May, 2011

Finding the Key Players in Online Child Exploitation Networks

Bryce Westlake, *Simon Fraser University*
Martin Bouchard, *Simon Fraser University*
Richard Frank, *Simon Fraser University*

Running Head: ONLINE CHILD EXPLOITATION NETWORKS

Finding the key players in online child exploitation networks

Bryce G. Westlake
bwestlak@sfu.ca

Martin Bouchard*
mbouchard@sfu.ca

Richard Frank
rfrank@sfu.ca

School of Criminology,
Simon Fraser University
8888 University Drive,
Burnaby, BC, Canada
V5A 1S6
Phone: 778-782-8135
Fax: 778-782-4140
mbouchard@sfu.ca

January 30, 2011

*Corresponding author

Acknowledgements.

Partial funding for this project was provided by the International Cybercrime Research Centre, Simon Fraser University. The authors would like to thank the anonymous reviewers and Eric Beauregard for helpful comments made on an earlier version of this paper.

Abstract

The growth of the Internet has been paralleled with a similar growth in online child exploitation. Since completely shutting down all child exploitation websites is difficult (or arguably impossible), the goal must be to find the most efficient way of identifying the key targets and then apprehend them. Traditionally, online investigations have been manual and centered on images. However, we argue that target prioritization needs to take more than just images into consideration, and that the investigating process needs to become more systematic. Drawing from a web-crawler we specifically designed for extracting child exploitation website networks, this study 1) examines the structure of ten child exploitation networks and compares it to a control group of sports-related websites, and 2) provides a measure (network capital) that allows for identifying the most important targets for law enforcement purposes among our sample of websites. Results show that network capital – a combination between severity of content (images, videos, and text) and connectivity (links to other websites) – is a more reliable measure of target prioritization than more traditional measures of network centrality taken alone. Policy implications are discussed.

Introduction

The Internet has revolutionized the way that we interact with others, as well as how business is conducted. While this revolution has facilitated numerous positive areas of our lives, it has also facilitated activities with more negative consequences. The global reach and anonymity of the Internet has resulted in its quickly becoming a breeding ground for illegal activities. One such activity, that continues to grow exponentially (Wolak, Finkelhor, and Mitchell 2005; Loughlin and Taylor-Butts 2009), is child exploitation. This most commonly occurs through the distribution of child pornography. In this article, "child pornography" will refer to the actual content itself, while "child exploitation" will refer to the overall phenomenon, which includes child pornography.

Due to the unregulated and seemingly anonymous nature of the Internet, online pedophile networks have flourished (Akdeniz 1999). Durkin (1997) outlines three ways that sex offenders have been able to utilize the Internet: for dissemination, social networking, and sexual communication with children. Dissemination involves the distribution of child pornographic images, videos, or textual stories, while social networking refers to communication with other pedophiles through chatrooms and newsgroups. Chatrooms and newsgroups are also used to misrepresent oneself as a youth, for the purpose of sexual communication with children. This article focuses on the first use, dissemination. As of 2009, the United Nations estimates that there are more than four million websites containing child pornography, and that 35 percent of these websites depict serious sexual assault, while 70 percent involve children under the age of eight (Engeler 2009).

Despite large investments of law enforcement resources, by global governments, and extensive monetary contributions by private organizations,[1] online child exploitation is nowhere near under control. In fact, McLaughlin (2004) estimates that less than one percent of all online pedophiles are apprehended. This is not to say that existing efforts have been futile. Image databases, such as the United States Division of Criminal Justice Services Database,[2] the International Child Sexual Exploitation Image Database (ICSE-

[1] This includes organizations specifically examining child exploitation, such as the National Child Exploitation Coordination Centre in Canada, the Child Exploitation and Online Protection Centre in the United Kingdom, and the National Centre for Missing and Exploited Children (NCMEC) in the United States, as well as corporations who conduct business online, such as Microsoft, Google, and Mastercard.

[2] The database consists of over 8000 images that are used to aid in the detection and removal of child pornography on social networking sites such as Facebook and MySpace (Office of the Attorney General 2010).

DB),[3] and the National Child Victim Identification Program (NCVIP),[4] have aided in the identification of child pornography websites and the rescuing of children being victimized. In addition, cross-national law enforcement operations, such as Avalanche, Blue Orchid, and Cathedral have resulted in the apprehension of some of the largest and most influential players of the time (see Krone 2005 for a summary of international police operations). However, as the problem continues to grow, new targets emerge.

Identifying the Key Players in Online Child Exploitation Networks

Law enforcement agencies are faced with many challenges in regards to online child pornography, not the least of which is to find the most appropriate targets to investigate. These priority targets have commonly been described by social network analysts as the "key players" (Sparrow, 1991; Borgatti 2003; Easton and Karaivanov, 2009; Schwartz and Rouselle, 2009; Malm and Bichler, in press). How one defines and measures the key player is both a matter of debate and of context. For example, a target may be important because it commits the most crime, because it possesses the most information on other network members, or because it holds a crucial position in the network that cannot be replaced. Law enforcement agencies may draw on multiple measures and criteria to identify the key players they need to target. For example, Borgatti (2006) notes that law enforcement typically has two primary goals when investigating criminal networks. The first goal is disruption, where the most connected players within the network are targeted. The three most common social network analysis methods to determine this are degree, closeness, and betweenness centrality measures. Degree centrality refers to the number of direct connections (ties) a player (node) has with all others within a network. Closeness centrality examines how close node A is to node B, in comparison to how close node C is to node B. If the path from node A to node B is shorter, then A has a higher closeness value. Betweenness centrality is the ability of one node to act as a broker between two other nodes.

Two issues with using traditional centrality measures to identify the key players within a network are design and group selection (Borgatti 2003). The design issue occurs in networks where several nodes are highly connected. In this type of network, the node with the highest degree, closeness, and/or betweenness

[3] Created in March 2009, and funded by the European Commission, the ICSE-DS is housed by the United Nation's International Criminal Police Organization, and contains more than 520,000 images (INTERPOL 2010).

[4] Developed in 1999, and launched in 2003, the NCVIP is one of the largest databases of child pornography in the world. In March of 2005, it was merged with the NCMEC database and is now jointly maintained (Ministry of Labour and Citizens' Services 2006).

may not be the most optimal target. If all the other nodes, connected to the most central node, are also highly connected, the removal of the most central node may do little to disrupt the overall network. However, if another node is targeted, which is less connected overall, but connected to other nodes that rely on it for their connection to the overall network, the overall disruption to the network can be greater. As for the group selection issue, it pertains to the problem of network redundancy. Sometimes the centrality of a node is dependent on another node—in this situation, the removal of both nodes is redundant, as the removal of just one of them has the same impact on the overall network.

The second goal of law enforcement when investigating criminal networks is intelligence collection, whereby particular players are targeted in order to maximize knowledge of the overall network (Schwartz and Rouselle 2009). The most common SNA methods for determining this are degree and closeness centrality measures. However, Borgatti (2006) argues that degree centrality is only useful when examining direct connections between nodes. Put another way, degree looks at each node separately and does not take into account the most optimal strategy for finding, and possibly removing, a group of nodes. That is, degree is less useful when trying to understand the entire network. Turning our attention to closeness centrality, although it may be useful as a measure of the minimum total "steps" separating a specific node from every other node in the network, this analysis can be misleading, as information degrades as it passes through people (Borgatti). Consider, for example, node A, who has six direct connections and can connect to every other node through 24 steps, and node B, who takes longer to go through the network (30 steps) but who has 10 direct connections. With the goal of intelligence gathering in mind, targeting node B (lower closeness score, but more direct connections) would be the right move—the information obtained could be more accurate and timely. Put another way, closeness fails to weight the quality of the information that is obtained as an individual moves further away from the originating source. Given that timely information can provide law enforcement with a better opportunity for disruption, for child exploitation networks and many other criminal networks, timeliness (or lack thereof) should be taken into consideration when identifying key players.

Translating these two goals of law enforcement to online child exploitation networks, disruption translates into a focus on websites' direct *connectivity* to one another, while the goal of intelligence collection translates into a focus on website content (i.e. *severity*) and how that content is shared across the network. The key players of online child exploitation networks should display both characteristics: they will be both highly central within their networks, and will display the most harmful content.

The centrality, or connectivity, of a child exploitation website is key to the circulation of information and content across the network—one of the main

raisons d'être of child exploitation networks. As noted previously, much of the effort to address online child pornography has focused on the presence of known images; however, Krone (2004) suggests that the focus of law enforcement should also be on the linkages between websites, and the offenders' reliance on these networks. Indeed, Beech et al. (2008) point out that child pornography networks tend to be very well organized, with systems of trade (usually pictures and video), mechanisms for circulating information (commonly through the links between websites), and methods of inclusion and exclusion of network members. In this article, we measure connectivity by the amount of links to, and from, a given website. In the absence of a direct measure of web traffic, these links provide the closest publicly available measure of exposure. We are interested in exposure and potential—sites that are linked to more often are more likely to be followed by users, just like someone who is a member of multiple groups and multiple networks has a higher likelihood of being known to more people.

The importance of content severity also cannot be neglected. The most connected websites do not necessarily host the most harmful content (and vice versa). Measuring severity of content represents an analytical challenge. Taylor et al. (2001) argue that the severity of a child pornography collection can be measured by three components: size of collection, presence of new and/or private material, and age of the children depicted. In conjunction with the Combating Paedophile Information Networks in Europe, Taylor et al. created a scale to measure picture severity. This scale comprises 10 levels of severity, beginning with children in underwear or swimsuits (1), continuing to sexually or provocative poses (5), and ending with sadism and/or bestiality (10). Although the number of images and the image content are important, other factors should also be examined. Krone (2004) points out that the individual's engagement with the material is also important.[5] For the purposes of online collections, this can take the form of the descriptions (text) posted with the image or fictitious stories about sexual exploitation. As a result, it is important that measures of severity take into account more than just the presence of images. Therefore, in the current study, we propose a measure of severity that takes into account text, images, and videos.

Current Study

As the problem of child pornography online continues to grow, and law enforcement resources are further strained, it has become imperative that

[5] Engagement can include the way the content is categorized or maintained by the offender, how long it has been in their possession, and whether the content is foreign or from their own sexual exploitation of a child (Taylor and Quayle 2003, as cited in Krone 2004).

resources be allocated in the most efficient manner. As previously discussed, Durkin (1997) states that the Internet has aided sex offenders in three key ways: aiding dissemination, social networking, and sexual communications with minors. In the current study, we look to improving current law enforcement strategies, while examining the phenomena of dissemination and social networking. We argue that these improvements must take two forms: (1) increasing the automation of searches, and (2) refining the systems being used to identify, and prioritize websites/targets. This prioritization is especially important given the size of the problem: targets are plentiful, and with limited resources, the priority should be given to the most harmful targets—the key players. Drawing on recent advances in social network analysis we develop a measure called *network capital* (NC) to identify the key players in online child exploitation networks, by focusing both on *severity* (harmfulness of the content) and *connectivity* (how exposed and easy to find is the content).

As discussed above, there are numerous technological products that aid in finding child pornography online; however, one of the key issues with these is that they still require substantial human intervention. Although we cannot get away from the manual component of child pornography searches, steps can be taken to increase the automation process of these searches, and to decrease the direct contact police officers have with child exploitative content. By decreasing the amount of material officers have to go through, we can increase the amount of time an officer can spend on investigating individuals, as well reduce the amount of content they have to examine. Therefore, within the current study, we propose a web-crawling tool that can be used to automate the process of searching websites for child pornography, and provide statistics on user-selected attributes from each website. These statistics can then be used to target key players.

In addition to automating the process of searching for content, we look to determine who the key players are within online child exploitation networks. This examination is done on the World Wide Web—arguably the medium providing the broadest coverage and most visible means for obtaining child pornography (Krone 2005).[6] Again, by properly identifying the key players we can aid law enforcement through the optimization of resource allocation. However, an important methodological challenge, before finding those key players, is to analytically define the network and its boundaries. Currently, there is a lack of

[6] It is important to acknowledge that the distribution of child exploitive content is rampant in other Internet domains such as Internet Chat Relay, Newsgroups, Web-chat, and Peer-to-Peer Networks (Carr 2004). However, as these domains operate at the micro (individual) level, while the WWW operates at the macro level, our study focuses solely on the WWW.

research exploring the structural nature of online child exploitation networks. Therefore, within the current study, we develop a method to extract child exploitation networks, map their structure and analyze their content. This is done by looking at how child exploitation networks differ from non-child exploitation networks. To determine this, we compare 10 child exploitation website networks with 10 non child exploitation networks.

Our final objective is to uncover the structure of online child pornography networks, and to identify their "hardcore key players": websites whose removal would result in the greatest reduction in network capital. This knowledge can then be used by law enforcement to make effective decisions on the methods that would have the greatest impact on the network. That is, the prioritization of targets to only those websites that are both highly connected and that also display the most harmful content. Although it would be naïve to suggest that online child exploitation can be completely eradicated, social network analysis provides a means of understanding the structure and vulnerability of online networks. In turn, this could greatly improve the effectiveness of law enforcement.

Methods

Network Extractor

For this paper, we used a custom-written crawler called Child Exploitation Network Extractor (CENE) (for the algorithm, see Appendix). CENE is a computer program that automatically browses the World Wide Web and collects information about the pages it visits. CENE starts the crawling process at a user-specified webpage, retrieves the page, analyzes it, and recursively follows the links out of the page. The crawling process is performed according to rules, and terminates when certain user-specified criteria are met. The result of the crawl is a network structure containing information about the content of the websites, and the linkages between them.

For each webpage extracted, certain statistics about the content of the webpage, such as frequency of keywords and count of images or videos, are collected. When the crawling process is complete, the statistics are aggregated up to the website level. For example the statistics for the node www.website.com are calculated from the statistics collected from all pages on that website. This is what makes up the network for this study.

Since the Internet is extremely large and a crawler would most likely never stop crawling, we implemented three conditional limits into CENE to keep the crawling process under control and network relevant. To keep the network

extraction time bounded, a limit was put on the number of pages retrieved (for this paper, it was 250,000). The network size was also fixed at a specific number of websites (200). The webpages were retrieved in such a way that each website was sampled equally, or as close to equality as possible. Finally, in order to give CENE some boundaries for the crawl and guide the network extraction process to a relevant network, a set of keywords were defined. This set consisted of 63 keywords in total, broken up into three groups. The first were words commonly used by the Royal Canadian Mounted Police (RCMP) to find illegal content containing children; most of them code words used by pedophiles online, and also included in other studies of online child pornography (Le Grand et al. 2009). The second group was words that could be classified as "softcore", and which may or may not be found on child pornography websites. These words were boy, girl, child, love, teen, variants of Lolita, twink, young, bath*, pre/post pubescent, innocent, smooth, and hairless. The third group of words was labeled as "hardcore" and included the following thirteen: sex, penis, cock, vagina, pussy, anus, anal, pedo/paedo, oral, virgin, naked, and nude. Although the focus of this study was on the most harmful content, it was important to collect a broader range of keywords for comparative and network extraction purposes. The distinction between hardcore and softcore words was based on the explicit focus on sexuality. That is, words that could be found under different, non-child pornography settings were classified as softcore.

For the crawler to include a given webpage in the analysis, it had to contain at least seven of the sixty-three unique keywords. If this criterion was met, the webpage was assumed to be relevant to the network, the statistics on that webpage calculated, and the links pointing out of the page recorded for later visit. If the page didn't meet this criterion, it was assumed to be irrelevant, and was discarded with no further links being followed from that page. This classification of webpages into the child-exploitation category based on the number of keywords has a drawback: if the threshold is too low, false-positives might be included in the analysis. Manual verification showed us that seven keywords distinguished well between child-exploitation webpages and regular webpages. In addition to keywords, the number of videos and images on each node was also recorded. To minimize irrelevant images (e.g., emoticons and logos), those smaller than 150×150 pixels were excluded; no criterion was set for videos. Where the crawler tried to follow a broken link, or if the webpage could not be retrieved for any other reason (such as a timeout or password protection), the webpage was considered inaccessible and discarded.

To construct the statistics of each node in the network, the links between websites were tracked, as well as the occurrence of each keyword aggregated to the website level. Thus, all web pages on a website contributed to the statistics for

that website. This allowed for the construction of a coherent network, complete with statistics assigned to both the websites and links.

Network Structure and Content

Ten web pages were chosen as starting points for crawls. Five websites were blogs ("Seed-Blog") and five were non-blogs ("Seed-Site"). These websites were selected using two methods. Four were selected from a list of known child pornography websites, provided by the Royal Canadian Mounted Police's Integrated Child Exploitation (ICE) unit, while the other six were selected through Google searches using known child pornography search terms (e.g., lolita, realkiddy, pthc, and nymphet).[7] This process involved inputting the keywords and manually verifying the websites suggested by Google. Once a website was found to contain child pornography, it was selected as a starting (seed) website. An attempt was made to select websites that unambiguously showed children, in order to keep the web crawling as close as possible to what would have been a manual search by an individual. As for the definition of child pornography, this included websites that did not necessarily contain hardcore content, but that presented children in sexually provocative ways. This could be done through sexual objectification, softcore videos and/or images, and obscene conversations depicting sexual activities with children. Bulletin board based forums were included as starting websites for seed-site networks, however, the forum had to require no registration to view posts on the website. If unique-user identification was required, the website was excluded from the analysis. Bit-torrent websites were also used as potential starting points for seed-site networks. The starting websites for blog networks were strictly of the blog genre. While the starting websites contained child pornography, not all of the links necessarily led to other (genuine) child pornography—even if seven of our keywords could be found. However, the information (keywords, images, videos, and connections) collected from each website allows us to rank websites according to their likelihood of containing child pornography, thus facilitating target prioritization for law enforcement agencies.

The two methods of website selection were chosen to mimic the process a person might take when searching for child pornography. The four websites provided to us by the ICE unit were selected to represent an individual being given a known website by another person. The six Google-discovered websites were selected to mirror the process an individual might go through if they went to search for the content themselves, with no other information. Using the web-

[7] The selected website was the first that came during each search undertaken, which met our requirement. An attempt was made to select boy-centered and girl-centered websites equally.

crawler, we can map out all the possible routes an individual might take through the network to obtain content[8]. This is important, as we want to obtain a good representation of what a potential user might do.

Due to the continually growing importance of blogs within online social media (Furukawa et al. 2007), we categorise the starting websites into blogs and sites.[9] This distinction between blogs and sites is important, as it may have implications for the content and structure of child exploitation networks. This is because blogs provide two benefits over sites when distributing child pornography. First, blogs are usually hosted (freely) by a third party, and require very little knowledge of web design, or computers in general. That is, the hosts typically have specific templates to follow that allow a non-technical person to easily set up a blog. Although similar templates exist for sites, there are additional steps required, such as purchasing a domain name and setting up a name server to point to the computer that is hosting the content. By contrast, many blog hosts, such as Blogger, LiveJournal, and Sensualwritter provide members with free space to post their blogs.

Second, blog website hosts provide increased anonymity as they rarely monitor the blogs created by their users. With the common requirement of only a username and an email address to create an account, the perceived chances of being apprehended for distributing child pornography through a blog is considered a priori lower than for other types of websites. For sites, a user would have to be wary of detection by law enforcement, as their name would be attached to the website registration information. Although each blog host has terms of service, it is usually the responsibility of patrons to report a blog containing copyrighted or illegal material. In fact, one of the largest hosts, Blogger/Blogspot, specifically state in their terms that they do not monitor blogs (Google 2010). Therefore, if a blog is found to be publishing illegal content, shutting it down acts as little more than a nuisance for the creator, as there is very little preventing the user from making another account and continuing their blog under a different, but usually similar, name. In other words, the anonymity of online blogs may lead to more freedom to network a certain type of content. In addition, the very nature of blogs seems to make them more likely than other types of websites to link to other blogs or websites. For example, Ali-Hasan and Adamic (2007) found that blog networks tend to have reciprocal ties between 27–43 percent of the time. Of course, a user will be more inclined to follow the content (child pornography) than pay much attention to whether it is hosted on a blog or website ("blog" or

[8] A user may or may not follow those links to find content, we examine the potential for users to do so given exposure.

[9] Non-blogs includes picture and video galleries and freely accessed chat forums; however, it excludes chatrooms and membership-based websites. Although these are important avenues to explore, they are beyond the scope of the current study.

"site"). Therefore, although each starting website for the crawl corresponds to the network type (i.e., blog for blog networks and site for website networks), it is not guaranteed that all of the websites within each network are of the same network type. That is, blogs will link to other blogs, but they will also link to other sites and vice versa with sites. Thus, our network type differentiation refers to the starting website rather than all the websites within the network.

To compare the structure and content of child pornography networks to other types of networks, five blog and five non-blog sports networks were also analyzed as a control. The starting websites were selected based on a list of the most popular sports websites (Top Sites Blogs, 2010) and the most popular sports blogs (Technorati, 2010). Like the child exploitation networks, each of the sports networks consisted of 200 nodes, a maximum of 250,000 web pages, and images greater than 150×150 pixels. Although the child pornography keywords were collected, they were not a criterion for website selection.

In addition to comparing child exploitation websites to non-child exploitation websites, we also want to see if there is something inherently different with the structure of the networks when we start with a blog versus starting with a site. The comparison between child exploitation networks starting with a blog or site will be made on several measures. The first is whether the websites are boy or girl centered, with a simple measure counting the number of times each word appears in a website (a website will be considered as boy centered if the word "boy" appears more times than the word "girl"). The second measure is the prevalence of hardcore content: a website with a higher prevalence of hardcore words (sex, penis/cock, vagina/pussy, oral, virgin, lolita, naked/nude, and anus/anal) compared to softcore words[10] will be classified as hardcore focused. The distinction between hardcore and softcore words is based on the explicit focus on sexuality. Words that could be found under different, non-child pornography settings are classified as softcore. Third, the network types will be compared on reciprocity (ie how often, when website A references website B, does website B reference website A). This measure informs us on how cohesive the network is. Finally, the network types will be compared on their overall density; another measure of cohesiveness. Density refers to the proportion of direct connections present, between websites, in relation to all possible network connections: each website being connected to every other website (Izquierdo and Hanneman 2006).

Determining the Key Players: Measuring Network Capital

[10] Softcore words, later referred to as "other words", were: boy, girl, child, love, teen, young, bathing, innocent, paedo/pedo, pre/post pubescent, and hairless/smooth.

Network capital is a term derived by Schwartz and Rouselle (2009), which takes into account the resources available to each node, the cohesiveness of the network, and the relationships between nodes. The more an individual node contributes to network capital, the more central/key the node is to the overall network. The formula itself is an extension of Borgatti's (2006) method for identifying key players. We follow Schwartz and Rouselle's network capital formula and adapt it to the specific context of online child exploitation networks by incorporating severity of content and website connectivity. By utilizing the adapted formula, we are able to meet both goals of law enforcement mentioned previously: detection and intelligence. *Network_capital* is calculated as follows:

$$\frac{Node\_Severity + Node\_Connectivity}{N + [N(N-1)RSL]}$$

Where:

N	total number of nodes in the network
RSL	resource sharing level.

Network capital comprises two key components: *node_severity* scores and *node_connectivity* scores. Schwartz and Rouselle (2009) describe *node_severity* to refer to the resources available to a given node that may, or may not, be shared with the rest of the network. In the current study, we define *node_severity* scores as the summation of three resources: the number of keywords, images, and videos per web page.[11] Analyzing the number of images and videos is not the same as analyzing the nature of their content (the way that keywords do). This study is limited by the assumption that a larger volume of images/videos is more harmful than a smaller one.

Each of the resources is standardized against the highest scoring node, for each resource, within the network. This means that for each of the three resources, the website with the highest number of images, videos, or words, receives a score of 1.0, with all other websites ranging between 0 and 1.0. The individual *node_severity* score is then the average of the three resources.[12] Therefore, the formula for calculating *node_severity* scores is as follows:

[11] Although the "type" of activity within the content is important when discussing severity, this was not possible for this study without viewing each video and image and scoring it.

[12] Because of the way the attribute weighting is derived, it is more tailored to intra- rather than inter-network comparisons.

$$\sum_{n=1}^{NAW_i} \frac{AW_{ni}}{NAW_i}$$

Where:

i	node
$AW_{1i}, AW_{2i}, AW_{3i}$	weighted number of keywords (AW_1), images (AW_2), and videos (AW_3), ranging from 0.0 to 1.0.
NAW_i	number of resources (3)

The second component of network capital is *node_connectivity*. This refers to the contribution made by a node to the overall network based on the direct connections it has to other nodes within the network, and the amount of resources it has available.[13] This is multiplied by the percentage of its resources made available by the node to the rest of the network[14] and then multiplied by any link-weighted values. In the current study, we included only one link weight: number of times node A references node B. The number of times node A references node B effectively increases the exposure of node B and thus adds to their connectivity.[15] The link weights have been standardized between 0.0 and 1.0 using the same method as the resource weights. Therefore, the formula for node_connectivity is as follows:

$$\left[\left(\sum_{n=1}^{NAW_i} \frac{AW_{ni}}{NAW_i}\right) * RSL\right](LW_{ij})$$

Where:

i	node
$AW_{1i}, AW_{2i}, AW_{3i}$	weighted number of keywords, images, and videos, ranging from 0.0 to 1.0.

[13] Although Schwartz and Rouselle's (2009) formula included indirect connections, they have been excluded in this study because the small size of the networks (200 nodes) resulted in all nodes being either one or two steps away from every other node. In other words, we hold indirect connections as a constant for the purpose of this study.

[14] For the purposes of the current study, we assumed a resource sharing level of 1.0. This is because all of the websites within our networks are open (do not require a membership) and thus all resources we have recorded are available to all others who go to the website.

[15] A key component of Schwartz and Rouselle's (2009) formula is the inclusion of isolates, but the logic of the web crawler makes it impossible to find true isolates (i.e. a website that connects to no other, and vice versa). This is a limitation of the crawler, as some isolates may lack connectivity yet host harmful content and be important targets for law enforcement.

NAW_i	number of resources (3)
RSL	the resource sharing level
LW_{ij}	weighted number of times node i references node j, ranging from 0.0 to 1.0

Results

We start by comparing the child exploitation networks to the control networks. In general, the control networks were easier to construct than the child exploitation networks. In total, of the pages we attempted to visit only 66 percent of the child exploitation seed-blog web pages were active (ie no broken links, or expired domains), while even fewer (56 percent) were active for the seed-site networks (Table 1). Conversely, the control seed-blog and seed-site networks were 98 percent active. These discrepancies might be the result of the relative legality of content within the networks. Due to the illegal nature of the child exploitation networks, they run a higher likelihood of being shutdown, while a blog about baseball does not run similar risks. One way of verifying the legality of websites being a contributing factor to their accessibility is through our measure of hardcore words per page. That is, if specific words are used by law-enforcement and other agencies to target potential child exploitation websites, the lack of these words would result in their being ignored by law enforcement. As shown in Table 1, the hardcore keywords were found at a much higher frequency per page within child exploitation networks (71 occurrences per page for seed-blog and 350 for seed-site networks) than sport-related networks (around four hardcore words per page). The *t*-tests presented in Table 2 show that these differences are statistically significant. From that point of view, it suggests that the words selected for the web crawler are good predictors of child pornography content. However, the degree to which the keywords distinguish between child pornography websites and legal pornography websites is unclear: this issue is discussed in detail later. Control networks were much lower in videos and images per page, in comparison to child exploitation networks. This may not be that surprising, as we have outlined previously that one of the key purposes of online child exploitation networks is to exchange content. Therefore, it would lead one to believe that, in comparison to an average sports website, websites devoted to child exploitation would have higher rates of videos and images per page.

Table 1 also compares the two samples of five seed-blog and five seed-site child exploitation networks (labeled based on starting point). Websites dedicated

to girls and hardcore content were more common when the seed was not a blog than when it was (t=2.22, P=0.03; t=5.48, P=0.00). Although using a blog as a starting point led to a larger number of valid nodes (138 to 60), networks with a site as a starting point averaged more web pages and images per node (see Tables 3 and 4). Seed-site networks had more hardcore words per web page, while seed-blog networks had more videos per web page. The significantly higher number of hardcore words per web page found in site networks might account for the smaller network sizes. That is, the excessive number of words may result in search engines indexing those websites more, therefore making it easier for the websites to be found and shutdown.

Table 1. Mean totals for child exploitation and control networks.

		Seed-Blogs (C.E.) n = 5	**Seed-Sites (C.E.)** n = 5	**Seed-Blogs (Controls)** n = 5	**Seed-Sites (Controls)** n = 5
Nodes (Valid)	Total (Across all networks)	688	299	938	917
	Average/network	137.6	59.8	187.6	183.4
Web pages (n)	Final	890,827	725,532	1,250,594	1,250,584
	Valid (% Valid)	588,632 (66.08)	409,622 (56.46)	1,230,817 (98.42)	1,222,470 (97.75)
	Per Node	855.57	1,369.97	1,312.17	1,333.12
Website focus	Boy (%)	62.8	55.2	–	–
	Hardcore (%)	39.2	57.9	–	–
Hardcore words (n)	Per Node	60,289.85	479,878.96	4,955.50	6,406.58
	Per Page	70.47	350.28	3.78	4.81
Other words (n)	Per Node	121,706.79	331,672.88	8,641.25	11,874.32
	Per Page	142.25	242.10	6.59	8.91
Videos (n)	Per Node	1,967.02	1,058.74	220.14	208.32
	Per Page	2.30	0.77	0.17	0.16
Images (n)	Per Node	9,122.45	16,354.48	1,559.45	2,028.88
	Per Page	10.66	11.94	1.19	1.52

Note: the networks are differentiated based on whether the starting point (or seed) was a blog or a regular website.

Table 2. Summary of t-tests comparing child exploitation networks to each other, and to control networks.

	CE Seed-Blog to CE Seed-Sites	Control Seed-Blogs to Control Seed-Sites	CE Seed-Blogs to Control Seed-Blogs	CE Seed-Sites to Control Seed-Sites
Pages per node	–4.01**	–0.40	–7.67**	0.30
Hardcore words per page	–10.33**	–3.65**	25.15**	14.95**
Videos per node	1.63	0.31	4.30**	2.22*
Videos per page	3.94**	0.48	10.31**	10.14**
Images per node	–2.33*	1.67	7.77**	4.83***
Images per page	–1.10	–0.87	17.87**	17.40**

*P<0.05; **P<0.01

Despite the wide range in hardcore words found, the number of videos per web page and images per web page were relatively low (0.00 to 1.76 and 1.85 to 22.89, respectively). The discrepancies between hardcore words and videos/images might be, in part, due to the nature of seed-site networks. As many of the websites within the site networks were community forum based, few would have had images, and especially videos, embedded within the page. Instead, they would have provided links within the forum post, which would allow a person to obtain the actual content. In addition, one of the starting websites used (Site C) was bit-torrent based.[16] In this situation, we would find the keywords identifying a file (or topic); however, the file would either not be physically located on the website or would require special access (e.g., password). Therefore, it might be that videos and images are just as, or even more, prominent in seed-site networks, however, the content is not as directly present and available as it is with seed-blog networks.

Table 3. Descriptives for five child exploitation seed-blog networks.

		Blog A	Blog B	Blog C	Blog D	Blog E
Nodes (n)	Valid	145	157	163	111	112
	Hardcore (%)	13.1	31.8	30.7	73.0	62.5
	Boy (%)	93.1	91.1	88.3	7.2	1.8
Web pages (n)	Final	250,031	176,829	220,417	109,257	134,293
	Valid	152,987	122,930	158,391	70,915	83,409
	Per node	1,055.08	783.00	971.72	638.87	744.72

[16] Bit-torrent is a peer-to-peer file sharing protocol that allows the user to download parts of a file from multiple people at once. This allows two people with an incomplete file to share the parts they already have with the rest of the network, while they are downloading the remaining parts.

Pages on starting node (n)		5,118	512	142	433	1,315
Nodes connected to by starting node (% of 200 nodes)		109 (54.8)	90 (45.2)	89 (44.7)	6 (3.0)	179 (89.9)
Hardcore words (n)	Per node	48,696	52,228	60,522	70,818	75,827
	Per Page	46.15	66.70	62.28	110.85	101.82
Other words (n)	Per Node	143,342	140,795	180,853	28,3601	73,371
	Per page	135.86	179.82	186.12	44.39	98.52
Videos (n)	Per node	3,395.79	2,097.11	2,736.39	467.41	301.39
	Per page	3.22	2.68	2.82	0.73	0.40
Images (n)	Per node	9,027.46	8,393.59	12,544.65	2,619.67	11,731
	Per page	8.56	10.72	12.91	4.10	15.75

Seed-blog networks were found to be more reciprocal than seed-site networks (18 vs 9 percent), although the reciprocity rates were lower in our study. Despite the difference in reciprocation, the density scores were comparable; about seven percent (seed-sites) to eight percent (seed-blogs) of potential ties were present in the networks.

Table 4. Descriptives for five child exploitation seed-site networks.

		Site A	Site B	Site C	Site D	Site E
Nodes (n)	Valid	162	24	46	36	31
	Hardcore (%)	54.9	100.0	8.7	80.6	87.1
	Boy (%)	92.6	100.0	23.9	2.8	9.7
Web pages (n)	Final	250,154	207,909	87,199	87,199	109,882
	Valid	182,604	116,549	34,011	18.022	58,436
	Per node	1,127.19	4,856.21	739.37	500.61	1,885.03
Pages on starting node (n)		5,197	5,571	1,010	85	263
Nodes connected to by starting node (% of 200 nodes)		18 (9.0)	199 (100.0)	10 (5.0)	23 (11.6)	104 (52.3)
Hardcore words (n)	Per node	63,762.41	3,313,712.75	10,941.57	207,767.72	1,472,330.39
	Per page	56.57	682.37	14.80	415.03	781.06
Other Words	Per node	73,247.96	2,239,022.63	61,786.87	118,565.97	853,448.61
	Per page	64.98	461.06	83.57	236.84	452.75

Videos	Per node	1,532.45	256.92	1,304.63	53.97	5.90
	Per page	1.36	0.05	1.76	0.11	0.00
Images	Per node	4,746.14	111,176.54	2,344.61	1,016.22	42,207.81
	Per page	4.21	22.89	3.17	2.03	1.85

Finding the Key Players

The main objective of this study is to find the key players in online child exploitation networks. The connectivity, severity, and network capital scores for each network are presented in Tables 5 and 6. First note that seed-site networks have much higher network capital scores than seed-blog networks, indicating that there are several very high scoring websites in the smaller seed-site networks. After multiplying by 1000 (to ease interpretation), the network capital scores for seed-blogs ranged from 0.38 to 1.0 (blog E), while for seed-site networks, it ranged from 0.66 to 33.9 (site B). The reason for these differences in ranges might be due to the variation in sizes between the two types of networks, as well as the smaller overall size of seed-site networks: with fewer valid nodes, extreme values have more leverage on network capital. However, this is not necessarily a negative. Extreme nodes are easier to find as they stand out more, as shall be seen below.

Table 5. Network capital descriptives for the five seed-blog child exploitation networks.

		Blog A	**Blog B**	**Blog C**	**Blog D**	**Blog E**
Connectivity score	Mean	0.005	0.008	0.006	0.006	0.006
	Median	0.000	0.001	0.001	0.000	0.000
	Max	0.146	0.205	0.136	0.285	0.230
Severity score	Mean	0.048	0.055	0.064	0.066	0.100
	Median	0.030	0.028	0.035	0.041	0.063
	Max	0.390	0.370	0.380	0.490	0.420
Total network capital (x1000)		0.38	0.41	0.43	0.69	1.00
Removal of top five scores (% change)		0.30 (21.1)	0.36 (12.2)	0.39 (9.3)	0.55 (20.3)	0.91 (9.0)
Removal of random five scores (% change)		0.37 (2.6)	0.39 (4.9)	0.42 (2.3)	0.67 (2.9)	0.94 (6.0)

Table 6. Network capital descriptives for the five seed-site child exploitation networks.

		Site A	**Site B**	**Site C**	**Site D**	**Site E**
Connectivity score	Mean	0.077	0.067	0.004	0.001	0.053
	Median	0.000	0.000	0.000	0.000	0.000
	Max	0.548	2.437	0.552	0.078	1.538
Severity score	Mean	0.077	0.252	0.098	0.072	0.270

	Median	0.040	0.162	0.051	0.030	0.179
	Max	0.630	0.670	0.670	0.820	0.670
Total network capital (x1000)		0.66	33.90	2.53	2.11	19.78
Removal of top five nodes (% change)		0.52 (21.2)	23.8 (29.8)	1.44 (43.1)	0.79 (62.6)	15.01 (24.1)
Removal of random five nodes (% change)		0.60 (9.1)	25.64 (24.4)	2.52 (0.0)	2.01 (4.7)	19.78 (0.0)

Overall, nodes within seed-site networks proved to host more harmful content than when we used a blog as a starting point to crawl the networks. With those general observations in mind, a lot of variation was still found within the networks. Several nodes act as outliers, that is, they have much higher network capital scores than others. These nodes, the key players, should be the focus of law enforcement agencies. Network capital proved to be a useful tool in finding the key players, as it helped highlight nodes that were both highly connected, and were hosting the greatest amount of harmful content given the assumptions of our measure.

The distribution of network capital is illustrated in Figures 1A (blog A network) and 1B (site A).[17] Both figures show a wide distribution in network capital scores, especially for the larger seed-blog network (Figure 1A). The figures also illustrate how the top 10 network capital scores (denoted in red) clearly stand out from the rest of the network. Combined with Table 7 it is clear that even the top five scores (Node ID: 179, 29, 51, 141, and 119) stand out from the next top five (38, 140, 158, 182, and 174). For site A (Figure 1B), the distinction between overall contributions to network capital is still evident; however, it is in three clustering sections with the top two nodes (164 and 63) clearly contributing more than any other nodes, the bottom three nodes clustering together (69, 84, and 71), and the remaining, in the middle group.

[17] Although 10 networks were analyzed, we selected these two networks for representative reasons. The rest of the networks followed the same patterns. Site A was selected to represent the site networks as it was the largest and would provide the clearest picture.

Figure 1A. Network capital for Blog Network A with top 10 scores highlighted in red.

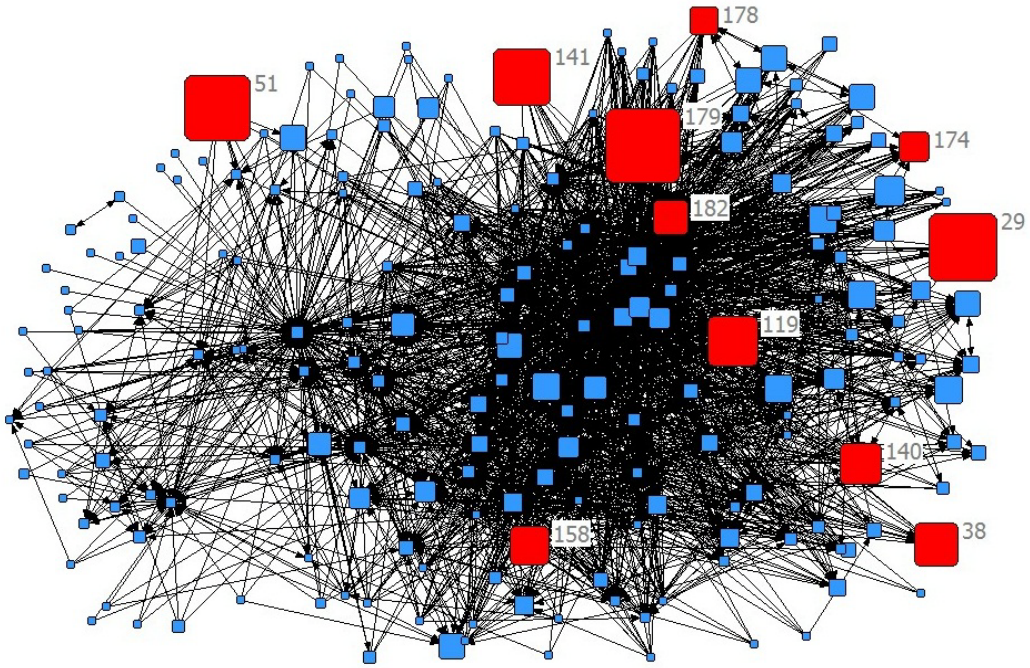


Figure 1B. Network capital for Site Network A with top 10 scores highlighted.

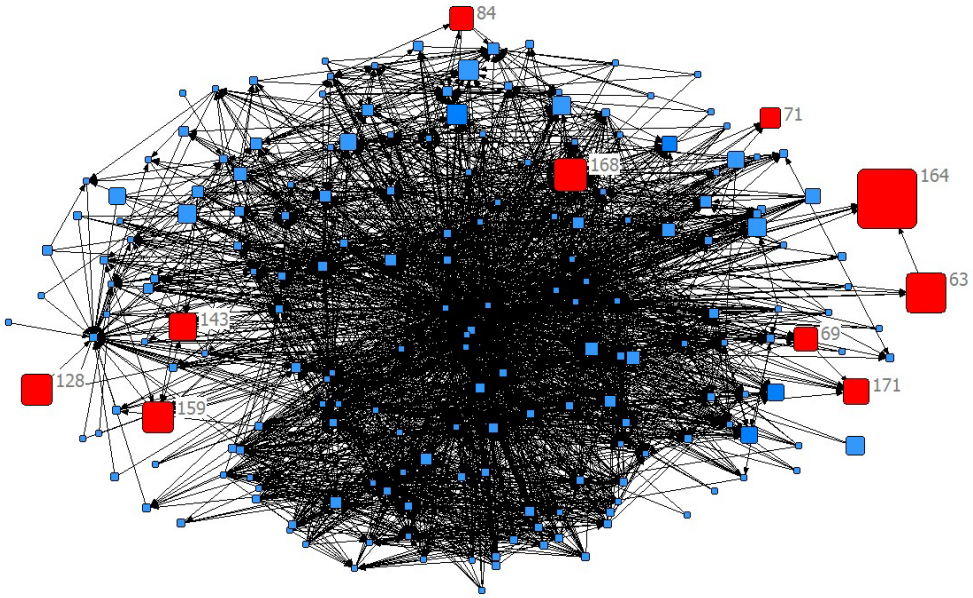


Figure 2A. Severity scores with top five severity scores (red) and connectivity scores (yellow) highlighted for Blog Network A.

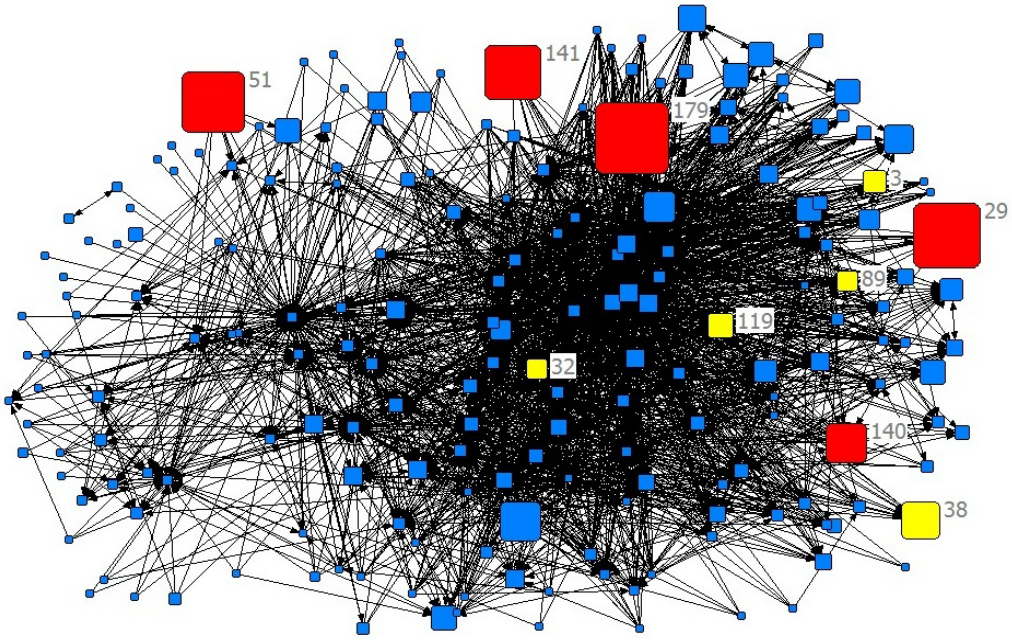
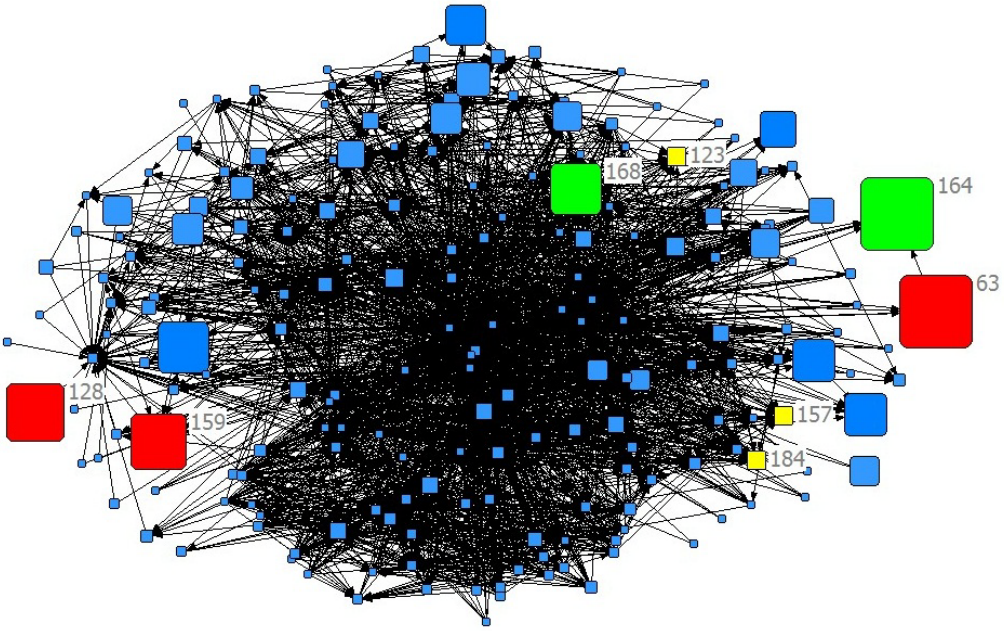


Figure 2B. Severity scores with top five severity (red), connectivity (yellow), and severity + connectivity (green) highlighted for Site Network A.



The importance of using network capital, as opposed to any of the two measures (severity and connectivity) alone, is illustrated in Figures 2A and 2B, and described in Table 7. Figure 2A shows all nodes in seed-blog network A, sized by their content severity score. The five nodes highlighted in yellow represent the top five connectivity scores, while the red nodes represent the top five severity scores. Both Figure 2A and Table 7 show that nodes high in content severity do not necessarily score highly for connectivity. For example, the node with the highest severity score (179) is not in the top ten for connectivity; and vice versa (node 119). In fact, none of the nodes score in the top 10 for both severity and connectivity. Figure 2B provides a similar illustration as 2A, but with seed-site network A. In Table 7, we see that the same node (164) is highest in severity and connectivity. This translates more generally into a stronger association between the two scores for sites, although they do not match exactly. It is entirely plausible that the priorities of law enforcement agencies are such that they might decide to focus on severity only. Our approach suggests that doing so will miss important targets. That is, they will miss the nodes that are the most connected. This is important as the nodes that are the most connected should be able to distribute content the fastest, to the most users.

Table 7. Top ten nodes (by Node ID number) for network capital, severity, and connectivity, for Blog A and Site A networks

BLOG A			SITE A		
Network Capital	**Severity**	**Connectivity**	**Network Capital**	**Severity**	**Connectivity**
179	179	119	164	164	164
29	29	3	63	63	184
51	51	89	168	128	157
141	141	32	128	159	168
119	140	38	159	168	123
38	158	130	143	143	171
140	38	145	171	171	121
158	182	121	69	69	120
182	174	59	84	84	122
174	178	126	71	71	103

Removing Key Players from Networks

One reason for finding the key players in any network is simply to provide a more efficient way to prioritize targets, as is suggested in this study. Another reason is to find the targets that, if removed, can create the most "disruption" in a network. To examine the importance of key players in our networks, we: (1) calculated total network capital scores for the networks as a whole (Tables 5 and 6); (2) removed the top contributors to network capital based on their network capital score;[18] and (3) re-calculated the new network capital score for the entire network and compared it to the initial score. Tables 5 and 6 show that removing the top five network capital contributors from each network, network capital was reduced an average of 14.4 percent for blogs and 36.2 percent for seed-site networks. The reduction was consistent for seed-blog networks, ranging from a 9–21 percent reduction, while seed-site networks varied from 21–63 percent. These reductions seem to be the result of network size, as networks with the lowest number of valid nodes (blogs D and E, and sites B–E) had the most significant reductions in network capital; with the exception of blog E and site E. There was no specific correlation between network reduction and standard deviation, as the websites with the largest standard deviations (blog E, sites B and E) did not have the largest network capital reduction. This suggests that network capital reduction is

[18] The five node removal proceeded in stages. First with the removal of the top contributor, then a re-calculation of the new network to find the new top target, etc. Schwartz and Rouselle argue that multiple combinations of a set of nodes (in our case the top five) need to be calculated to determine which are the best to target (due to redundancy). We only took the top five because connectivity redundancy, amongst the top five nodes, was low. This meant that the most optimal removal combination, within each network, involved removing the top contributor to network capital, followed by the next highest, followed by the next. Note that if redundancy within each network's top five network capital scores had been high, the method of removing the top five contributors to network capital may not have been the most optimal.

more sophisticated than simply taking the outliers within a network. Instead, it is a combination of multiple factors that account for the discrepancies.

Although the removal of the top five contributing nodes to network capital resulted in a significant reduction in most of the networks, is this reduction different to that which results from removing random nodes? Using a random number generator, we randomly selected five nodes from each network to be removed. The results are presented in Tables 5 and 6. With the exception of Blog Network E, selecting the five top network capital scores resulted in at least a four-fold greater percentage decrease in network capital. Although still lower, the difference found in Blog Network E was 50 percent. Like the blog networks, the site networks also had a greater reduction in network capital, when the top contributors were selected over a random sample; however, the amount of reduction varied. For instance, in Site Network B, the percentage change was small (29.8 vs 24.4 percent), while in Site Network C, the change was large (43.1 vs 0.0 percent). Obviously, removing the top five contributors to network capital will result in a greater reduction than five randomly selected nodes; however, the fact that, in most networks, the difference was substantial tells us three things. First, the use of network capital successfully differentiates nodes within a network. Second, that within each network, there are sites that truly stand out compared to others (the key players). Third, that police investigations need to be targeted, and not random, as targeted attacks can have a much greater impact on the overall network.

Discussion

The Internet has provided the social, individual, and technological circumstances needed in order for the production, distribution, and consumption of child pornography to flourish (Taylor and Quayle 2003). This study provides a snapshot of the structure of this online world through an analysis of the networks formed by connections between child exploitation websites. The main goal was to design a method to find, within these networks, the websites that should be prioritized by law enforcement agencies involved in combating child pornography—the key players. These objectives were accomplished by: (1) designing a web crawler to extract online child exploitation networks, and (2) by adapting the measure of "network capital" first created by Schwartz and Rousselle (2009) to the context of child exploitation networks. Instead of focusing strictly on connectivity or exposure within the network, our measure of network capital also took into account the severity of content (i.e, keywords, images, and videos). In doing so, we began to address Krone's (2004) request to incorporate the linkages between websites as well as Taylor, Holland, and Quayle's (2001) work on the severity of website content.

First, the web crawler we designed was found to be suitable in properly identifying child exploitation websites. The clear differences found in hardcore content between child exploitation and non-child exploitation websites increase the validity of our strategy. As will be discussed below, the web crawler still generates false positives, which invites further refinement of the tool. Another limitation is that we currently have no way of verifying whether the images and videos found in those websites always display exploited children. While it could be thought that website owners would find ways to avoid being detected by a web crawler of the type we have used, child exploitation websites turned out being easy to find. Despite the fact that much of the content is illegal, our initial Google searches found countless websites linking to child erotica and child pornography. Previous research by Wolak et al. (2005) found that only 20 percent of online child pornography arrestees used sophisticated tools for hiding illegal content, while Carr (2004) found that only 25 percent did. Therefore, the problem is not just simply websites that host the material, but also the ease with which the material can be obtained: this is because the problem is not simply distribution of the content, but rather its use in lowering the inhibitions, or normalizing the behavior, for prospective victims.

Second, we found that the structure of online networks pertaining to child exploitation were different from networks of websites without that content. More specifically, child exploitative websites had more dead links and pages, as well as higher frequencies of images and photos per page. This is to be expected, as prior research has noted how child exploitation websites exist for the very purpose of exchanging content (Tremblay 2006; Beech et al. 2008).

Despite having two "types" of seed-networks (blogs and sites) it is unclear what percentage of the websites within each network actually coincides with the type of network: this is why we have refrained from interpreting the networks as being composed of "blogs" or "sites".

When comparing the two different types of child exploitation websites based on starting point, it was found that the number of images per page was equal; however, seed-blog networks had more videos per page. It is unclear whether this is the result of the structure of the network (i.e. as influenced by our starting point) or something entirely different. Within blogs, the content is embedded directly onto the page; however, with sites, the content is sometimes linked to indirectly. That is, the site may provide a link to another website (possibly a blog) that hosts the content, or a link to a peer-to-peer program. The small number of networks that could be extracted for this study does not allow us to make a clear statement on whether blogs act more as hosts/producers and sites as distributors.

One of this study's key objectives has been to determine whether a modified version of Schwartz and Rouselle's (2009) network capital measure

could aid in identifying the key players within online child exploitation networks. Our results suggest that it could. We found that websites with harmful content had varying degrees of exposure in the derived networks, and that the measure of network capital was able to discriminate properly between targets that met (or didn't meet) both criteria. This demonstrates the utility of a network approach in target prioritization when targets are in abundant supply, as is the case for child exploitation websites. Note that we adapted the measure network capital from another domain (e.g. organized crime), and that similar adaptations to other fields are both possible, and desirable: the measure has always been intended to be of general use to any social network analysis (Schwartz and Rouselle 2009). The concept of network capital is also worthy of further theoretical refinements (e.g. Wellman and Frank 2001), as it may prove more useful than the more general concept of social capital in certain network contexts.

Finally, given that the websites with the highest severity scores did not necessarily coincide with the websites that were the most connected, emphasizes the need to incorporate both factors into the analysis. As for the reason why the most connected websites were not the websites with the highest severity, the answer is not clear. One possible reason for this is that the most connected websites are the most likely to be discovered. Another possible reason is that some websites may focus on providing content directly, while others might focus on connecting individuals to content. This might be because of personal preference or for legal reasons: those that are not directly hosting the material, but are telling people where they can find it, might feel they are less liable or likely to be apprehended.

Our findings also support the need for targeted attacks on networks. Recall that network capital consists of text, images, videos, and connections to other websites. This means that when targeted tactics are employed, a larger percentage of text, images, videos, and connections to others are impacted—or removed—from the overall network. As has been suggested earlier, the extensive amount of content located on the Internet means that the likelihood of eradicating the problem of child exploitation online is nil. Therefore, steps need to be taken to maximize the current efforts by law enforcement and private organizations. Our findings suggest that the use of a network capital measure can aid in maximizing the impact of efforts to disrupt online child exploitation networks.

In regard to the most optimal removal strategy, although Schwartz and Rouselle (2009) argue that multiple combinations of a set of nodes (in our case the top five) need to be calculated in order to determine which are the best to target (due to redundancy), we only took the top five. The reason for this was that connectivity redundancy, amongst the top five nodes, was low. This meant that the most optimal removal combination, within each network, involved removing the top contributor to network capital, followed by the next highest, followed by

the next. However, if redundancy within each network's top five network capital scores had been high, the method of removing the top five contributors to network capital may not have been the most optimal. For instance, if the top two contributors to network capital had high connectivity between one another, it may be redundant to remove both nodes from the network; or not the most optimal removal strategy. Regardless, the comparison between the removal of the top five contributing nodes (targeted) and the removal of five random nodes (random) provides support for the use of a network capital measurement as an improvement to existing strategies. That is, node targeting appears to be an effective tool in combating online child exploitation.

Research and Policy Implications

A key policy aspect to law enforcement is devising policing methods that reduce cost, improve existing methods, and improve officers' ability to continue doing their job. By creating a web crawler that reduces the amount of hours officers need to spend examining possible child pornography websites, and determining whom to target, we believe that we have touched on each of these key areas of law enforcement policy. An automated process has the added benefit of aiding to keep officers in the department longer, as they would not be subjugated to as much traumatic content (Perez, Jones, Englert, & Sachau, 2010) . However, there are still areas in which what we have proposed here can be improved.

The first step for future research is to further refine the web crawler. In order to develop a tool that can be used by law enforcement, and/or private organizations, several modifications need to be made. Despite being a considerable improvement over a manual analysis of 300,000 web pages, the web crawler's functionality could be improved to allow for efficient analysis of larger, yet still-relevant, networks. Doing so will not simply bring us closer to the true size of the full online child exploitation network, but also, we expect, to some of the more hidden (e.g. password/membership protected) websites. However, it is important to note that there are privacy issues with examining password-protected domains. In Canada, for instance, this type of research would require a warrant to search the protected files. Therefore, despite the desire to investigate these locales, in practice this might be difficult or impossible to achieve.

The current version of the web crawler is designed to find and catalog any target closely related to online child exploitation. As such, it is equally likely to retrieve websites hosting hardcore child pornography content, as well as websites sitting on the fence of child erotica and adult pornography. There are two reasons for this. First, the web crawler was unable to examine membership websites (e.g.,

pay websites and member-based chat forums).[19] Although it is unclear what percentage of World Wide Web content requires a membership for access, Beech et al. (2008) point out that child pornography websites often have self-regulatory methods, such as membership, for excluding individuals. Ergo, the ability to analyze websites that require a membership is clearly an area of need. However, this does not negate the value of researching publicly accessible/open websites for three reasons. First, the websites provide some security for the viewer, as no personal information (e.g., credit card) is required. Second, these websites are probably being used as an initial starting location for most individuals. Third, although child erotica may be viewed as less severe, Wolak, Finkelhor, and Mitchell (2005) found that 79 percent of individuals arrested in the United States for child pornography possession also possessed child erotica images,[20] while only one percent of these arrestees possessed *just* child erotica. This positive association suggests that the individuals visiting websites containing hardcore content are probably also visiting websites containing child erotica content.

The second value of this research is that, although continued refinement of keyword and video content analyses are important, the primary focus of future study improvements will relate to the images. Within the current study, the web crawler totaled the number of images on a website meeting a specific size criterion (greater than 150 by 150 pixels); however, it did not distinguish between the image content. As previously mentioned, much of the law enforcement to date has focused on investigating images; the primary reason being that a "hash function" exists for images, and not for videos. This refers to the mathematical process of taking a large piece of data and reducing it to it into a single "hash value" (Howard 2004), which act as a form of encryption and can be used to authenticate the content of an image (Hoffman 2010). According to Hardy and Kreston (2004), the chances of two files having the same hash value, but different content, is one in 10^{38}. Therefore, utilizing known hash values,[21] instead of just the total number of images, would help improve the validity of our severity

[19] These forums typically exist for people to chat and exchange content outside of the visible website channels (Tremblay 2006).

[20] Child erotica refers to image content that does not contain nudity and is not specifically for sexual purposes. This includes young children in provocative poses, costumes, or form-fitting clothing such as bathing suits.

[21] Currently many law enforcement agencies, the RCMP included, will calculate the hash value for any and all images they encounter, and store both the image, the classification/severity of the image, and the hash information in a database. For any computer/hard-drive/collection-of-files law enforcement encounters during an investigation, they can automatically scan the hard-drive, calculating all hash values for all files, and check each hash value against the database. In this fashion law enforcement can immediately identify the severity of all known images, and manually process only the unknown ones. With the help of this hash database, CENE could similarly identify the severity of all known images as it crawls the web.

measurement.[22] Of course, the development of hash values for videos would be a welcome addition to combat online child pornography.

Finally, as the decentralized nature of the Internet making combating child exploitation difficult, it becomes more important to introduce new methods to address this. This is especially true online, where there is little to no face-to-face contact between users. Therefore, social network analysis measurements, in general, can be of great assistance to law enforcement investigating all forms of online crime—not just online child exploitation. However, an important point to consider is that the individual(s) "running the organization" (i.e. operating the website) may not actually be the key players. As websites allow for contributions by multiple people, the individual operating the website may only act as a broker, while other individuals may actually be the key players on whom law enforcement should focus their attention. In other words, shutting down a website and apprehending the operator may not do much to reduce the problem. This leads to the need to combine social network analysis methods, such as measurements of network capital, with other methods to create a multi-dimensional approach to combat online crime.

[22] Nevertheless, it may still be important to incorporate child erotica images into future analyses. In this scenario, our severity measure would include two image values: known child pornography, and "other".

Appendix. The Child Exploitation Network Extractor Algorithm.

Algorithm CENE(*StartPage, PageLimit, WebsiteLimit, Keywords(), BadWebsites(),minImageWidth, minImageHeight*)	
1:	*Queue*() ← {*StartPage*}
2:	*KeywordsInWebsiteCounter*() ← 0, *LinkFrequency*() ← {}, *WebsitesUsed*() ← {}, *FollowedLinks*() ← {} //initialize variables
3:	**while** |*FollowedPages*| < *PageLimit* **and** |*Queue*| > 0
4:	*P* ← *Queue*(1), D_P ← domain of *P* //start evaluating next page in queue
5:	**if** $D_P \notin$ *WebsitesUsed*() and |*WebsitesUsed*| < *WebsiteLimit* then
6:	*WebsitesUsed*() ← *WebsitesUsed*() + D_P
7:	**if** $D_P \in$ *WebsitesUsed*() and $D_P \notin$ *BadWebsites*() then //evaluate this page
8:	*PageContents* ← Retrieve page *P VideoCounter* ← *0, ImageCounter* ← *0*
9:	*FollowedPages* ← *FollowedPages* + *P*
10:	**if** *PageContents* contains *Keywords()*
11:	*KeywordsInWebsiteCounter*() ← get frequency of all *Keywords*()
12:	*LinksToFollow*() ← all {href} elements in *PageContents*
13:	**for each** *L* **in** *LinksToFollow*() **if** *L* links to an image *ImageContents* ← *retrieve image I //if the link leads to an image* **if** *width(ImageContents)* > *minImageWidth* **and** *height(ImageContents)* > *minImageHeight* **then** *ImageCounter* ← *ImageCounter + 1* *//count only if the image is big enough* **else if** *L* links to a video //*if the link leads to a video VideoCounter* ← *VideoCounter + 1*
14:	**if** *L* ∉ *Queue*() **and** *L* ∉ *FollowedPages* *Queue*() ← *Queue*() + L D_L ← domain of *L*
15:	*LinkFrequency*(D_P, D_L) ← *LinkFrequency*(D_P, D_L) + 1 *VideosInWebsite*(D_P) ← *VideosInWebsite*(D_P) + *VideoCounter* *ImagesInWebsite*(D_P) ← *ImagesInWebsite*(D_P) + *ImageCounter*
16:	*KeywordsInWebsite*(D_P) ← *KeywordsInWebsite*(D_P) + *KeywordsInWebsiteCounter()*
17:	**return** *WebsitesUsed(), KeywordsInWebsite(),LinkFrequency(), VideosInWebsite(), ImagesInWebsite()*

References

Akdeniz, Y. 1999. *Sex on the Net: The Dilemma of Policing Cyberspace*. Reading, UK: Garnet Publishing.

Ali-Hasan, N., and L. Adamic. 2007. Expressing Social Relationships on the Blog Through Links and Comments. ICWSM2007, Boulder, CO.

Beech, A.R., I.A. Elliott, A. Birgden, and D. Findlater. 2008. The Internet and Child Sexual Offending: A Criminological Review. *Aggression and Violent Behavior* 13: 216-228.

Borgatti, S. 2003. The key player problem. In: R.Breiger, K.Carley and P.Pattison (Eds) *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers.* Washington DC: National Academy of Science Press, pp. 241-252.

Borgatti, S. 2006. Identifying Sets of Key Players in a Social Network. *Computational and Mathematic Organization Theory* 12: 21-34.

Durkin, K.F. 1997. Misuse of the Internet by Pedophiles: Implications for Law Enforcement and Probation Practice. *Federal Probation* 61: 14-18.

Easton, S.T., and A.K. Karaivanov. 2009. Understanding Optimal Criminal Networks. *Global Crime* 10: 41-65.

Engeler, E. 2009. September 16. UN Expert: Child Porn on Internet Increases. The Associated Press. http://abcnews.go.com/Technology/wireStory?id=8591118

Frank, R., B.G. Westlake, and M. Bouchard. 2010. The Structure and Content of Online Child Exploitation. Paper presented at the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, D.C.

Furukawa, T., M. Ishizuka, Y. Matsuo, I. Ohmukai, and K. Uchiyama. 2007. Analyzing Reading Behavior by Blog Mining. Proceedings of the 22nd National Conference on Artificial Intelligence (AAA1'07) 2: 1353-1358.

Google. 2010. *Blogger: Terms of Service.* http://www.blogger.com/terms.g

Hardy, R.L., and S.S. Kreston. 2004. *Geeks With Guns, or How I Stopped Worrying and Learned to Love Computer Evidence*. Paper presented at the South African Professional Society on the Abuse of Children National Conference. http://www.sapsac.co.za/geeks.pdf

Hoffman, S. 2010. An Illustration of Hashing and its Effect on Illegal File Content in the Digital Age. *Intellectual Property and Technology Law Journal* 22 (4). https://lawlib.wlu.edu/works/426-1.pdf

Howard, T.E. 2004. Don't Cache Out Your Case: Prosecuting Child Pornography Possession Laws Based on Images Located in Temporary Internet Files. *Berkeley Technology Law Journal* 19: 1227-1273.

INTERPOL. 2010. *Crimes Against Children.* http://www.interpol.int/public/children/default.asp

Izquierdo, L.R., and R.A. Hanneman. 2006. *Introduction to the Formal Analysis of Social Networks Using Mathematica.* http://www.luiz.izquierdo.name

Krone, T. 2004. A Typology of Online Child Pornography Offending. *Trends and Issues in Crime and Criminal Justice* 279: 1-6.

Krone, T. 2005. International Police Operations Against Online Child Pornography. *Trends and Issues in Crime and Criminal Justice* 296: 1-6.

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

Loughlin, J., and A. Taylor-Butts. 2009. Child Luring Through the Internet, 2009. Juristat 29 (1) (Cat No. 85-002-X). Ottawa, ON: Statistics Canada.

Malm, A.E., J.B. Kinney, and N.R. Pollard. 2008. Social Network and Distance Correlates of Criminal Associates Involved in Illicit Drug Production. *Security Journal* 21: 77-94.

McLaughlin, J. 2004. Cyber Child Sex Offender Typology. http://www.ci.keen.nh.us/police/typology.html

Ministry of Labour and Citizens' Services. 2006. *Detecting Pornographic Images on the Network.* Victoria, British Columbia: Information Security Branch: Office of the Chief Information Officer.

Perez, L.M., Jones, J., Englert, D.R., & Sachau, D. (2010). Secondary tramatic stress and burnout among law enforcement investigators exposed to disturbing media images. *Journal of Policing and Cimrinal Psychology, 25,* 113-124. doi: 10.1007/s11896-010-9066-7.

The Office of the Attorney General. 2010. Attorney General Cuomo Announces Groundbreaking Initiative to Enable Social Networking Sites to Eliminate Thousands of Images of Child Pornography. www.ag.ny.gov

Schwartz, D.M., and T. Rouselle. 2009. Using Social Network Analysis to Target Criminal Networks. *Trends in Organized Crime* 12: 188-207.

Taylor, M. 1999. The Nature and Dimensions of Child Pornography on the Internet. Paper presented at the Combating Child Pornography on the Internet. http://www.stop-childpornog.at

Taylor, M., G. Holland, and E. Quayle. 2001. Typology of Paedophile Picture Collections. *The Police Journal* 74: 97-107.

Technorati. 2010. Sports blogs. http://technorati.com/blogs/directory/sports

Top Sites Blog. 2010. Top 11 Most Popular Sports Websites. http://topsitesblog.com/best-sports-websites/

Tremblay, P. 2006. Convergence Settings for Nonpredatory 'Boy Lovers'. In: R.Wortley and S.Smallbone (Eds) *Situational prevention of child sexual abuse*. Monsey, New York: Criminal Justice Press, pp.145-168.

Wellman, B., and K. Frank. 2001. Network Capital in a Multi-Level World: Getting Support in Personal Communities. In: N.Lin, K.Cook and R.Burth (Eds) *Social Capital: Theory and Research*. Chicago: Aldine DeGruyter, pp. 233-273.