$).
[b]Statistically different compared with non-CE sexuality networks ($p < .01$).
[c]Statistically different compared with sports networks ($p < .01$).

Advances in video recording technology and increases to download (and upload) speeds may correspond to an increase in the amount of CE videos being distributed going forward. Despite this hypothesis, Table 3 shows that distribution remains heavily image-based, suggesting that CE images can still be an effective criterion to find child exploitation websites. Networks beginning with a CE-seed website averaged 21 images per webpage whereas non-CE sexuality and sports networks averaged six and five, respectively. Although CE-seeded networks and non-CE sexuality networks displayed similar rates of videos per webpage, when non-CE sexuality networks were divided into pornography and SE, the results differed. Within SE-seeded networks, webpages averaged 0.07 videos (similar to sports-seeded networks), whereas within pornography-seeded networks webpages averaged 0.44 videos. The abundance of images in CE-seeded networks may reveal (a) the continuing tendency toward image-based distribution, as it is easier; (b) a slower movement toward CE video distribution; or (c) that videos increase risk of detection, through possible recognition of video features. Regardless, the higher rate of videos per webpage in pornography-seeded networks suggest that, over time, CE distribution may include more video distribution.

Individual characteristics are important for describing websites, however, websites and their corresponding network do not operate independently. From the structure of a network, information regarding individual website behavior and patterns can be determined. Table 4 summarizes the change in network cohesion across the three network genres, at Wave 1 and Wave 10. Sports networks are divided between blog-seed and site-seed as the two significantly differed on hyperlinking and each network cohesion measure. Significant differences between network-genres (and seed) are noted.

Hyperlinks to and from a website can be viewed as proxy measures for a website's popularity (incoming hyperlinks) and embeddedness (outgoing) within a larger network. For illegal websites, connections to and from other websites potentially increase the risk of detection, but also the opportunity to attract consumers. CE-seeded

Table 4. Network Cohesion Measures at Wave 1 and Wave 10.

	CE networks	Non-CE sexuality networks	Sports blogs	Sports sites
Hyperlinking at Wave 1				
Outgoing (median)	23[a,b,c] (10)	15 (8)	45 (17)	15 (10)
Incoming (median)	23[a,b,c] (20)	15 (8)	45 (17)	15 (10)
Density (%)				
Wave 1	39[a,b,c]	5	15	5
Wave 10	7[a,b,c]	5	14	5
Clustering coefficient				
Wave 1	0.43[b]	0.42	0.59	0.45
Wave 10	0.43[c]	0.40	0.48	0.38
Reciprocity (%)				
Wave 1	23[b]	24	55	24
Wave 10	22	23	36	19

Note. CE = child sexual exploitation.
[a]Statistically different compared with non-CE sexuality networks ($p < .01$).
[b]Statistically different compared with sports blogs ($p < .01$).
[c]Statistically different compared with sports sites ($p < .01$).

networks hyperlinked more than non-CE sexuality-seeded and sports site-seeded networks, and less than sports blog-seeded networks (Table 4). While the median number of outgoing and incoming hyperlinks did not vary within non-CE sexuality and sports networks, they did for CE-seeded networks (10 and 20). This suggests that within CE-seeded networks there are websites that act as directories, connecting users to all websites, and those that act as suppliers, connecting to only selective websites. Combined with similarities in reciprocated hyperlinks, to comparison networks, it appears directory-based websites hyperlink to all suppliers, whereas suppliers do not hyperlink to one another but *do* hyperlink back to some directory-based websites. This conclusion is supported by the network density and clustering coefficient. CE-seeded networks had a higher proportion of all possible network connections present (density), which held until Wave 10, but were equally as likely as comparison networks to have two websites that were connected to one another be connected to a common third (clustering). In other words, more websites acted as hubs (brokers) within CE-seeded networks, compared with within non-CE sexuality and sports networks. This finding has implications for how we conceptualize distribution and competition within online illegal networks and removal strategies.

Discussion

Automated data collection for IMR is an efficient, cost-effective technique and can be useful for collecting data on topics of a sensitive or graphic nature (e.g., CE). Although useful, these tools come with the caveat that by not looking at the data directly, there are questions

regarding the validity of the data collected. In the current study, we began to address questions surrounding validity by designing a webcrawler to collect data on networks surrounding CE-seeds and compare the presence of CE criteria to similar (non-CE sexuality) and dissimilar (sports) networks. From the topic-specific criteria, selected to guide the webcrawler, we (a) provide recommendations for improving the reliability of selection criteria for CE investigation and researchers and (b) outline how CE-seeded website networks differ from comparisons and what these differences tell us about how they function.

Research into the distribution of CE material in cyberspace has primarily used image hash value databases and keywords to identify CE websites and content. Our results show that images from all three of our database's categories were prevalent in CE-seeded networks and minimally in non-CE-seeded networks (Table 2). *Code* keywords—those suggested as identifiers by past researchers—were equally prevalent across all three genres of networks; however, other "non-code" keywords were more frequent in CE networks. Based on these results, we have recommendations for improving selection criteria. First, specialized databases (e.g., hash values) are valid criteria for identification. However, their reliability, and hence usefulness, is contingent on the completeness of the database. For CE research, the abundance of images being distributed coupled with the simplicity of changing a hash value means that hash value databases, alone, are not a reliable criterion (see also Westlake et al., 2012). Second, Internet is a fast-paced, evolving environment. Subsequent research cannot rely solely on the findings of previous research for choosing selection criteria. Of the 27 *code* keywords from past research (LeGrand et al., 2009; Steel, 2009; Vehovar et al., 2009), more than half (16) were minimally present on *any* website. Even code keywords that persist are not particularly useful, if used alone, as we found them on other genres of websites. A manual examination of the websites with abnormally high rates of code keywords found that their abundance was the result of the HTML code. For example, on fantasysnews.cbssports.com, "pthc" was a part of a coding reference to "depthchart." On several non-CE sexuality websites, the use of code keywords was actually evidence of a security measure. Like the criterion we placed on CENE, to exclude "safe" websites, some websites had a script built into their HTML code that scanned a user's post and would filter "bad words." In many cases, these included keywords such as "paedo." Although identified in isolated cases, it is a limitation that we address in more detail below. As CENE examined the HTML source code for a webpage, it registered the code keyword's presence despite its presence being an attempt by the website to exclude the keyword. Therefore, we believe that keywords are a useful selection criterion provided they (a) reflect the current, not historic, landscape of the topic of research; (b) cover a wide range of related keywords (e.g., thematic); and (c) are used more as an inclusion criterion rather than an exclusion criterion, as some keywords have different meanings in different contexts, fall out of favor, or are adopted by other sub-groups over time.

For researchers and companies attempting to detect and remove CE content from the Internet, our findings have important implications. Any search for CE-related content needs to include a multi-criterion approach that is updated regularly. In addition, the variety of content available suggests that research, and identification, needs to focus on specific types. This may be type of exchange platform (e.g., public/private

website or peer-2-peer networks), victim (e.g., age or sex), or media (e.g., videos or keywords). The graphic nature of the material being examined has also been a hurdle for continued research into online CE-related distribution. From our research, those studying or combating (allied professionals) online child exploitation have evidence that criteria can be used to autonomously identify CE-related sources (websites) using automated data collection techniques. This extends beyond distribution, as researchers examining discussion forums (e.g., Fortin & Corriveau, 2015; Tremblay, 2006) or other user networks can modify the criteria to specific topics. For example, add specific keywords to target "boy love" forums, trafficking, or live webcam broadcasts.

Our second objective was to identify how CE-seeded networks differed from similar and dissimilar genre networks. Hash values and the frequency of keywords easily distinguished sports networks (dissimilar) from CE-seeded networks but were less effective at distinguishing networks with overlapping focus (sex). While the higher quantity of (all) images in CE-seeded networks, compared with non-CE sexuality networks, point to CE distribution still being heavily dominated by images rather than transitioning to videos—an important finding regarding the evolution of online web-based CE distribution—there was one finding that we wanted to pay particular attention. Networks beginning with a CE website differed from comparison networks in hyperlinking practices (Table 3) and subsequent network cohesion (Table 4). Specifically, CE-seeded networks were more densely connected, but were structured with hubs that controlled connectivity and thus information. This practice, unique to CE-seeded networks, may be a survival tactic. If consumers are aware of "hubs" that list websites, then distribution-based websites can use these hubs to redirect previous consumers to the new website address, should their original website be removed by control agencies. Hubs also provide a location where distribution-based websites can inform the larger community of their available content and, by extension, increase traffic to their website. Equally as interesting, the proportion of hyperlinks reciprocated and the degree of clustering between subsets of websites within their larger network was similar across all three network genres. This could mean that outside of hubs CE-seeded networks operate similarly to other, legal, online networks and that CE websites may compete with one another at equivalent rates to non-CE websites. This potential idea of competition may appear to counter previous research suggesting a communal aspect to CE (e.g., Beech et al., 2008; Estes, 2001; Tremblay, 2006) and cybercrime in general (Basamanowicz & Bouchard, 2011; Dupont, 2013; Holt, 2007); however, we argue that the two are not incongruent. While the Internet has transitioned CE from a solo crime to more community forum-based, for virtual interactions and distribution, this does not mean that competition is completely non-existent. Instead, we argue that there may be more competition between illegal websites than currently suspected and that it is similar to rates found amongst legal websites.

Expanding beyond individual distribution and understanding how the larger network functions as a whole has implications for control efforts by law enforcement and private agencies (Krone, 2004). More specifically, research into the network structure facilitates identification of the most effective methods for disruption and for how new content is circulated. For those researching online CE distribution, the linkages formed between distributors may be as important as the individual distributors themselves,

given the communal aspect of this cybercrime. Coupled with the automatic detection techniques proposed, this research provides an important framework for future research into CE distribution at the micro and macro levels.

Limitations

The validity of an automated webcrawler tool is contingent on the criteria used to guide the crawling process. We discuss the limitations that result from the starting (seed) websites, inclusion criteria (keywords, media, and websites), and comparison genres. Among the 10 CE-seed websites, eight were identified as boy-focused. It is likely that these seed websites contributed to approximately 80% of websites within CE-related networks being classified as boy-focused. In the two networks beginning with a girl-focused seed, less than 20% of websites were classified as boy-focused, and *Child Exploitation* images were statistically less frequent.

Given that our study used a Canadian definition of CE material, our criteria do not necessarily translate to research where the definition of CE material is more or less stringent. In addition, the use of individual keywords as a criterion, rather than groups of keywords, can result in false positives. A manual verification of CE-seeded and non-CE-seeded websites with high frequencies of code keywords found that they were used in different contexts. This was most evident within the HTML source code of a webpage, where references to automated scripts or to non-related files (e.g., videos) resulted in a sports website, for example, appearing to include code keywords. Nevertheless, we believe that the patterns found within this study have global applicability. While countries have differing CE laws, the general context is similar and the issue of cybercrime is not unique to any location. Although individual CE keywords will be used in non-related environments, sets of keywords (including sex-oriented and thematic) will still be more prevalent among CE websites.

Increases to the capability of video recording devices and download speeds point to the potential rise in the number of CE-related videos being distributed in cyberspace and/or live webcam performances by sexually exploited children. As of yet, no database of known CE videos exists. Given this, our webcrawler did not include any video-based criteria. While we collected descriptive information on the number of videos being distributed, we were unable to verify if any were CE-related. The increase in video distribution, across all genres, on the Internet (e.g., YouTube©) point to the need for a video-based database, and criterion for CE research and investigations. Specific to the current research, the additional storage space required to display videos, compared with images, may point to video-based distribution being more prevalent on private websites or networks, where users pay for access. Growth in the number of live webcam performances adds further complexity to the collection of data on CE-related websites as the webcam stream (a) may not be directly connected to the website; (b) would regularly include "new" content, yet be interpreted by a webcrawler as one video, given that it would be a static webcam connection; and (c) would not be cataloged in any video database. Therefore, this study was focused on the distribution of CE-related images and limited with regard to the implications for how CE video-based networks function.

The current study examined publicly accessible websites, excluding websites that required a password or registration, to access hidden areas, and those found on the *Deep Web*, which is comprised of dynamic webpages not indexed by search engines (Bergman, 2001). The targeting of our study to public websites points to a typical limitation of many automated webcrawler tools. Given the simplicity of their build, many are unable to automatically access private/hidden data. The added complexity of coding a webcrawler to register an account or "verify" that it is human may be beyond the capabilities of many social scientists. This inherent limitation affects the validity of any data collected from their use. However, this does not limit the use of webcrawlers as a strategy for gathering data on protected networks. Although the scope would not be as large, a webcrawler can be launched from within a password protected area allowing for complete data collection from one source (e.g., discussion forum website). Continued interdisciplinary partnerships are necessary to integrate methods for accessing more secure/private data. This is especially true for researchers interested in underground criminal activities on the Deep Web.

Finally, the two comparison genres selected for this study were sexuality and sports. These were chosen because we felt that an initial step in the validation process was to select a similar and dissimilar genre, to determine which criteria failed to distinguish between legal and illicit sexuality-based genres, as well as from other genres (e.g., sports). Although SE seed websites were included, a limitation of this study was that child-related seed websites were not included. However, the inclusion of sports genre provided a "control" comparison that aided in determining whether the website and network characteristic effects found were better attributed to differences between legal (sexuality and sports) and illicit (CE) genres or differences between sexuality-based (sexuality and CE) and non-sexuality-based (sports) genres. In addition, it helped nuance the finer details of criteria selection, as evidenced by the frequency of CE code keywords within sports networks, which may have not been as clearly evident without the inclusion of a dissimilar genre. As a result, the findings from this study provide a useful base for the continued refinement of selection criteria, and the potential problems that may arise (e.g., webpage coding).

Conclusion

Growth in automated data collection, for cybercrime research, has occurred without the necessary, vigorous, validation of these techniques. Comparing the composition of a series of networks beginning with known CE websites, to those derived from non-CE sexuality and sports websites, we (a) provided recommendations for improving the validity of automated webcrawler tools for CE research, and (b) identified criteria and characteristics that can be used to distinguish CE-seeded networks from non-CE-seeded networks. Using criteria selected from previous research on the topic of online CE distribution, we found that image databases was a valid criterion but their lack of completeness limit reliability, whereas keywords was a reliable criterion but the constant evolution of the Internet limit their validity. Comparing CE networks to non-CE sexuality and sports networks, we found that websites within CE-seeded networks

were larger (than sports websites) and more image-based with different hyperlinking properties, whereas the CE networks were more dense but equal in clustering and reciprocity suggesting the presence of hubs.

The use of automated data collection tools, such as webcrawlers, provide a great advantage to researchers as they are more efficient and require minimal intervention. However, for automated data collection techniques to be useful in any discipline, prior research into the field in question must be completed to ensure that the best and most recent selection/inclusion criteria are chosen. Failure to select relevant criteria, or relying on one criterion rather than several, can lead to high rates of false positives and an inaccurate representation of the current landscape. While their validity and reliability is dependent on the criteria used and their objective, webcrawlers are a great asset for third-party companies (e.g., Blogger©), to ensure that their terms of service are being adhered to by clients, or to software engineers enforcing copyright laws or identifying security vulnerabilities being traded on the black market. In these situations, automated data collection tools can be a first line of defense, by conducting an initial scan of the source, to flag potential issues, that can then be manually verified by humans. The customizability of automated data collection tools mean companies would be able to select the best criteria to fit their needs, or what they are trying to identify, thereby proving to be a reliable criterion.

For researchers and organizations combating the distribution of CE material in cyberspace, the graphic nature of some of the material viewed can have substantial impacts on the retention and psychological health of employees (Bourke & Craun, 2014; Krause, 2009). Continued improvements to the validity and reliability of the criteria selected may aid with prolonging the careers of those investigating the crime and increase the number willing to research the topic. Together, better detection techniques and strategies can be developed while the criminal processes involved in distribution can be better understood and explained.

Acknowledgments

The authors thank research assistant Ashleigh Girodat for her contributions to data collection and analysis.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The authors thank the Social Science and Humanities Research Council for funding this project (# 435-2012-0336).

Note

1. Although automated data collection tools can be used to index a singular website, they are just as often used to index larger networks. One example, though with a slightly different purpose, are those used by search engines to index websites throughout the world.

Although the network limits are sometimes partly (Burris, Smith, & Strahm, 2000; Chau & Xu, 2008) or fully pre-determined (Dykstra & Sherman, 2013; Saari & Jantan, 2013), autonomous webcrawlers have also been used to study unknown criminal networks. For example, Layton, Watters, and Dazeley (2011) used automated data collection techniques to analyze phishing campaign networks. Allodi, Shim, and Massacci (2013) explored the online black market trading of tools that could be used to exploit computer system vulnerabilities. Kanich and colleagues (2011) scanned and monitored spam-advertising websites. Where automated data collection techniques have been most prominent are in the study of online terrorist networks (Ball, 2016; Chen, 2012; Fu, Abbasi, & Chen, 2010; Zhou, Reid, Qin, Chen, & Lai, 2005).

References

Allodi, L., Shim, W., & Massacci, F. (2013). Quantitative assessment of risk reduction with cybercrime black market monitoring. In *Security and Privacy Workshops 2013 IEEE* (pp. 165-172). San Francisco, CA: IEEE.

Almutairi, A., Parish, D., & Phan, R. (2012). *Survey of high interaction honeypot tools: Merits and short-comings*. Retrieved from http://www.cms.livjm.ac.uk/pgnet2012/Proceedings/Papers/1569604821.pdf

Armstrong, H. L., & Forde, P. J. (2003). Internet anonymity practices in computer crime. *Information Management & Computer Security*, *11*, 209-215.

Babchishin, K. M., Hanson, R. K., & van Zuylen, H. (2015). Online child pornography offenders are different: A meta-analysis of the characteristics of online and offline sex offenders against children. *Archives of Sexual Behavior*, *44*, 45-66.

Ball, L. (2016). Automating social network analysis: A power tool for counter-terrorism. *Security Journal*, *29*, 147-168. doi:10.1057/sj.2013.3

Basamanowicz, J., & Bouchard, M. (2011). Overcoming the Warez paradox: Online piracy groups and situational crime prevention. *Policy & Internet*, *3*, Article 4. doi:10.2202/1944-2866.1125

Beech, A. R., Elliott, I. A., Birgden, A., & Findlater, D. (2008). The Internet and child sexual offending: A criminological review. *Aggression and Violent Behavior*, *13*, 216-228.

Bergman, M. K. (2001). The deep web: Surfacing hidden value [White paper]. *Journal of Electronic Publishing, 7*(1). doi:10.3998/3336451.0007.104

Bossler, A. M., & Burruss, G. W. (2011). The general theory of crime and computer hacking: Low self-control hackers? In T. Holt & H. Schell (Eds.), *Corporate hacking and technology driven crime: Social dynamics and implications* (pp. 38-67). Hershey, PA: IGI Global.

Bossler, A. M., & Holt, T. J. (2010). The effect of self-control on victimization in the cyberworld. *Journal of Criminal Justice*, *38*, 227-236.

Bouchard, M., Joffres, K., & Frank, R. (2014). Preliminary analytical considerations in designing a terrorism and extremism online network extractor. In V. Mago & V. Dabbaghian (Eds.), *Computational models of complex systems* (pp. 171-184). Cham, Switzerland: Springer.

Bourke, M. L., & Craun, S. W. (2014). Secondary traumatic stress among Internet crimes against children task force personnel. *Sexual Abuse: A Journal of Research and Treatment*, *26*, 586-609.

Burris, V., Smith, E., & Strahm, A. (2000). White supremacist networks on the Internet. *Sociological Focus*, *33*, 215-235.

Callanan, C., Gercke, M., De Marco, E., & Dries-Ziekenheiner, H. (2009). *Internet blocking*. Dublin, Ireland: Aconite Internet Solutions.

Canadian Criminal Code. (1985). *R.S.C., 1985, c. C-46 s163*. Retrieved from http://laws-lois.justice.gc.ca/eng/acts/C-46/FullText.html

Carr, A. (2004). *Internet traders of child pornography and other censorship offenders in New Zealand*. Wellington, New Zealand: Department of Internal Affairs. Retrieved from http://www.dia.govt.nz/Pubforms.nsf/URL/entirereport.pdf

Chang, L., & Krosnick, J. A. (2009). National surveys via RDD telephone interviewing versus the Internet: Comparing sample representativeness and response quality. *Public Opinion Quarterly*, *73*, 641-678. doi:10.1093/poq/nfp075

Chau, M., Shiu, B., Chan, I., & Chen, H. (2007). Redips: Backlink search and analysis on the web for business intelligence analysis. *Journal of the American Society for Information Science and Technology*, *58*, 351-365.

Chau, M., & Xu, J. (2008). Using web mining and social network analysis to study the emergence of cyber communities in blogs. *Terrorism Informatics*, *18*, 473-494.

Chen, H. (2012). *Dark web: Exploring and data mining the dark side of the web*. New York, NY: Springer. doi:10.1007/978-1-4614-1557-2

Choi, K. C. (2008). Computer crime victimization and integrated theory: An empirical assessment. *International Journal of Cyber Criminology*, *2*, 308-333.

Chow-White, P. A. (2006). Race, gender, and sex on the net: Semantic networks of selling and storytelling sex tourism. *Media, Culture & Society*, *28*, 883-905. doi:10.1177/0163443706068922

Décary-Hétu, D., & Dupont, B. (2012). The social network of hackers. *Global Crime*, *13*, 160-175.

Décary-Hétu, D., Morselli, C., & Leman-Langlois, S. (2012). Welcome to the scene: A study of social organization and recognition among warez hackers. *Journal of Research in Crime & Delinquency*, *49*, 359-382.

De Maeyer, J. (2013). Towards a hyperlinked society: A critical review of link studies. *New Media & Society*, *15*, 737-751.

Dillman, D. A. (2007). *Mail and Internet surveys: The tailored design method* (2nd ed.). New York, NY: John Wiley.

Dupont, B. (2013). Skills and trust: A tour inside the hard drives of computer hackers. In C. Morselli (Ed.), *Crime and networks* (pp. 195-217). New York, NY: Routledge.

Dykstra, J., & Sherman, A. T. (2013). Design and implementation of FROST: Digital forensic tools for OpenStack cloud computing platform. *Digital Investigation*, *10*, S87-S95. doi:10.1016/j.diin.2013.06.010

Elliott, I. A., Beech, A. R., Mandeville-Norden, R., & Hayes, E. (2009). Psychological profiles of Internet sexual offenders: Comparison with contact sexual offenders. *Sexual Abuse: A Journal of Research and Treatment*, *21*, 76-92.

Estes, R. J. (2001). *The sexual exploitation of children: A working guide to the empirical literature*. Philadelphia, PA: National Institute of Justice.

Evans, R. D., Forsyth, C. J., & Wooddell, G. (2000). Macro and micro views of erotic tourism. *Deviant Behavior*, *21*, 537-550.

Fortin, F., & Corriveau, P. (2015). *Who is Bob_34?* Vancouver, Canada: University of British Columbia Press.

Fournier, R., Cholez, T., Latapy, M., Chrisment, I., Magnien, C., Festor, O., & Daniloff, I. (2014). Comparing pedophile activity in different P2P systems. *Social Sciences*, *3*, 314-325.

Frank, R., Westlake, B. G., & Bouchard, M. (2010, August). The structure and content of online child exploitation. *Proceedings of the 16th ACM SIGKDD Workshop on Intelligence and Security Informatics*, Washington, DC.

Fu, T., Abbasi, A., & Chen, H. (2010). A focused crawler for Dark Web forums. *Journal of the American Society for Information Science and Technology*, *61*, 1213-1231.

Garton, L., Haythornthwaite, C., & Wellman, B. (1997). Studying online social networks. *Journal of Computer-Mediated Communication, 3*(1). doi:10.1111/j.1083-6101.1997.tb00062.x

Gillespie, A. A. (2011). *Child pornography: Law and policy*. New York, NY: Routledge.

Grabosky, P., Smith, R. G., & Dempsey, G. (2001). *Electronic theft: Unlawful acquisition in cyberspace*. Cambridge, UK: Cambridge University Press.

Heckathorn, D. D. (2007). Extensions of respondent-driven sampling: Analyzing continuous variables and controlling for differential recruitment. *Sociological Methodology*, *37*, 151-207.

Hewson, C., Yule, P., Laurent, D., & Vogel, C. M. (2003). *Internet research methods: A practical guide for the social and behavioral sciences*. London, England: Sage.

Higgins, G. E., & Marcum, C. D. (2011). *Digital piracy: An integrated theoretical approach*. Raleigh: Carolina Academic Press.

Hogan, B. (2008). Analyzing social networks via the Internet. In N. Fielding, R. Lee, & G. Blank (Eds.), *The SAGE handbook of online research methods* (pp. 141-161). London, England: Sage.

Holt, T. J. (2007). Subcultural evolution? Examining the influence of on and offline experiences on deviant subcultures. *Deviant Behavior*, *28*, 171-198.

Holt, T. J., Blevins, K. R., & Burkert, N. (2010). Considering the pedophile subculture on-line. *Sexual Abuse: A Journal of Research and Treatment*, *22*, 3-24.

Internet Watch Foundation. (2014). *IWF Operational Trends 2014*. Retrieved from https://www.iwf.org.uk/resources/trends

Iqbal, F., Fung, B. C. M., & Debbabi, M. (2012). Mining criminal networks from chat log. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology* (Vol. 1, pp. 332-337). doi:10.1109/WI-IAT.2012.68

Joffres, K., Bouchard, M., Frank, R., & Westlake, B. G. (2011, September). Strategies to disrupt online child pornography networks. *Proceedings of the 2011 EISIC—European Intelligence and Security Informatics* (pp. 163-170), Athens, Greece.

Kanich, C., Chachra, N., McCoy, D., Grier, C., Wang, D., Motoyama, M., . . . Voelker, G. (2011, August). No plan survives contact: Experience with cybercrime measurement (Article 2). *CSET'11 Proceedings of the 4th Conference on Cyber Security Experimentation and Test*, Berkley, CA.

Karpf, D. (2012). Social science research methods in Internet time. *Information, Communication & Society*, *15*, 639-661. doi:10.1080/1369118X.2012.665468

Kontostathis, A., Edwards, L., & Leatherman, A. (2010). Text mining and cybercrime. In M. Berry & J. Kogan (Eds.), *Text mining: Applications and theory* (pp. 149-164). Chichester, UK: John Wiley.

Krause, M. (2009). Identifying and managing stress in child pornography and child exploitation investigators. *Journal of Police and Criminal Psychology*, *24*, 22-29.

Krone, T. (2004). A typology of online child pornography offending. *Trends & Issues in Crime and Criminal Justice*, *279*, 1-6.

Latapy, M., Magnien, C., & Fournier, R. (2013). Quantifying paedophile activity in a large P2P system. *Information Processing and Management*, *49*, 248-263.

Layton, R., Watters, P., & Dazeley, R. (2011). Automatically determining phishing campaigns using the USCAP methodology. In *eCrime Researchers Summit, 2010* (pp. 1-8). Los Alamitos, CA: IEEE.

LeGrand, B., Guillaume, J., Latapy, M., & Magnien, C. (2009). *Technical report on dynamics of paedophile keywords in eDonkey queries*. Measurement and analysis of P2P activity against paedophile content project. Retrieved from http://antipaedo.lip6.fr/T24/TR/kw-dynamics.pdf

Maimon, D., Alper, M., Sobesto, B., & Cukier, M. (2014). Restrictive deterrent effects of a warning banner in an attacked computer system. *Criminology*, *52*, 33-59.

Marín, J. M. F., Naranjo, J. Á. M., & Casado, L. G. M. (2015). Honeypots and honeynets: Analysis and case study. In M. M. Cuz-Cunha & R. M. Portela (Eds.), *Handbook of research on digital crime, cyberspace security, and information assurance* (pp. 452-482). Hershey, PA: IGI Global.

McGuire, M., & Dowling, S. (2013). Cyber-dependent crimes. In Home Office (Ed.), *Cyber crime: A review of the evidence*, Research Report 75 (pp. 4-34). London, England: Home Office.

Medina, A., Matta, I., & Byers, J. (2000). On the origin of power laws in Internet topologies. *ACM SIGCOMM Computer Communication Review*, *30*, 18-28.

Milrod, C., & Monto, M. A. (2012). The hobbyist and the girlfriend experience: Behaviors and preferences of male customers of Internet sexual service providers. *Deviant Behavior*, *33*, 792-810. doi:10.1080/01639625.2012.707502

Monto, M. A., & Milrod, C. (2014). Ordinary or peculiar men? Comparing the customers of prostitutes with a nationally representative sample of men. *International Journal of Offender Therapy and Comparative Criminology*, *58*, 802-820.

O'Halloran, E., & Quayle, E. (2010). A content analysis of a "boy love" support forum: Revisiting Durkin and Bryant. *Journal of Sexual Aggression*, *16*, 71-85.

Park, H. W. (2003). Hyperlink network analysis: A new method for the study of social structure on the web. *Connections*, *25*, 49-61.

Patchin, J. W., & Hinduja, S. (2011). Traditional and non-traditional bullying among youth: A test of general strain theory. *Youth & Society*, *43*, 727-751.

Provos, N., & Holz, T. (2007). *Virtual honeypots: From botnet tracking to intrusion detection*. Boston, MA: Pearson Education.

Ray, J. V., Kimonis, E. R., & Seto, M. C. (2014). Correlates and moderators of child pornography consumption in a community sample. *Sexual Abuse: A Journal of Research and Treatment*, *26*, 523-545.

Rice, S. R., & Ross, M. W. (2014). Differential processes of Internet versus real life sexual filtering and contact among men who have sex with men. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace, 8*, Article 1. doi:10.5817/CP2014-1-6

Roberts, J. W., & Hunt, S. A. (2012). Social control in a sexually deviant cybercommunity: A cappers' code of conduct. *Deviant Behavior*, *33*, 757-773. doi:10.1080/01639625.2012.679894

Rodriguez, M. G., Leskovec, J., & Scholkopf, B. (2013). Structure and dynamics of information pathways in online media. *Proceedings of the Sixth ACM International Conference on Web Search and Data Minin*g (pp. 23-32). New York, NY: ACM.

Rutgaizer, M., Shavitt, Y., Vertman, O., & Zilberman, N. (2012). Detecting pedophile activity in bittorrent networks. In N. Taft & F. Ricciato (Eds.), *Passive and active measurement* (pp. 106-115). Berlin, Germany: Springer.

Saari, E., & Jantan, A. (2013). E-Cyborg: The cybercrime evidence finder. *2013 Proceedings of the 8th International Conference on Information Technology in Asia (CITA)* (pp. 1-6). Kota Samarahan, Malaysia: IEEE.

Salganik, M. J., & Heckathorn, D. D. (2004). Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology*, *34*, 193-240.

Schonlau, M., van Soest, A., Kapteyn, A., & Couper, M. (2009). Selection bias in web surveys and the use of propensity scores. *Sociological Methods & Research*, *37*, 291-318. doi:10.1177/0049124108327128

Seigfried, K. C., Lovely, R. W., & Rogers, M. K. (2008). Self-reported online child pornography behavior: A psychological analysis. *International Journal of Cyber Criminology*, *2*, 286-297.

Seto, M. C., Hanson, R. K., & Babchishin, K. M. (2011). Contact sexual offending by men with online sexual offenses. *Sexual Abuse: A Journal of Research and Treatment*, *23*, 124-145.

Seto, M. C., Hermann, C. A., Kjellgren, C., Priebe, G., Svedin, C. G., & Langstrom, N. (2015). Viewing child pornography: Prevalence and correlates in a representative community sample of young Swedish men. *Archives of Sexual Behavior*, *44*, 67-79.

Shropshire, K. O., Hawdon, J. E., & Witte, J. C. (2009). Web survey design: Balancing measurement, response, and topical interest. *Sociological Methods & Research*, *37*, 344-370. doi:10.1177/0049124108327130

Spitzner, L. (2003). *Honeypots: Tracking hackers* (Vol. 1). Reading, UK: Addison-Wesley.

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

Tremblay, P. (2006). Convergence settings for nonpredatory "boy lovers." In R. Wortley & S. Smallbone (Eds.), *Situational prevention of child sexual abuse* (pp. 145-168). Monsey, NY: Criminal Justice Press.

Tretyakov, K., Laur, S., Smant, G., Vilo, J., & Prins, P. (2013). Fast probabilistic file fingerprinting for big data. *BMC Genomics*, *14*, S2-S8. doi:10.1186/1471-2164-14-S2-S8

van Wijk, A., Nieuwenhuis, A., & Smeltink, A. (2009). *Behind the scenes: An exploratory investigation into the downloaders of child pornography*. Arnhem, ND: Bureau Beke.

Vehovar, V., Ziberna, A., Kovacic, M., Mrvar, A., & Dousak, M. (2009). *Technical report on an empirical investigation of paedophile keywords in eDonkey P2P Network*. Measurement and analysis of P2P activity against paedophile content project. Retrieved from http://anti-paedo.lip6.fr/T24/TR/keywords-vv.pdf

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge, UK: Cambridge University Press.

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

Westlake, B. G., Bouchard, M., & Frank, R. (2012, August). *Comparing methods for detecting child exploitation content online*. Paper presented at the European Intelligence and Security Informatics Conference, Odense, Denmark.

Wiseman, J. (1996). *SM 101: A realistic introduction*. San Francisco, CA: Greenery Press.

Wolak, J., Finkelhor, D., & Mitchell, K. J. (2005). *Child-pornography possessors arrested in Internet-related crimes: Findings from the National Juvenile Online Victimization Study*. Retrieved from http://www.missingkids.com/en_US/publications/NC144.pdf

Wolak, J., Liberatore, M., & Levine, B. N. (2014). Measuring a year of child pornography trafficking by US computers on a peer-to-peer network. *Child Abuse & Neglect*, *38*, 347-356.

Wurtele, S. K., Simons, D. A., & Moreno, T. (2014). Sexual interest in children among an online sample of men and women: Prevalence and correlates. *Sexual Abuse: A Journal of Research and Treatment*, *26*, 546-568.

Zhou, Y., Reid, E., Qin, J., Chen, H., & Lai, G. (2005). US domestic extremist groups on the Web: Link and content analysis. *IEEE Intelligent Systems*, *20*, 44-51.