Analyzing Online Child Exploitation Networks: An Examination of Severity and Connectivity

by

Bryce Garreth Westlake
B.A. (Psyc/Soci), University of British Columbia, 2007

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Arts

In the
School of Criminology

SIMON FRASER UNIVERSITY

Spring 2011

APPROVAL

Name: **Bryce Garreth Westlake**

Degree: **Master of Arts**

Title of Essays: **Analyzing online child exploitation networks: An examination of severity and connectivity**

Examining Committee:

Chair: **Neil Boyd, LL.M.**

Martin Bouchard, Ph.D.
Senior Supervisor

Eric Beauregard, Ph.D.
Supervisor

William Glackman, Ph.D.
Supervisor

Anthony Beech, Ph.D.
External Examiner
University of Birmingham

Date Defended/Approved: 18 April 2011

Declaration of
Partial Copyright Licence

ABSTRACT

The continued growth of child pornography distribution online has resulted in the need for new innovative tools to combat the problem. Since shutting down all child exploitation (CE) websites is arguably impossible, the goal must be to find the most efficient way of identifying the key targets and then apprehend them. Using a web-crawler we specifically designed for extracting CE networks, we 1) examined the structure of ten CE networks and compared them to a control group of sports-related networks, and 2) provided a measure (network capital) that allowed for better identification of the most important targets, within each network, for law enforcement purposes. Results show that network capital –a combination of content severity (images, videos, and text) and connectivity (links to other child pornography websites) – was a more reliable measure of target prioritization than traditional methods currently being used. Implications for future research and law enforcement practices are discussed.

Keywords: Child Exploitation; Internet; Social Network Analysis

DEDICATION

To my lovely and supportive wife Shannon, without whom I would not have driven, persevered, and completed my education.

To my grandmother who passed away during the writing of this thesis. Her influences on my life are great and her memory will not be lost.

ACKNOWLEDGEMENTS

Thank you to Dr. Martin Bouchard who saw the potential and desire that I had in this topic, as well as school in general, and worked hard to push me to reach and go beyond my potential.

Thank you to Dr. Delroy Paulhus who gave me the opportunity to conduct research as an undergraduate and provided me with the research experience needed to get in to, and be successful in, graduate school. To this day, you are still a person that I look to within my academic career.

Thank you to my committee for their support. Dr. William Glackman, for his assistance with obtaining a senior supervisor. Dr. Eric Beauregard, for his support and editing of publications leading up to this thesis as well as the thesis itself. Dr. Anthony Beech, whose work helped provide the foundation for the research that I conducted and wrote about in this thesis.

Finally, I want to thank my family, especially my Mother and Sister, who have always been a strong support system and believed in my abilities to achieve what I have thus far, well before I had the confidence in myself.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1: INTRODUCTION

The Internet has revolutionized the way that we interact with others, as well as changed how we conduct business. There are just under two billion individuals worldwide using the Internet, with 266 million users being from North America (Internet World Stats, 2010). Of the two billion users, adolescents and college students make up the largest proportion (Gross, 2004; Mitchell, Finkelhor, & Wolak, 2007; Technology Quick Response Team, 2005). In Canada, Sinclair and Sugar (2005) state that 64 percent of households have access, with British Columbia being the highest at 71 percent. However, through access at home and at school, 90 percent of youth regularly use the Internet (Frontline, 2008). As much as the Internet 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 it quickly becoming a breeding ground for illegal activities. One such activity that continues to grow is child exploitation: commonly through the distribution of child pornography. In the current study, child pornography will refer to the actual content itself, while child exploitation will refer to the overall phenomenon that includes child pornography.

Currently a global definition of child pornography does not exist (Sinclair & Sugar, 2005). In fact, there is controversy as to whether the term *pornography* should even be used to describe what is actually sexual abuse (Carr, 2004b; Taylor & Quayle, 2003). Nevertheless, there is some consensus on the type of material that would be classified as child pornography; or child sexual abuse, depending on the overarching definition used. However, there is still ambiguity and disagreement regarding artificial material made using technology (Jensen, 2004). In Canada, section 163.1 of the Criminal Code (1985) covers access, possession, production, distribution and selling of child pornography and defines child pornography to include all visual representations, whether made by mechanical or electronic means. This includes material manufactured that does not involve real children.

Previous research has found that the most searched topic on the Internet is 'sex' (Cooper, 1998; Freeman-Longo & Blanchard, 1998). Although the vast majority of individuals who use the Internet for sexual pursuits do so in a safe and legal way (Cooper, Scherer, Boies, & Gordon, 1999; Griffiths, 2000), the anonymity of the Internet has resulted in a growing percentage who sexually solicit youth (Mitchell, Finkelhor, & Wolak, 2003). Due to the unregulated and seemingly anonymous nature of the Internet, online pedophilic networks have flourished (Akdeniz, 1999). Durkin (1997) outlines three ways that sex offenders have been able to utilize the Internet: 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 correspondences with other pedophiles through chat rooms and newsgroups. Chat rooms and newsgroups are also used to misrepresent oneself as a youth, for the purpose of sexual communication with children. The current study focuses primarily on the first, dissemination, and indirectly on the second, social networking. What makes the problem of child exploitation worse is the ease with which one can obtain illegal pornographic material (Spink, Ozmutlu, & Lorence, 2004; Young et al., 2000). Searching the words 'boy', 'teen', or 'child', brings up countless websites and photos of youth in sexually exploitive roles (Mitchell, et. al., 2007; Young, 2005).

As of 2009, the United Nations estimates that there are more than four million websites containing child pornography and that 35% of the websites depict serious sexual assault, while 70% involve children under the age of eight (Engeler, 2009). Much of the existing efforts to curb online child exploitation have taken the form of Internet chat room stings and injunctions against online groups seen to be facilitating the proliferation of child sexual abuse (e.g., *North American Man-Boy Love Association*, *Pedophile Information Network*, *Freespirit* and *BoyChat*). At times, this process has come against opposition from those who

argue Internet stings are a form of entrapment[1] (Fulda, 2005). In addition, website owners often find loopholes, arguing that their websites are merely support forums that do not host exploitative material and that they cannot be held responsible for the private messages people send back and forth, that may or may not contain information on obtaining illegal material[2].

As online child exploitation is seen as a global issue, the United Nation's International Criminal Police Organization (INTERPOL) has taken a leading role in addressing the problem (INTERPOL, 2010). One of the key ways that INTERPOL has aided is through the creation of the International Child Sexual Exploitation image database, which contains all known sexually explicit photos of children. The database is used by various police organizations to aid in identifying children being abused. Additionally, INTERPOL collaborates with the Internet Related Child Abuse Material Project and the Virtual Global Taskforce to help coordinate multi-country investigations and spread awareness of the problem.

Despite large investments by global governments and private organizations[3], online child exploitation is nowhere near under control. In fact,

[1] One such example is the FBI posting fake links to explicit images of children and then raiding the homes of those who clicked on the links (McCullagh, 2008).

[2] For instance, one of the most well know sites 'Free Spirits' state that "the sites linked from these pages are operated by private citizens exercising their right to free speech under the U.S. Constitution and Universal International Human Rights Convention" (Hooked,2001).

[3] This includes organizations specifically examining child exploitation, such as the National Child

McLaughlin (2004) estimates that less than one percent of all virtual pedophiles are apprehended. This is not to say that existing efforts have been futile. Image databases, such as United States Division of Criminal Justice Services Database[4], International Child Sexual Exploitation Image Database (ICSE-DB)[5], and the National Child Victim Identification Program (NCVIP)[6], have aided in the identification of child pornography websites, and the rescuing of children currently being victimized. For example, in 2001, a thirteen-country operation, organized by the British National Crime Squad, resulted in the arrest of 107 suspected members of the *Wonderland Club*; the largest Internet pedophile ring (Reuters, 2001). This resulted in the conviction of seven individuals and the confiscation of 750,000 images and 1,800 videos, containing 1,263 identifiable children[7]. In January 2011, a sixty-four-year-old British Columbia man was arrested for possession of child pornography. This resulted in the seizing of ten computers, several digital and web cameras, thirty hard drives, twenty-five other

---

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©.

[4] The database consists of over 8,000 images that are used to aid in the detection and removal of child pornography on social networking sites such as Facebook© and MySpace© (The Office of the Attorney General, 2010).

[5] Created in March of 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).

[6] 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).

[7] The children in the images and videos ranged from three months to sixteen years. The majority were under the age of ten with many being two or three-years-old.

storage devices, and multiple cds and dvds. In total around one million images of children were confiscated (O'Brien, 2011). In September of 2010, a sixteen-year-old girl was videotaped and photographed being gang-raped in Pitt Meadows, British Columbia. The photos were subsequently placed on the Internet and circulated via social media websites and word of mouth. A sixteen-year-old boy and a nineteen-year-old boy were both charged with producing and distributing child pornography (Saltman, 2011). Despite these successes, the prevalence of these types of stories in the media never seems to decrease.

As previously noted, there is no global consensus regarding what constitutes child pornography. Although the crimes of child exploitation cross international borders, the varying laws, in different countries, result in cross-national investigations being difficult. However, there have been several high profile operations[8], such as Avalanche, Blue Orchid, and Cathedral, which have resulted in the apprehension of some of the largest and most influential players of the time (see Krone, 2005 for summary of international police operations). Cross-national investigations tend to be successful when there are similarities in laws regarding child pornography; namely law congruencies in Australia, the United Kingdom, and Canada. Recently, in an investigation that lasted almost one year and involved the social media website Facebook©, eleven individuals total, three

[8] The most recent operation, Operation Rescue, resulted in the removal of well-known boy-love website boylover.net. Although still expected to rise, to-date there have been 184 arrests and 230 children rescued -the highest ever in any investigation (The Hague, 2011).

in Australia, six in the United Kingdom, and two in Canada, were arrested for child exploitation (Cross, 2010). However, even in this example, although the investigation was joint, each country dealt with their own offenders differently and according to their own national laws.

As the growth of online child pornography outpaces law enforcement resources, the primary goal going forward must be to improve the existing strategies and methods for combating the problem. With so many websites containing child sexual abuse images (and videos), and the limited resources available to various organizations to combat the problem, there needs to be continued efforts to automate and simplify the process of selecting and prioritizing targets for the purpose of criminal investigation. With the cessation of online child exploitation unlikely, the focus needs to be on the *severity* and *exposure* of the content rather than simply the *presence* of the content. Therefore, within the current study we look to address the following three issues. 1) Improve on the existing technology being used by law enforcement through the creation of a web-crawling tool that can automatically search websites for child pornography, thereby reducing the amount of time and contact needed to carry out manual searches. 2) Explore the structure of child exploitation networks using social network analysis tools and see how they differ from other types of online networks. 3) Find the best method to identify the key players within child exploitation networks, through the creation of a formula that weights both the

content (words, images, videos) as well as the connections to others (links out of a website).

2: LITERATURE

2.1 Child Exploitation Online

The growth of the Internet has resulted in a substantial increase in research aimed at understanding various online networks (e.g., Garton, Haythornthwaite, & Wellman, 1997; Kempe, Kleinberg, & Tardos, 2003; Smith & Kollock, 1999; Wellman et al., 1996). However, most of the research to date has focused on social networking sites such as Facebook© and MySpace©, and has stopped short of investigating the structure of child exploitation networks. This is despite the online child pornography business grossing more than three billion (American) dollar each year (TopTenREVIEWS, 2004).

Prior to the Internet, child exploitation could have been viewed as more of a solitary crime with very sparse networks (Beech, Elliot, Birgden, & Findlater, 2008). Although images and videos were transferred through the mail, the speed of the exchange was low and the chance of getting caught sending material was high. More importantly, it was difficult for people to get in contact with one another. However, the advent of the Internet has changed the crime of child exploitation and sexual abuse (Durkin, 1997; Tremblay, 2006). The anonymity of the Internet coupled with the ability to find like-minded individuals has made the

Internet a popular medium through which collectors and traders of child pornography can congregate (Akdeniz, 1999; Durkin, 1997; Taylor & Quayle, 2003). With websites such as *Bliss* and *Rene Guyon Society* outwardly supporting relationships between young children and adults, the ease at which material can be obtained and shared has grown exponentially. In fact, according to the National Society for the Prevention of Cruelty to Children, each week, more than 20,000 child exploitive images are added to the Internet (Frith, 2003).

2.1.1 Child Pornography under Canadian Law

In Canada, child pornography is covered under section 163.1 (corrupting morals) of the Canadian Criminal Code (1985). The law covers any visual representation (e.g., photographic and film) of any individual under the age of eighteen depicted in a sexual activity or whose sexual organ or anal region is the dominant focus. The law also includes any description, presentation or representation of such acts through written or auditory means. In the highly cited 2001 case of *R. v. Sharpe*, Chief Justice McLachlin modified the definition of child pornography in Canadian law. McLachlin stated that child pornography excluded any written, visual, or auditory representation of sexual activity, if it were created for personal use, and was exclusive to the creator (no distribution). In 2002, the Canadian Criminal Code was again amended to directly address the proliferation of child pornography online (Media Awareness Network, 2010).

The word 'transmission' was incorporated into the definition of distribution to include email and postings on websites. Also, intentionally accessing child pornography through an Internet browser was added.

In addition to the amendments to section 163.1, section 172.1 was added to the Criminal Code (Bill C-15A, 2002). This section criminalized the online luring of any person who was *believed*[9] to be under the age of 18, for the purposes of sexual exploitation[10]. If found guilty, the accused is subject to a fine of $2 000 and/or six months imprisonment; this increases to five years in prison for an indictment. Currently, accessing or possessing child pornography carries a sentence of up to five years, while distributing and/or making child pornography is subject to a maximum sentence of ten years. However, with the push by the Canadian government to provide all schools with Internet access Sinclair and Sugar (2005) argue that there is not enough being done to provide a safe environment online. Although it is unclear how many children are being victimized online, the statistics regarding exposure to pornography and sexual solicitation is high (Ropelato, n.d.). In fact, almost half (48%) of youth aged eight to thirteen reported visiting a website with adult content (Greenfield, 2004), while

[9] In R. v. Alicandro (2009), Alicandro was convicted under s. 172.1 for online child luring, despite claiming that he believed the undercover officer posing as a thirteen-year-old girl was over the age of eighteen. This set the precedent that the perceived age of the victim was irrelevant.

[10] In R. v. Legare (2009), it was determined that online luring did not need to involve an attempt to lure the child into a face-to-face meeting.

Mitchell, Wolak, and Finkelhor (2008) found that 53% of youth online experienced sexual harassment or solicitation.

2.1.2 Combating Online Child Exploitation

The fight against child exploitation is a strongly advocated issue with plenty of support; however, it is an enormous undertaking that takes countless hours and resources that are not necessarily available. For instance, in Britain it was found that two percent of identified commercial child pornography sites were still in operation over a year after they were first detected (Johnson, 2008). Although two percent might not seem like much, one could argue that once any child pornography website is found, law enforcement should be able to shut it down immediately.

In Canada, Sinclair and Sugar (2005) found that of the 555 reports received by Cybertip, from September 26th, 2002 to September 25th, 2003, 428 involved online sexual exploitation of a minor. Of the 428 cases, 187 were forwarded to law enforcement, which resulted in 39 websites in Canada, and 61 elsewhere, being shutdown. In the United States, Wolak, Finkelhor, and Mitchell (2005) noted that during a twelve month period –July 1st, 2000 to June 30th, 2001- CyberTipline received over 19,000 reports of child pornography. Over that same period, there were approximately 1,700 arrests for Internet child pornography possession. The discrepancy between the number of reports and the number of

follow-up investigations, both in Canada and the United States, support the claim that law enforcement resources are strained and are unable to keep up with the amount of reports of child exploitation that they receive. This emphasizes the need to improve investigation techniques that reduce the strain on police resources.

Understanding the immense number of hours and resources that go into finding sexually explicit material online, two well-known Internet companies, Google© and Microsoft©, have created software programs to aid in detecting child pornography. In conjunction with the National Center for Missing and Exploited Children (NCMEC), Google© adapted a pattern recognition program, originally used on Youtube© to detect copyrighted material, to aid in searching through files and detecting child pornographic images (Shiels, 2008). This program has aided the NCMEC "in organizing and indexing…information so that analysts can both deal with new images and videos more efficiently and also reference historical material more effectively" (Baluja, 2008, para. 6). Meanwhile Microsoft©, while also working with NCMEC, has created a similar program known as PhotoDNA (Microsoft, 2009). PhotoDNA has aided in two key ways. First, it has been able to detect modified versions of known child pornography images. Second, it has been able to analyze large quantities of potential images in a short period of time.

In Canada, the most successful tool has been The Child Exploitation Tracking System (CETS). Officially launched on April 7, 2005, CETS was developed jointly by Microsoft Canada, the Royal Canadian Mounted Police (RCMP) and the Toronto Police Service. Using CETS, police agencies can manage and analyze huge volumes of information in powerful new ways, such as cross-referencing obscure data relationships and using social-network analysis to identify communities of offenders (Microsoft, 2005). The program contains the ability to permit investigators to easily import, organize, analyze, share, and search information from the point of detection right through the investigative phase. As of late 2006, Microsoft had contributed more than seven-million-dollars to the development of CETS. In June of 2007, Microsoft released CETS 2.0 (Microsoft, n.d.). Although the continual support of companies such as Google© and Microsoft© are vital, technology is continually advancing and as a result, online child pornography distributors are finding new ways to both hide and exchange content. Therefore, there is a need to be continually improving on existing technology and discovering innovative ways for combating the problem.

In addition to finding content being distributed around the Internet, improving detection tools can aid in finding children who are currently being abused. Wolak et al. (2005) found that one in six investigations, starting with child pornography allegations, resulted in dual offending charges. That is, of those investigated for possessing child pornography, one in six was also abusing a

child offline. Therefore, the need for tools that improve the ability to detect content being distributed online is not only about the physical content, but also about discovering cases of children currently being abused. Thus, improving the tools that are used by law enforcement can aid in achieving two important goals: removing content from the Internet and rescuing children being abused.

Despite software tools such as CETS, there are still plenty of gaps in detection. Wolak et al. (2005) note that there needs to be an increase in the training and resources made available to conduct these investigations as well as improvements to computer forensic tools. Additionally, Krone (2004) notes that although image detection is important, perhaps the most important factor in law enforcement is the reliance on networks by many offenders. Krone suggests that there needs to be more attention paid to the linkages between users and websites, as these linkages are just as important as the content being distributed. The current research looks to address Wolak et al. call for improvements to detection tools as well as Krone's request for the inclusion of linkages in analyses.

2.1.3 The Online Offender

Although a large percentage of online offenders are white, single men, over the age of twenty-five (Burke, Sowerbutts, Blundell, & Sherry, 2001; Loughlin & Taylor-Butts, 2009), Wolak et al. (2005) found that offenders vary greatly in regards to socio-economic status and education. In their National

Juvenile Online Victimization Study (NJOVS), Wolak et al. found that only eleven percent of offenders had prior arrests for sex-related crimes. Together these findings point to a need to profile the content and connections more than the type of offender as it would appear that offenders differ greatly in their demographic make-up.

When it comes to image content, according to the NJOVS, one-third of those arrested were categorized as distributors of child pornography[11] (Wolak et al, 2005). Overall, fourteen percent of arrestees had more than 1,000 images, nineteen percent possessed images of children aged three or younger, while thirty-nine percent had images of children aged three to five. Eighty percent of offenders had pictures involving sexual penetration, while 71 percent had images depicting sexual activities involving adults and children. Summarily these statistics show that the range in both quantity and content of the images varies greatly and therefore, efforts need to be made to find offenders that are high in both. Although 79 percent of those arrested possessed both softcore (non-nude) and hardcore content, only one percent of arrestees possessed strictly softcore content. Again, this finding emphasizes the need to investigate a range of content as one typically leads to the other. Finally, only fifteen percent of offenders had

[11] Distributors were more likely to have large collections (more than 1,000 images) and images of children under the age of six.

photos of both girls and boys. This suggests that offender preferences are mostly specialized to one sex and therefore investigations need to be tailored accordingly.

Although images are very common, video content is also an issue. Within their study, Wolak et al. (2005) found that 39 percent of those arrested possessed both images and videos of children involved in sexual activities. As previously discussed, much of the police efforts and technology to-date have focused on image content. Despite the importance of images, Wolak et al. finding suggests that videos are similarly important and need to be incorporated into future detection tools.

Within Wolak et al. (2005) study, there were some positive findings. First, of the over 1,700 offenders arrested, none were acquitted and most plead guilty immediately. This finding suggests that the key to future enforcement is to work on improving our ability to find perpetrators, as once they are found they tend to plead guilty to the charges. This goal can be better achieved through an automation process of finding child pornography, instead of officers needing to manually search for the content. In other words, automation can free up time for law enforcement agents to focus on apprehending perpetrators rather than finding the content. Second, only twenty percent of offenders used any sort of method to hide their collections, while only ten percent were classified as extremely knowledgeable computer users. That is, many users do not use any sophisticated method of protecting their collections. Therefore, elaborate tools that can break

through various barriers or encryptions, used by an offender, do not need to be the priority. They are important, of course, however, for a large portion of offenders, these types of tools are unnecessary. These two findings suggest that the key to apprehending those proliferating child pornography is to create simple tools that can analyze large amounts of material/data over short periods of times; like PhotoDNA.

2.2 Social Network Analysis

Social Network Analysis (SNA) is a method of examination that looks at the "relational patterns of nodes (actors) and connections (ties) based on mathematical computations" (van Hulst, 2009, pg. 103). Primarily derived from the mathematical analysis known as graph theory, SNA is common in many research fields (Wasserman, Faust, Iacobucci, & Granovetter, 1994). In business SNA is used in marketing (Kempe, Kleinberg, & Tardos, 2003), organizational behaviour (Balkundi & Kilduff, 2006), and, more recently, international relations (Hafner-Burton, Kahler, & Montgomery, 2009). SNA is also commonly used in health research, to examine the networks of infections (Ferguson & Garnett, 2000; Keeling & Eames, 2005). Morselli (2009) notes that it should not be looked at as a theoretical framework but instead an analytical framework that allows for the

analysis of social interaction themes[12]. In other words, it should be looked at as a tool for presenting the correlation between various individuals (or things) within a network, and not a causal explanation of the relationship or the network structure.

Overall, SNA has begun to play a more prominent role in criminological research, through the expansion to areas such as violence, police socialization, and co-offending patterns (Easton & Karaivanov, 2009; Morselli, 2009). However, SNA is most commonly used to examine the network structure of gangs (McGloin, 2005; Papachristos, 2009), the drug trade (Malm & Bichler, in press; Malm, Kinney, & Pollard, 2008; Morselli, 2001; Natarajan, 2006) and terrorist organizations (Krebs, 2002; Medina & Hepner, 2008; Xu & Chen, 2008). Through these and other studies, SNA has been shown to be a valuable tool for criminologists and law enforcement for the purposes of scenario building, risk assessment, hypothesis testing, network destabilization, identifying aliases, supporting decisions for law enforcement resources, and evidence for prosecution (van Hulst, 2009). When it comes to network destabilization, SNA is typically used to determine who the key players within a network are, and whose removal would result in the greatest fragmentation (or destabilization) of the overall network (Borgatti, 2006; Malm & Bichler, in press). Carley (2006) argues that

[12] Borgatti & Lopez-Kidwell (in press) counters this argument by stating that social network analysis can be looked at as a theoretical framework and that it can be broken down into two different types: bond and flow.

this destabilization can be achieved through the control of financial, informational, or personnel flow both in and out of the network.

Although destabilizing networks is important, *who* is identified as a key player can be greatly influenced by the size of the network being analyzed. More generally, network properties can change extensively, both qualitatively and quantitatively, depending on the size of the network (Butts, 2009). For example, with Internet networks, invalid or 'dead' websites can severely affect the size of the network. Coinciding with size, Sparrow (1991) points out that fuzzy network boundaries can also influence SNA findings. More specifically, it is difficult to determine where a network starts and ends and who should be excluded and included within the network.

When relating this to the Internet and child exploitation, the question can be raised whether websites that do not contain child pornography, but link to such content, should be included or excluded from the network? Although they are not directly involved in the distribution of the content, they are aiding in its distribution through their association with the website(s). When it comes to child exploitation, the best example may be the website *Freespirts*, which does not contain any (known) child pornography but supports those, and hosts some, that do engage in child exploitive activities.

2.2.1 Determining the Key Players in Online Child Exploitation Networks

When examining a network from a criminological perspective, one of the most important components is the key players. This is because targeting specific individuals will result in a greater reduction in crime compared to simply targeting the most chronic offenders, or someone at random (Easton & Karaivanov, 2009). However, defining what constitutes a key player is far more difficult than it may first appear (Borgatti, 2003; Schwartz & Rouselle, 2009). Carley and Hill (2001) point out that when determining key players, the network is comprised of more than just node-to-node interactions. Instead, attributes such as what each node knows, how information is distributed, and how assignments within the network are dispersed have to be taken into consideration. More specifically, the relationships between nodes are as important as the connections themselves (van Hulst, 2009). Easton and Karaivanov note that criminal networks are structured by the participants themselves and that awareness of law enforcement techniques may alter the structure of the network. Therefore, as previously discussed, it is important to understand that the networks are dynamic, not fixed, and thus key player identification is more than just identifying who is the most central to the network, as the network will typically adapt to law enforcement policies (Carley, 2006). Thus, initiatives for finding key players need

to be tailored to the specific properties of the given network (Malm & Bichler, in press).

Another reason why identifying key players in criminal networks is difficult is because law enforcement typically has multiple criteria to identify the key players they need to target. For example, Borgatti (2006) notes law enforcement typically has two primary goals when investigating criminal networks. The first goal is disruption, whereby law enforcement targets the most connected players within the network. The three most common SNA ways to determine connectivity are degree, closeness, and betweenness centrality measures (Freeman, 1979). Degree centrality refers to the number of direct connections (ties) a player (node) has with all others within the 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 for one node to act as a broker between two other nodes.

There are two issues with traditional centrality measures being used to identify the key players within a network: 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 betweeness 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. In other words, centrality measures are not optimized for finding key players (Borgatti, 2006). 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 one of them has the same impact on the overall network.

The second goal of law enforcement is intelligence collection, whereby law enforcement targets players in order to maximize its knowledge about the overall network (Schwartz & Rouselle, 2009). The most common SNA methods for determining this is through degree and closeness centrality measures. However, Borgatti (2006) argues that degree centrality is only useful when examining direct connections between nodes. Although closeness centrality sums the minimum total 'steps' that a specific node needs to travel through to reach every other node in the network, Borgatti states that this analysis can be misleading. As information is passed through people, it becomes degraded and some of that information becomes lost. 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. Timely information can provide law enforcement

with a better opportunity for disruption. For child exploitation, as for many other crimes, timeliness (or lack thereof) has great consequences.

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 website is key to the circulation of information and content across the network. In fact, it is one of the main raison d'être of child exploitation networks. As previously noted, much of the efforts to address online child pornography have focused on the presence of known images; however, Krone (2004) suggests that the focus of law enforcement needs to also be on the linkages between websites and the offender's 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 study, we measure connectivity by the linkages between websites.

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, Holland, and Quayle (2001) argue that the severity of a child pornography collection is comprised of three components: size of collection, presence of new and/or private material, and age of the children depicted. Although the number of images, and the content within them, are very important, it is also important to examine other factors. Krone (2004) points out that the engagement an individual has with the material is also important. For the purposes of online collections, this can take the form of the descriptions (text) posted with the image. In the current study, we propose a measure of severity that takes text, images, and videos into account, which emphasizes Taylor et al. first component of severity: size of collection.

2.3 Current Study

As the problem of child pornography online continues to grow, and law enforcement resources are further strained, it becomes more imperative that resources be allocated in the most efficient manner. Durkin (1997) states that the Internet has aided sex offenders in three key ways: dissemination, social networking, and sexual communications with minors. The current study focuses on improving law enforcement strategies while examining the phenomena of dissemination and social networking. It is argued 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. The 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 SNA, we develop a measure to identify the key players in online child exploitation networks that focuses on both *severity* (how harmful is the content) and *connectivity* (how exposed and easy to find is the content).

As previously discussed, there are numerous technological products that aid in finding child pornography online; however, one of the key issues with these are 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 as 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). 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 any 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. More specifically, this is done by looking at how blog and site child

exploitation networks differ from one another as well as how they differ from non-child exploitation networks.

Finally, the vulnerability of child exploitation networks will be explored utilizing the various vulnerability measures common in SNA. The purpose of this is to determine whether shutting down the websites leads to the fragmentation of the overall network, or whether it does little to disrupt the flow of material. Our objective is to uncover the structure of online child porn networks, and to identify their 'hardcore key players': websites whose removal would result in the greatest reduction in network capital (NC). 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 highly connected websites that also display the most harmful content. Although it would be naïve to suggest that online child exploitation could be completely eradicated, SNA provides a means of understanding the structure and vulnerability of online networks. In turn, this could greatly improve the effectiveness of law enforcement.

3: DATA AND METHODS

The objectives of this study are as follows. First, to create a web-crawler that could be used to map online child exploitation networks. Second, to understand how child exploitation networks differ from non-child exploitation networks. Third, to determine how child exploitation blog networks differ from site networks. Finally, to create a measurement of NC, that incorporates both the severity of the content found on a webpage with the connections to other webpages, which can be used to better target the key players within online child exploitation networks. We propose a method to undertake this analysis by extracting networks of websites, and their features, then creating measures to determine the severity of content on each website and its importance within the network. Finally, we remove the top contributors to NC in order to see how it affects overall NC.

3.1 Network Structure

3.1.1 Child Exploitation Network Extractor (CENE)

We use a custom-written web-crawler called Child Exploitation Network Extractor (CENE)[13]. CENE starts the crawling process at a user-specified webpage, retrieves the page, analyzes it, and recursively follows the links out of the page. During this process, in order to construct a coherent network for analysis, the web-crawler establishes the links between websites and collects statistics on the type of content on the webpages hosted on that website. The crawling process is not random, but is done according to rules, and terminates when certain user-specified criteria are met. The algorithm we designed to do this is described in Figure 1. A variety of starting locations were used to extract multiple networks for comparison purposes. For each network extracted, features were collected about the content of the pages and the links between them. The statistics were then aggregated up to the website level. For example the features for www.website.com were calculated from the statistics collected from all pages on that website.

A few conditions were used to keep the network manageable in size and relevance. Since the Internet is extremely large and a crawler would most likely never stop crawling, we had to implement limits into CENE in two ways. First, to

[13] CENE was created by Dr. Richard Frank, a computing scientist, while I acted as a consultant for the crawler guiding and keywords.

keep the network extraction time bounded, a limit was put on the number of pages retrieved (*PageLimit* – line 3). For this study, the limit was set to 250,000 pages. Second, the network size was fixed at a specific number of websites (*WebsiteLimit* – line 5). For this study, the limit was set to 200 websites to make the networks both meaningful and analyzable[14]. This was done for two reasons. First, to ensure that the networks were focused on websites dealing only with the specified topic and second, to mimic possible patterns that a typical 'user' might go through, when trying to find child pornography. The end result of this process was a network where all the websites in the network were sampled approximately equally, with ($^{PageLimit}/_{WebsiteLimit}$) pages being sampled per website. Despite including 200 websites and 250,000 webpages, the networks collected during this study should not be viewed as complete networks, but rather samples of larger networks.

In order to keep the network extraction process relevant, and on the chosen topic, a set of websites (*BadWebsites*) and a set of keywords (*Keywords*) were also defined (discussed in section 'keywords'). *BadWebsites* contained websites known to be safe and assumed to not host any pages relevant to child exploitation. Examples of these websites included www.microsoft.com and www.google.com. Without these made explicit, the crawler could wander into a

[14] Networks of 100 websites proved to lack any meaning as they ended up being primarily ego networks, while networks of more than 200 websites proved to be difficult to manage because of size as well the time it took to download/crawl the network.

search-engine leading it completely off topic and making the resulting network irrelevant to the specified topic. *Keywords* also gave CENE some boundaries which guided it during the exploration. For the crawler to include the page being analyzed, at least seven keyword from *Keywords* had to exist (line 10). If the keywords existed on the page, the page was assumed to be relevant to the network and the statistics on that webpage were calculated (line 11). In addition to statistics regarding the keywords, statistics pertaining to the number of images and videos were also collected (line 13). The links pointing out of the page were also retrieved (line 12) and added to the queue of pages to visit – if they had not been visited yet (lines 14-16). If however no keywords existed on the page, it was discarded and no further links were followed. In instances where links pointed to a webpage which did not exist (called a broken, or dead, link), or if the webpage could not be retrieved for any other reason (such as a timeout); the webpage was considered inaccessible and was discarded. In order to construct the features of the network, the links between websites were tracked (line 17), as well as the occurrence of each keyword (line 18). When the crawling process was complete, the statistics were aggregated up to the website level (line 18). For example the statistics for the node www.website.com were calculated from the statistics collected from all pages on that website. Thus, all pages on a website contributed to the features for that website. This allowed for the construction of a coherent network, complete with features assigned to both the websites and links (line 19).

Based on the keywords, and set of websites CENE could not explore, the network constructed remained on topic. Finally, the CENE retrieved up to twenty-five pages in parallel, thereby requiring between three to five days to extract.

3.1.2 Type of Website

This study included twenty networks. Ten were child exploitation networks, with five beginning with user-generated post websites, hence forth referred to as *Blogs*, and five beginning with traditional interlinking-page websites, hence forth referred to as *Sites.* The other ten networks were non-child exploitation networks, specifically sports related, which also consisted of five *Blog* networks and five *Site* networks. The twenty networks were compared within type (e.g., Blog A vs. Blog B), between website type (Blog vs. Site), and across website domains (child exploitation vs. non-child exploitation). Although child exploitive content can be found via other Internet Protocols, such as peer-to-peer, Internet chat relay, and newsgroups, they require different types of analysis and hence were not included in this study. The exclusion of these different Internet Protocols is discussed in further detail later.

Due to the continually growing importance of blogs within online social media (Furukawa et al., 2007; Mitchell, Wolak, Finkelhor, & Ybarra, 2008), we

categorise websites into two broad categories: blogs[15] and sites[16]. 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, when distributing child pornography, blogs provide two benefits over sites. First, blogs provide a more efficient and cheap way to distribute material. The very nature of blogs seems to make them more likely than other types of websites to link to other blogs or websites (Ali-Hasan & Adamic, 2007). Also, many blog hosts such as Blogger©, LiveJournal© or Sensualwritter© provide members with free space to post their blogs. This eliminates the financial capital someone would need to set-up their own website, as well as the knowledge needed to design the website.

Second, blog website hosts provide increased anonymity. 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 cautious of detection by law enforcement, as their name would be attached to the website registration information. Although each blog host has terms of services (TOS), it is usually

[15] Gruhl, Guha, Liben-Nowell, and Tomkins (2004) state that weblogs, or blogs for short, are a form of online journal whereby the user can publish their thoughts and feelings to the Internet. Blogs can be a part of other websites, such as Facebook and Twitter, or websites dedicated specifically to blogs, such as Blogger/Blogspot or LiveJournal.

[16] Sites 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 – and will be at a later date–, they are beyond the scope of the current study.

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 TOS that they do not monitor blogs (Google, 2010). Therefore, if a blog is found to be publishing illegal content and is shut down, it may be 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 added security that is inherent with online blogs may lead to more freedom to network a certain type of content.

3.1.3 Starting Websites

Child Exploitation Websites:

Ten websites pertaining to child pornography were chosen as starting points. 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)[17]. This process involved inputting the keywords and manually verifying websites that Google© suggested. Once a website was found to contain child pornography, it was selected as a starting website. As for the definition of child pornography, this included

[17] The selected website was the first that came up in the search engine, which met our requirement. An attempt was made to equally select boy-centered and girl-centered websites.

websites that did not necessarily contain hardcore content, but 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 starting websites were included in 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. In addition, bit-torrent websites were also used as potential starting points for site networks. The starting websites for blog networks were strictly of the blog genre, as described previously.

The two methods of website selection were chosen to mimic the process a person might take searching for child pornography[18]. The four ICE websites were selected to represent an individual being given a known website from a friend or another pedophilic individual. The six Google© found websites were selected to mirror the process an individual may go through if they went to search for the content themselves with no other information. Using the web-crawler, we can map out all the possible routes an individual might take through the network to obtain content. This is important, as we want to get a good representation of what a potential user might do. Of course, a user will be more inclined to follow the

[18] It is important to note that we are measuring all potential paths a user might take while searching for child pornography. It is highly unlikely that a user will take all possible paths or look at all the content; however, the potential is there. The key is understanding that each of these websites are connected and therefore, it is possible that a user might go to each or a subset.

content (child pornography) than the type of website (blog or site). Therefore, although each starting website corresponds to the network type (i.e., blog for blog networks and site for site networks), without visually inspecting each website, 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 is more about the starting website rather than all the websites within the network. Having said that, using the criteria of the word *blog* in the website title, or being a sub-domain of blogspot, nibblebit, tumblr, milkboys, sensualwritter, pichack, or wordpress, 22.5% (225 out of 1000) of the websites within the blog networks and only 4.1% (41 out of 1000) of site network websites met this criteria for classification as a blog. Additionally, thirty-two of the fourty-one blogs found within the site networks were located within Site Network A. Although 22.5% is not all of the websites, there is a clear difference between the site and blog networks in the percentage of blogs. Finally, it is important to note that given the strict criteria used, four of our five starting blog websites did not meet the criteria. As we know these are blogs, this suggests that the numbers we found (22.5%) is actually considerably lower than what percentage of the websites are actually blogs.

Control Websites:

To compare the structure and content of child pornography networks to other types of networks (or *control* websites), five blog and five non-blog sports networks were also analyzed. 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 x 150 pixels. Although the child pornography keywords were collected, they were not a criterion for website selection. The control networks were constructed to help validate the findings in our child exploitation networks. This was done two ways. First, the control networks were used to compare the structure of the, illegal, child exploitation networks to legal networks: sports. Second, the content of the child exploitation and non-child exploitation networks were compared to determine if there was a difference in what was contained on each type of network (i.e., keywords, videos, and images).

3.2 Network Content

3.2.1 Keywords

There were sixty-three keywords included in this study, that were broken up into three groups. The first group was words commonly used by the RCMP to

find illegal content containing children[19]. The second group was words that could be classified as 'softcore', 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.

As previously discussed, for the crawler to include a given webpage in the analysis, it had to contain at least seven unique keywords. If the criterion was met, the webpage was assumed to be relevant to the network topic (child exploitation) and statistics about that webpage were calculated and the links found on the webpage were followed. This strategy of classifying a webpage into the child exploitation category based on the number of keywords has a drawback. If the threshold is set too low, it is possible that a given non-child exploitation

[19] As these words are used by law enforcement to find child pornography, we were requested to suppress the list of words. However, for example purposes, we were allowed to share a few: pthc (pre-teen hardcore), realkiddy, and lolita. In addition, many of these keywords can be found within Le Grand et al., (2009) study of child pornography keywords within *eDonkey* peer-to-peer networks.

webpage will go above the threshold and be considered child exploitation material. This form of error, called a false positive, was minimized by selecting a high number of keywords: seven. Manual verification showed us that seven keywords did distinguish well between child exploitation webpages and regular webpages.

3.2.2 Videos and Images

In addition to collecting statistics on keywords, information regarding the number of videos and images on each website was also collected. These two attributes were selected because they are the most common material that is transferred in online child pornography networks (Wolak et al., 2005). Therefore, there was a logical thought process of analyzing each of these attributes on each website.

Typically, a website contains many small images that would not be classified as 'photos' but would be counted as such by the CENE. Therefore, to prevent the inclusion of irrelevant images (e.g., emoticons and logos), graphics smaller than 150 by 150 pixels were excluded from the count. Like the keywords, the videos and images were aggregated up to the website level. Unlike the images, there was no criterion set for the videos. The reason for this was because there was no way to properly filter out irrelevant content, while still leaving the relevant videos. Of course, these criterions have their own limitations. As we are

not manually verifying every image and video collected on each website, we cannot be completely sure that all the content would meet a judicial or our own operational definition of child pornography. This means that websites identified as being high in child pornography may have either no child pornography or have extensive adult pornography instead. The implications of this are discussed in further detail below and steps are outlined on how to better resolve the issue of false positives within the networks.

3.2.3 Website Connections

For each website crawled, each link found, leading to another website, was investigated by the web-crawler. In addition to scanning the new webpage for keywords, images, and videos, the number of times the new website was linked to, by the original website was also recorded. This meant, for example, that it was possible for website A to link to website B 9,000 times. Instead of simply recording whether a relationship existed between website A and website B, we decided to record how often that relationship was present. The reason for this was that we believed that the number of times two websites link to one another is of great importance. If website A contains 1,000 webpages, but only links to website B once, then the likelihood of an individual finding website B is decreased. However, if website A links to website B 1,000 times, on those 1,000 pages, than the likelihood that an individual would be able to travel to website B,

through website A, is greatly increased. That is, there is now, approximately, a link to website B on every webpage of website A, instead of one link somewhere in the 1,000 webpages.

3.2.4 Network Statistics

For each webpage and websites extracted, statistics about the number of keywords, videos, and images were collected. In addition, statistics regarding the number of valid websites, webpages, and connections to other websites were also collected. Through these, three key statistics were created. The first was whether a website was boy-centered, through the higher presence of the words boy, twink, penis, and cock, or girl-centered, through the higher prevalence of the words girl, Lolita, vagina, and pussy.

The second statistic was whether the website contained more references to hardcore content than softcore content. This was determined by whether the hardcore keywords, previously outlined, were more prevalent than the softcore keywords. If the hardcore keywords per page were higher than the softcore keywords per page, the website was classified as being hardcore-focused.

The final statistic involved using common SNA measures: density, clustering, fragmentation, reciprocity, and centrality. Combined these measures can be used to understand the structure of the child exploitation networks.

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 & Hanneman, 2006). This can be used to determine whether the network is closely packed with every website communicating with one another, or whether the network is more loosely packed with much less communication between the websites.

Clustering is measured through the *clustering coefficient* which examines the likelihood that if two websites are connected to a third website, that they will also be connected to one another (Malm & Bichler, in press). When compared to density, the coefficient lets us know whether the ties between websites is evenly distributed across the network or whether there are groups of websites that are more likely to cluster together and link to one another, but not to other websites.

Fragmentation refers to the percentage of the network connections that become disconnected as a result of the removal of a given website, from the overall network (Borgatti, 2003). Fragmentation informs us how a user's ability to travel through a network may be impacted by the removal of some websites. If the network is hard to fragment, it means that if a website is shutdown, it will do little to hamper a user's ability to find new websites to replace the one that was shutdown. Conversely, if the network is able to be fragmented, it impacts the networks efficiency to distribute content and suggests that law enforcement can disrupt the function of the network.

Reciprocity is about the proportion of websites that reference one another (Izquierdo & Hanneman, 2006). That is, if website A references website B, does website B also reference website A? Like density, reciprocity can help us to understand how websites within a network communicate with one another. More specifically, is there a global camaraderie between websites that results in them referencing one another on their websites or does the illegal nature of the activities result in the websites being more isolated and reducing the amount of reciprocal ties present.

Finally, *centrality* is measured through degree, closeness, and betweenness (Freeman, 1979). Degree is the most common form of centrality and can be measured through in and out-degree (Newman, 2003). For website A, in-degree centrality is the number of other network websites that link into website A. Meanwhile, out-degree is based on how many websites, website A links to. Recall that one of the main objectives of this study is to determine whether the most connected websites also have the most severe content. In-degree centrality tells us how popular a website is within the network. This is important when explaining content as, if the most severe websites are the most popular, then they should have the highest out-degree scores.

3.3 Determining the Key Players: Measuring Network Capital

NC is a term derived by Schwartz and Rouselle (2009) that 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 NC, 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 NC 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. Therefore *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	the resource sharing level

NC is comprised of 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 hardcore keywords, images,

and videos per web page[20]. 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, while all other websites are measured against that highest value and thus range between 0.0 and 1.0. The individual *node_severity* score is then the average of the three resources[21]. Therefore, the formula for calculating *node_severity* scores is as follows:

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

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 NC is *node_connectivity*. *Node_connectivity* refers to the contributions a node makes 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. This is multiplied by the percentage of its resources the node makes available to the rest of the network and then multiplied by any link

[20] 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 giving it a score.

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

weighted values. Although the original formula by Schwartz and Rouselle (2009) includes a resource sharing component, we made this a constant within the current study by assuming a resource sharing level of 100%, or 1.0. The reason for this decision was because all of the websites within our networks would be classified as open. That is, they did not require any special permission to access. As a result, any individual who accessed the website would have access to all resources available. Therefore, there was no reason to modify the resource sharing level. However, in online networks where certain permissions are required (e.g., membership), or where there are certain mechanisms in place to prevent some individuals from accessing specific content, a resource sharing level would have to be put in place. For this reason, we still included the resource sharing level component.

For 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[22]. The link weight was standardized between 0.0 and 1.0 using the same method as the resource weights. Although Schwartz and Rouselle's (2009) original formula included indirect connections, they were excluded in this study because the small network sizes (200 nodes) resulted in all nodes being either one or two steps away

[22] 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.

from one another. In other words, we held indirect connections constant. As a result of these modifications, 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.
NAW_i	number of resources (3)
RSL	the resource sharing level (1.0 for this study)
LW_{ij}	weighted number of times node i references node j, ranging from 0.0 to 1.0.

3.4 Analytic Strategy

Within the current study, four key research questions were investigated. The research questions are as follows:

1. Does the content of child exploitation networks differ from non-child exploitation networks? Does the content of blog child exploitation networks differ from site child exploitation networks?
2. Does the structure of blog child exploitation networks differ from site child exploitation networks?
3. Are the child exploitation websites that contain the most severe content also the websites that are the most connected?
4. Can key players within child exploitation network be determined using NC? What impact does the removal of the highest contributors to NC have to overall NC, within each network?

To answer the first research question, a descriptive analysis was performed comparing the characteristics of child exploitation networks to non-child exploitation networks, specifically sports related networks. Comparisons were also conducted on child exploitation blogs and sites. Each of the comparisons was analyzed at the network, website, and webpage levels. Additional t-tests were performed to determine whether the differences found across network types

-both child exploitation versus control and blog versus site- were significantly different.

The second research question was investigated using traditional social network analysis methods. This involved comparing the structure of child exploitation blogs and sites on reciprocity, density, clustering, fragmentation, and in/out-degree centrality. In addition to comparing across network types, comparisons were conducted within network types to determine whether the findings were consistent.

The third research question was addressed using the NC measure outlined in the methods. This involved comparing child exploitation blog and site networks on severity and connectivity as well as overall NC. In addition, comparisons, or differences, between severity and connectivity, across websites, were also analyzed. Finally, a comparison was made between in-degree centrality and NC, to determine whether the most outwardly connected and severe websites were also the most popular.

The final research question involved removing websites with the highest contributions to NC and determining the impact this had on the overall network's capital.

4: RESULTS

4.1 Network Content

4.1.1 Comparing Child Exploitation Networks to Control Networks

We start by comparing the child exploitation networks to the control networks (see Table 1). In general, the control networks were easier to construct than the child exploitation networks. Of the websites the CENE web-crawler attempted to visit, only 66% of the child exploitation blog network websites were active[23], while even fewer (56%) were active for the site networks. Conversely, the control blog and site networks were each 98% 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 are at a higher likelihood of being shutdown, while a blog about baseball does not run similar risks. One way of verifying this is through our measure of hardcore words per page. Also shown in Table 1, the hardcore keywords were found at a much higher rate per page within child exploitation networks (71 for blogs and 350 for sites) than sport-related networks (around four hardcore words per page). T-tests confirm

[23] For websites in the child exploitation networks, active refers to the website meeting our criteria of at least seven unique keywords. If the website did not contain at least seven keywords or was unable to be reached (dead, shutdown, timed-out), then it was considered inactive. For non-child exploitation networks, as we did not require the seven unique keywords, inactive simply referred to the website being unreachable by the web-crawler.

that these differences are statistically significant (see Table 2). This tells us that the words selected for the web-crawler were relatively good at distinguishing between websites of a sexual nature and non-sexual websites. 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.

Interestingly, control blog and site networks generally did not differ in the number of web pages, videos, or images per node or per web page. In addition, they 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.

4.1.2 Comparing Blog and Site Child Exploitation Networks

In Table 1, we compare the two samples of five blog and five site child exploitation networks. Websites dedicated to girls were more common in site networks than in blog networks ($t=2.22$, $p<0.05$). Although, in general, the levels of boy-centred and girl-centred websites were equal across network types. For blogs, the percentage of websites that were boy-centred was 62.8%, while it was slightly lower for sites at 55.2%. However, when we look at the individual

networks in Table 3 (blogs) and Table 4 (sites), we see a clear distinction within the networks. Blog Networks A, B, and C were primarily boy-centred, while Blog Networks D and E were primarily girl-centred. Within the site networks, a similar pattern arose. Site Networks A and B consisted of almost entirely boy-centred websites, while Site Networks C, D, and E were predominantly girl-centred. These findings suggest that although an individual is equally as likely to find boy-centred or girl-centred content in both network types, within a network, content is predominantly one sex. This supports Wolak et al. (2005) findings that offenders tend to fixate on one sex, with little overlap. Validating the web-crawler further, we found that the starting website for each network corresponded to the overall networks sex preference in all but one of the networks. That is, networks that were found to be boy-centred started with a boy website, while networks found to be girl-centred started with a girl website; with the exception of Site network B.

Overall, hardcore content was referenced more often in site networks ($t=5.48$, $p<0.00$). Within network types, the findings were more varied. For blog networks, the percentage of websites that were more hardcore focused ranged from 13.1% (Blog A) to 73.0% (Blog D). Meanwhile, for site networks, it ranged from 8.7% (Site C) to 100.0% (Site B). These findings suggest that although, in general, site networks have more hardcore content, the amount of hardcore websites is dependent on the individual network characteristics. That is, there are

groups of websites that are dedicated to hardcore content, and they stick together, while there are websites dedicated to more softcore content, who also stick together. Like with boy and girl-centred websites, the starting website seemed to dictate what type of network formed. In all but Blog network E, the starting websites hardcore/softcore rating corresponded with the overall networks rating. That is, if the starting website was labelled as hardcore so was the rest of the network; and vice versa. Again, this provides more validity for our crawler that it was able to remain on topic.

Table 1: Mean totals for blog and site child exploitation and control networks.

		Blogs (C.E.)	**Sites (C.E.)**	**Blogs (Controls)**	**Sites (Controls)**
Nodes (Valid)	Total	688	299	938	917
	Avg/Network	137.6	59.8	187.6	183.4
Number of Web Pages	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	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	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	Per Node	1,967.02	1,058.74	220.14	208.32
	Per Page	2.30	0.77	0.17	0.16
Images	Per Node	9,122.45	16,354.48	1,559.45	2,028.88
	Per Page	10.66	11.94	1.19	1.52

Table 2: Summary of T-tests comparing child exploitation networks to each other and control networks (Student's T reported).

	CE Blogs to CE Sites	Cont. Blogs to Cont. Sites	CE Blogs to Cont. Blogs	CE Sites to Cont. Sites
Number of 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<.05; **p<.01

Although blog networks were larger in average number of valid nodes (138 to 60), site networks averaged more web pages and images per node. Sites had more hardcore words per web page ($p<0.01$), while blogs had more videos per web page ($p<0.01$). 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 cataloguing those websites more, therefore making it easier for the websites to be found and shutdown.

Table 3: Descriptives for five child exploitation blog networks.

		Blog A	Blog B	Blog C	Blog D	Blog E
Nodes	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
Number of Web Pages	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
Pages on Starting Node		5,118	512	142	433	1,315
# of Nodes Starting Node Connects to (%)		109 (54.8)	90 (45.2)	89 (44.7)	6 (3.0)	179 (89.9)
Hardcore Words	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	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	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	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

Table 4: Descriptives for five child exploitation site networks.

		Site A	Site B	Site C	Site D	Site E
Nodes	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
Number of Web Pages	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		5,197	5,571	1,010	85	263
# of Nodes Starting Node Connects to (%)		18 (9.0)	199 (100.0)	10 (5.0)	23 (11.6)	104 (52.3)
Hardcore Words	Per Node	63,762	3,313,713	10,942	207,768	1,472,330
	Per Page	56.57	682.37	14.80	415.03	781.06
Other Words	Per Node	73,248	2,239,023	61,787	118,566	853,449
	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	4,746.1	111,176.5	2,344.6	1,016.2	42,207.8
	Per Page	4.21	22.89	3.17	2.03	1.85

4.2 Network Structure

Another objective of this study is to describe the structure of the networks derived from the web-crawl. This is important as it will help us understand how the networks function and what steps can be taken by law enforcement to best combat the networks functionality. For example, are the networks dense (high number of connections and reciprocity between websites), sparse (few connections between websites), and does it vary between blog and site networks? For law enforcement purposes, dense networks would suggest that the networks are highly connected and interact with one another regularly, while sparse networks would suggest more isolated networks and that law enforcement would have an easier time of fragmenting the existing network. Within this study, we derived the following measures: reciprocity, density, clustering coefficient, fragmentation, and in/out degree centrality. These measures are summarized in Table 5 (blogs) and Table 6 (sites). Confirming the findings of Ali-Hasan and Adamic (2007), blogs were found to be more reciprocal than sites (eighteen percent versus nine percent); although the reciprocity rates were lower in our study than those found by Ali-Hasan and Adamic. Within our control groups, the numbers were higher with blogs ranging from sixteen to 35 percent, averaging twenty-five percent, while sites ranged from nineteen to 33 percent, averaging twenty-six percent. The lower rates of reciprocation for child exploitation

networks are not surprising as the illegal nature of the website content should result in fewer reciprocal ties: a way to avoid detection. Nevertheless it is interesting to find that even within illegal domains, reciprocation follows the patterns of legal domains: blogs have higher rates than sites[24]. There were three exceptions to the patterns found: Blogs D and E, and Site A. Blogs A, B, and C, had reciprocity rates between twenty-one and twenty-three percent; however, Blog E, and to a lesser extent Blog D, had only half the reciprocity (eleven percent for Blog E and fourteen percent for Blog D).

The reciprocity findings seem to be correlated with the number of valid webpages within the network. Recall from Tables 3 and 4 that the total number of valid webpages was significantly lower for Blog Networks D and E, compared to the other three blog networks. Conversely, the number of valid webpages within Site Network A was considerably higher than the other four site networks. Therefore, the lack of reciprocity may actually be a lack of webpages searched. Then again, the average number of webpages per node did not differ across the networks; which may be a better indicator that the difference in reciprocity, found across network types, is actually present and not a case of total valid webpages.

[24] Although our child exploitation networks followed the patterns found by Ali-Hasan and Adamic, our non-child exploitation networks did not: blog and site networks were equally reciprocal. This might be the result of our networks not being solely blogs or sites and instead a mixture between the two. Of course, this has implications for our findings within child exploitation networks as the same argument can be as with the control networks. The findings are, however, partly supported by our validity check –previously discussed- measuring the percentage of websites within each network that were blogs based.

One possible explanation may be that the natural design of blogs makes them more likely to have reciprocity. That is, it is common-place for blogs to have what is called a 'blog roll', where the blog author can post links to other blogs they like. This type of structure is uncommon with non-blog websites and thus might be a contributing factor to the higher rates of reciprocity within the blog networks in this study.

The higher rates of reciprocity in blog networks might be a result of blog rolls, however, the lack of reciprocity within site networks may be for an entirely different, precautionary, reason. In the world of blogs there is little in repercussions for being found to have illicit material – besides getting shut down. However, for an independent website, the risk is a lot greater as individuals are tied to it through website registration and hosting services. This increased risk may limit the amount of reciprocal ties that are present. Furthermore, as search engines rank pages based on their popularity, having more links to a website increases its exposure on search engines, which in turn likely increases the possibility of being shut down.

Despite the difference in reciprocation across network types, the density scores were comparable; about seven percent (sites) to eight percent (blogs) of potential ties were present in the networks. Within network type, the findings were also fairly consistent with the maximum density for blogs being 0.12 (Blog B & C) and for sites, 0.09 (Site B). Recalling that density tells us the percentage

of all possible links between nodes that are present, the consistency in density found across networks suggests that the differences found for reciprocity are not the result of the number of valid webpages or nodes within each network. That is, site and blog networks have the same percentage of total possible links present, therefore, any differences found in reciprocity are the result of something other than the total number of pages or nodes.

Recall that the clustering coefficient is a measurement of a specific type of network density: density of triangular relationships within the network (Newman, 2003). There was a larger range in clustering for site networks (0.44 to 0.81) compared to blog networks (0.39 to 0.66). However, like overall network density (proportion of all available ties present) the average clustering coefficient was similar across network types. For site networks, the average coefficient was 0.60, while for blogs it was 0.51. This suggests that although there is a difference within network types, across types, site and blog networks are equally likely to have clusters of websites. However, the key finding is that within each network, the clustering coefficient was more than four times greater than the corresponding network density. A clustering coefficient that is equal to the network density states that the network is fairly random and that groups of nodes (in this case websites) do not cluster together. Considering that the clustering coefficient is four times greater, this indicates that there is indeed considerable clustering within each network. In other words, groups of websites all connect to one another and

are separate from other groups of websites within the same network. This is important because it suggests that although the networks are large, they tend to operate in smaller groups of websites. That is groups of websites all link to one another while those same websites do not link to large parts of the other websites within the network. This finding therefore has implications for law enforcement in regards to how they can best target key players. That is, within groups of websites, there are several key players that stand out from the rest of the websites.

Recall from earlier that we designed the networks to mimic the path a user might take if they were trying to find content. One of the ways that we can determine how much of an impact law enforcement can have on a users ability to navigate a network is to determine how easy, or hard, it is to fragment the network. As previously discussed, fragmentation scores range from 0.0 to 1.0 and represents the proportion of pairs of nodes that cannot reach one another. Proportions closer to 1.0 indicate that the network can be more easily fragmented.

In the networks with fewer valid nodes (Sites B to E, and Blog D), the fragmentation scores were very high, ranging from 0.87 to 0.97. Conversely, in the larger networks the range was 0.32 to 0.61. Thus, with the larger networks, nodes were more easily able to reach one another because they had more options (nodes) available. Put another way, in the larger networks there were multiple pathways that a user could take to get to more content. Therefore, if law enforcement were to shutdown a random website within Blog Network A, for

example, it would do little to impede a user's ability to access child pornography. However, there are some important, positive, findings for law enforcement. Although randomly shutting down a website seems to do little to disrupt the efficiency of network mobility, targeting specific websites can make a difference within some networks. For instance, in Blog D, the removal of the 'top' node resulted in a 34 percent fragmentation, while in Site A, it was sixteen percent, and in Site C, it was 57 percent. However, in Blog C, the reduction was only one percent, while for Blog A and Site B it was eight percent. In most networks, the law of diminishing returns was evident. That is, beyond the top node, or the top two in the rare circumstance, the amount of fragmentation that occurred, from the removal of a given node, was minimal. Then again, this might not be a negative. What it suggests is that in some circumstances, simply removing one or two nodes from a network can have a large impact on the ability to navigate the network. Therefore, these findings reinforce the need for targeted attacks by law enforcement and that removing specific websites can have a large impact on a user's access to material.

Table 5: Social network measures for Blog Networks.

	Blog A	Blog B	Blog C	Blog D	Blog E
Density (Ties)	0.07 (2917)	0.12 (4876)	0.12 (4846)	0.02 (988)	0.07 (2938)
Clustering Coefficient	0.50	0.51	0.51	0.39	0.66

Fragmentation		0.51	0.44	0.38	0.87	0.61
Reciprocity		0.22	0.21	0.23	0.14	0.11
Central-ization	In-Degree	73.95	26.07	22.11	12.66	20.86
	Out-Degree	17.38	80.61	74.12	28.31	84.50

Table 6: Social network measures for Site Networks.

		Site A	Site B	Site C	Site D	Site E
Density (Ties)		0.08 (3310)	0.09 (3415)	0.01 (434)	0.03 (1080)	0.07 (2780)
Clustering Coefficient		0.44	0.81	0.47	0.51	0.77
Fragmentation		0.32	0.89	0.89	0.97	0.87
Reciprocity		0.20	0.04	0.12	0.03	0.06
Central-ization	In-Degree	25.98	2.99	3.45	2.32	5.60
	Out-Degree	82.55	91.88	26.18	82.52	72.10

4.3 Network Capital

The main objective of this study is to find the key players in online child exploitation networks. Individually, connectivity and content are important elements for finding key players; however, on their own neither sufficiently identifies whom police should prioritize. Instead, both need to be taken into account simultaneously. This is achieved through NC: the combined use of connectivity (connections to other websites within the network) and content severity (number of hardcore words, videos, and images per webpage). The connectivity, severity, and NC scores for each blog and site network is presented in Table 7 (blogs) and Table 8 (sites). First, we examine connectivity scores between and within network types, followed by a similar analysis of severity and then NC. Second, we examine whether the same nodes (websites) are identified across all three measures: connectivity, severity, and NC. Finally, we examine whether the nodes with the most severe content, as identified by severity scores, are the most popular nodes within their own networks, as identified by in-degree centrality measures.

4.3.1 Connectivity

Network linkages in online child pornography networks is an area that has lacked sufficient research (Krone, 2004). However, the connections that websites make between one another can tell us a lot about how the network is structured

and how information (e.g., images) is being distributed across the Internet. Within this study, the mean connectivity scores across networks were low (0.001 to 0.008), with the exception of Site Networks A (0.077), B (0.067), and E (0.053). When comparing these three site networks to the other two site networks, the clear difference resided in the size of the network. That is, networks A, B, and E had more valid webpages than networks C and D. However, this does not explain the discrepancy for two reasons. First, all the networks are standardized and therefore their mean scores should be approximately equal. Second, it would make more sense that the smaller networks (C and D) would have higher means because of several extreme (large amounts of connections to several websites) nodes skewing the average. However, when comparing the three higher-mean networks to the two lower-mean networks, the proportion of high connectivity score nodes is greater in the three high-mean networks. Therefore, it would appear that these three networks differ from the other two networks in that they may contain more key players. Despite the differences found amongst site networks, there was no difference across blog networks, and Site Networks C and D means corresponded with the blog networks.

Although Site Networks A, B, and E differed from the other networks, in their mean connectivity scores, all ten networks had equal median scores (<0.001). This finding supports the hypothesis that within each network, there are

several highly connected nodes that act as key players, or ‘mega-websites’, within the overall network. These nodes tend to be connected to many other websites within the network and they have multiple connections to each of those websites. Again, this provides evidence that, when properly selected, law enforcement can have an impact on the overall network. However, without knowing the content, it is unclear which of these websites would be the best to target.

4.3.2 Severity

Traditionally, the image content of a website would denote its severity; however, as we previously argued, it is important to take into account multiple forms of content (images, video, and text) as they all play important roles. When examining severity scores, the means across networks were consistent, with the exception of Site Network B and E. Again, it is not fully clear why this was the case; however, these two networks, by a substantial margin, had the highest median severity scores (0.16 and 0.18 respectively). These findings seem to coincide with the findings within connectivity that within these networks, there are a high number of key players, or ‘mega-websites’, that possess a lot of content. It is worth noting that when we looked at the range of max severity scores within each of the ten networks, they were much higher for site networks (0.63 to 0.82), than for blog networks (0.37 to 0.49). This finding suggests that within site networks the disparity between low severity nodes and high severity

nodes is greater. This conclusion is supported further by the median scores within each network. For blog networks, the distributions were more even than in site networks (median and mean scores are close in blog networks). Of course, the number of valid nodes might have also played a role, as Site Network A had the most valid nodes of all site networks, but also had the lowest max score (0.63). However, as previously discussed, Site Network A behaves more like a blog network than a site network. Therefore, it is unclear whether this finding is the result of the distribution within the network or whether it is a characteristic of the network size, or a combination of the two. Like connectivity, the severity of the content on a website can tell us a lot of vital information; however, alone it only tells us what is present at a fixed point in time and does not tell us anything about how, and who, that content may be circulated.

Table 7: Network capital descriptives for the five Blog CE 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 5 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 5 Scores (% Change)		0.37 (2.6)	0.39 (4.9)	0.42 (2.3)	0.67 (2.9)	0.94 (6.0)

Table 8: Network capital descriptives for the five Site CE 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 5 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 5 Nodes (% Change)		0.60 (9.1)	25.64 (24.4)	2.52 (0.0)	2.01 (4.7)	19.78 (0.0)

4.3.3 Network Capital

Although the connections between websites and the content contained on a website are important, alone, each only tells us part of the story. It is not just about the number of connections a website has, as a website with many connections, but no content, does not pose a threat. Conversely, a website with a lot of content, but few connections, is unable to quickly distribute child pornography to other websites/users. Therefore, it is important that we take into consideration both connections and content at the same time. This is because the

website high in content and connectivity has the ability to distribute the large amount of content it possesses, to a multitude of other websites/users. Therefore, law enforcement efforts need to equally balance both attributes and make decisions about who are the key players with both being considered.

Presented in Table 7 (blogs) and Table 8 (sites) are the NC scores for each of the ten child pornography networks. Each NC score was multiplied by 1,000 to ease the interpretation of the values. Scores ranged from 0.66 (Site A) to 33.9 (Site B); however, site networks had a higher mean NC score. Blog networks ranged from 0.38 (Blog A) to 1.00 (Blog E). Recalling that the NC scores are based on standardized measures, the difference found between network types informs us that there is a real-world difference in the nature of blog and site networks. There are two possible reasons for this difference. First, site networks tended to have fewer valid nodes, meaning that nodes very high in both severity and connectivity had more leverage on NC. This is partly supported by the connectivity and severity mean scores, which were higher for several site networks. Therefore, it may not simply be a case of severe and highly connected nodes having leverage, but instead that site networks, when compared to blog networks, contain more severe and highly connected nodes in general. Second, recall from Table 1 that site networks had a greater proportion of hardcore content websites. This supports the previous hypothesis of multiple extreme nodes within site networks. This is countered, in part, by the higher levels of videos in blog

networks; however, it would appear that the difference in hardcore websites has more influence than the difference in videos per webpage. Regardless of the reason for the difference between blog and site networks, overall, these findings suggest two things. First, that within blog networks, there are a few extreme nodes that tend to stand out from the rest of the nodes. Second, the higher proportions of hardcore websites found in site networks results in there being multiple key players present. However, are nodes found to be high in NC (denoting key player) also the nodes with the highest severity and connectivity scores?

The importance of using NC as opposed to connectivity or severity on their own, is illustrated in Figures 2 and 3, and described in Table 9. Figure 2 shows all nodes in Blog Network A, sized by their content severity score. The five yellow highlighted nodes represent the top five connectivity scores, while the red nodes represent the top five severity scores. Although illustrated in Figure 2, by looking at Table 9, we see that nodes high in content severity are not necessarily high in connectivity. For example, the node with the highest severity score (ID: 179) is not in the top ten for connectivity; and vice versa (node 119). In fact, none of the nodes are in the top ten in both severity and connectivity. Figure 3 provides a similar illustration as Figure 2, but with Site Network A. Again, nodes highest in connectivity are in yellow, while nodes highest in severity are in red. Nodes within the top five in both severity and connectivity (ID: 164

and 168) are highlighted in green. In Table 9, we see that the same node (164) is highest in severity and connectivity. Within Site Network A, severity seemed to play a bigger role in NC and there was some overlap between severity and connectivity. This finding might be the result of site networks operating differently than blog networks, or simply a function of smaller networks and thus fewer valid nodes from which to select. Regardless, these findings suggest that law enforcement agencies can take multiple approaches for combating child exploitation websites, based on their goal. If the goal is to remove content, then it might be best to focus primarily on the severity score, while if the goal is to disrupt/fragment the network, it might be best to focus primarily on the connectivity score. In the end however, we suggest that law enforcement still combine connectivity and content severity as targeting one over the other will result in missing important targets – more so with blogs than with sites.

Table 9: Top ten nodes for network capital, severity, and connectivity, for Blog A and Site A.

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

4.3.4 Comparing Severity to Website Popularity (In-degree)

We have shown that websites that have a lot of content are not necessarily the websites that reach out to other websites (high connectivity). However, this fails to answer the question whether websites high in content are more popular within their network. This question can best be answered by correlating the severity scores with in-degree centrality. Recall that in-degree centrality informs

us how often a website is linked to by another website. If severity decides popularity, in-degree and severity scores should be highly correlated.

Within blog networks, the correlation between in-degree and severity ranged from -0.13 to 0.15, with Blog E being the only positive correlation. Within site networks the range was from -0.27 and 0.60; however, only Site E was above 0.02 and only Site D was below -0.12. Only Site E (0.60) and Site B (0.02) were positively correlated. Across the ten child pornography networks, it is clear that severity and in-degree (popularity) are not correlated. In fact, it might be that popularity and severity are inversely correlated. Although the reason for this is unclear, a hypothesis can be derived.

Websites that become popular increase their risk of being shutdown as they become easier to find. As noted by previous research (Morselli, 2009), when given a decision between efficiency and security, a network will select security. Therefore, within child pornography networks, the most efficient network would involve having the websites with the most severe content also being the most highly connected (both in and out degree centrality). This, in turn, would optimize dissemination. However, to ensure the websites survival, the best strategy for the network would be to minimize the amount of incoming traffic to the 'mega-website'. This may coincide with Carley's (2006) point that covert networks adapt because of law enforcement practices. By minimizing public advertisement and relying on word of mouth or private advertising (e.g., private

messages on a discussion board or chat room), the network reduces outside individuals (e.g., law enforcement) from entering the network. This practice illustrates Beech et al. (2008) point that online child exploitation networks have built-in security measures to keep unwanted individuals out. Because of these known practices, it may not be a surprise that not only are the most severe websites *not* the most popular, but, in fact, these two measurements appear to be inversely related.

4.3.5 Removing the Key Players

NC is a measure that combines the content present on a website with the number of connections the website makes to other websites. This measure can be a valuable tool when trying to discover which websites should be marked as key players and the focus on law enforcement prioritization. However, the next step is to see how NC changes when we remove the top contributors to NC. To examine the importance of key players within online child pornography networks, we first calculated total NC for each of the ten networks (see Table 5 and 6). Second, we sequentially removed the top five nodes who contributed the most to NC. Finally, we re-calculated the new NC score, after the removal of each node, and compared the new value to the original NC value. Table 5 and 6 show that once the top five contributors were removed, NC was reduced an average of 14.4% for blogs and 36.2% for sites. The reduction was consistent for blogs,

ranging from a 9% to 21%, while sites varied from 21% to 63%. These reductions seem to be the result of network size, as networks with the lowest number of valid nodes (Blog D and E, and Site B, C, D, and E) had the most significant reductions in NC; with the exception of Blog E. There was no specific correlation between network reduction and standard deviation (SD), as the websites with the largest SD (Blog E, Site B, and E) did not have the largest NC reduction. This suggests that NC is 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 from removing random nodes? Using a random number generator, we randomly selected five nodes from each network to be removed. The results are presented at the bottom of Tables 7 and 8. With the exception of Blog Network E, selecting the five top scores resulted in at least a four times 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% versus 24.4%), while in Site Network C, the change was large (43.1% versus 0.0%). 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 differentiates nodes within a network. Second, that within each network, there are key players. Third, that police investigations need to be targeted and not random as we can have a much greater impact on the overall network with targeted attacks.

5: DISCUSSION

The Internet has provided the social, individual, and technological circumstances needed for the production, distribution, and consumption of child pornography to flourish (Taylor & Quayle, 2003). Within the current study, we looked to better understand 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, 2) comparing child exploitation networks to non-child exploitation networks as well as comparing child exploitation network types (blogs and sites), and 3) by adapting the measure of 'network capital' first created by Schwartz and Rousselle (2009) to the context of child exploitation networks. Instead of being focused strictly on connectivity or exposure within the network, our measure of network capital also took the severity of content into account. In doing so, we began to address Krone's (2004) request to incorporate the linkages between websites as well as Taylor et al. (2001) work on the severity of website content.

Using the CENE web-crawler, we were able to create networks pertaining to child exploitation. Although there were issues with false positives –which will be discussed below –the web-crawler appeared to do a suitable job of identifying websites containing child exploitive content (i.e., child pornography and child erotica). Although it is unclear what proportion of our websites were false positives, two steps were taken to increase the validity of our findings. First, we required the presence of seven keywords for a website to be included in the network. Second, we compared our child exploitation networks to our sports-related networks and found that there was a clear difference in hardcore content (keywords). Through these two steps we are more confident that the content we were targeting were of a sexual nature and focused on children.

As noted by Young (2005), and Mitchell et al. (2007) websites dedicated to child exploitation are not difficult to find. Our research proved to be no exception. Despite most of the content being illegal, a simple Google© search using both sexual and non-sexual terms resulted in countless websites linking to child erotica and child pornography. Previous research by Wolak et al. (2005) found that only twenty percent of online child pornography arrestees used sophisticated tools for hiding illegal content, while Carr (2004a) found that only twenty-five percent did. The reason for this may be the perceived anonymity of the Internet and the low likelihood of legal ramifications. Therefore, it is not surprising that the content was so readily available, even through search engines.

However, the ease of finding content leads us to suggest that companies like Google© can continue to play a key role in decreasing the amount of content available. As discussed previously, they, along with Microsoft©, are working on tools that can help find child pornography. By integrating these tools into their own search engines and programs, they can further reduce the amount of child pornography being distributed on the World Wide Web.

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

When comparing the two different types of child exploitation websites – blogs and sites- the number of images per page were equal, however, blogs had more videos per page. It is unclear whether this is the result of the structure of the network (i.e., our starting point) or something entirely different. For instance, within blogs, the content is embedded directly onto the page; however, with sites, the content is sometimes indirectly linked. That is, the site may provide a web-link to another website (possibly a blog) that hosts the content, or a web-link to a peer-to-peer program. The small number of networks extracted for this study

does not allow us to make a definitive statement if blogs act more as hosts/producers while sites as distributors. Given that some of the blogs we analyzed were run by the youth themselves, this hypothesis may hold some validity.

In addition to differences in the number of videos per page, site networks had higher levels of hardcore content, through both the number of hardcore words per page and the percentage of websites that had more hardcore words then softcore words. As search engines work on keywords, having more hardcore keywords provides benefits and disadvantages. Having more keywords results in more people being able to find the website; however, those people might be users or law enforcement officials. Therefore, it seems to be a kind of catch-22 situation. For people to access the website they need to know that it exists; however, for people to know that it exists requires some form of 'advertising', which increases the risk of being shutdown. As for blogs, focusing on softcore material may decrease the likelihood of being shutdown. If there is ambiguity as to whether the content is illegal or not, then the blog might be allowed to remain active longer. Couple that with uncertainty regarding what constitutes child pornography (Jensen, 2004; Sinclair & Sugar, 2005), focusing on softcore content seems like a form of security for blogs (but it may reduce web traffic).

Following the patterns found by Ali-Hasan and Adamic (2007), our blog networks had higher rates of reciprocity in comparison to site networks.

Although Ali-Hasan and Adamic examined legal online networks, it is interesting that even within the illegal context the pattern of reciprocity remained. Ali-Hasan and Adamic suggest that blogs act as relationship builders and that bloggers typically frequent the same blogs regularly. Therefore, it comes as no surprise that blogs have higher levels of reciprocity as there appears to be more of a community aspect to them. This might also explain why there were fewer dead links found within blog networks. Due to the group mentality, the blogs remained open for longer.

Although there was a difference in link reciprocation, blog and site networks were found to be equally dense (same proportion of all possible ties present); however, site networks clustered more than blog networks. The higher degrees of clustering may partly explain why site networks were harder to construct. Through clustering, groups of websites can more easily be targeted by law enforcement for removal. In addition, given the aforementioned findings of Ali-Hasan and Adamic (2007) it is somewhat surprising that clustering was lower in blogs. If blogs act as communities and people tend to venture the same few blogs, it would lead one to believe that clustering would be higher. Then again, the average clustering coefficient in the site networks was driven by Site B and E, who had twenty-four and thirty-one valid nodes respectively. Once these two networks are removed, blog and site network clustering was approximately the same. Therefore, clustering may be more of a function of network size then the

actual network construction. As future research is completed, and the methods are fine-tuned, the answer to this question may become more apparent.

We also began to address the need outlined by Krone (2004) for online investigations of child pornography to incorporate the linkages between users/websites. This is an important step forward because, to our knowledge, this has not been looked at previously. When it comes to combating child pornography, linkages are important as they can help us follow the route through which content is transferred. With this knowledge, we can retrace the steps an image has taken, as it has been distributed across the Internet, and possibly find the originating source. If it is content currently being produced, then we can halt the abuser and rescue the child who is being exploited. Within the current study, we found that the size of the network seemed to play a role in the connectivity scores for each network. For small networks, the mean connectivity scores were considerably higher than in the large networks. Despite the difference in mean scores, the median scores were consistent across all ten child exploitation networks. Therefore, the difference in mean scores might be the result of the network size, as mentioned, or the number of key players within the network. Regardless, by including connectivity in our measurement we created a better-rounded measurement of online child exploitation.

Along with connectivity, we examined content severity. As previously noted, online investigations primarily take the form of image searches. However,

as Wolak et al. (2005) found, videos were also a common form of transfer. Young (2005) also noted that using common keywords such as ‘boy’ and ‘child’, in search engines, would result in child exploitive content. Understanding that each of these attributes was different forms of content, we looked to improve on existing methods of detecting child pornography by incorporating all three. Our measurement of content severity was more robust and should improve interpretations of what constitutes a key player. There were some issues with false positives and thus additional steps need to be taken to improve in this regard. Nevertheless, the use of all three forms of severity is an improvement over existing methods and is a step in the right direction.

As outlined above, one of the key objectives of this study was 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 does. We found that websites with harmful content had varying degree of exposure in the derived networks, and that the measure of network capital was able to properly discriminate between targets that met (or not) both criteria. This demonstrates the utility of a network approach in target prioritization when targets are in abundant supply, as it is the case for child exploitation websites.

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, however, is that the most connected websites are the most likely to be discovered. Another possible reason is that 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.

When we removed the top five contributors to network capital for Blog networks, network capital was reduced by an average of 14.4%. For Site networks, the reduction was even higher at 36.2%. Comparing this to the removal of five random websites from each network, we found a substantial difference. For Blog networks, the average reduction was only 3.7% while for Site networks the reduction was only 7.6%. These two findings 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 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 network capital can aid in maximizing the impact that we can have on fragmenting online child exploitation networks.

In regards 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, 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 SNA method of network capital as an improvement to existing strategies. That is, node targeting is an effective tool in combating online child exploitation.

Despite SNA's overall benefits over less systematic approaches, law enforcement has been slow to adopt it and has a tendency to falter when making decisions about the structure of criminal networks (Clark, 2007). For instance, it is often assumed that the most important people are those running the organization. Baker and Faulkner (1993) point out that it is common for criminal enterprises to conduct business in a decentralized model. This decentralization process is a way to ensure security and the long-term viability of the criminal organization (Morselli, 2009). In other words, the decentralization process allows members of the organization to be replaced and for the network to still function efficiently. This process of decentralization makes law enforcement efforts more difficult as it can make finding the key players more challenging.

The Internet is decentralized by nature, which adds to the difficulty of combating online crime. This has aided in the long-term viability of online child pornography distribution. Because of the inherent difficulty in finding key players online, it becomes more important to introduce new methods that can aid at reducing the difficulty. This is especially true online, where there is little to no face-to-face contact between users. Therefore, SNA 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'* (operating the website) may not actually be the key players. As websites allow the contribution of multiple

people, the individual operating the website may only act as a broker, while other individuals may actually be the key players that 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, as the actual key player has not been apprehended. This leads to the need to combine SNA methods, such as network capital, with other methods to create a multi-dimensional approach to online crime.

6: CONCLUSION

In the current study, we examined the metaphorical street level market for online child pornography. Like the street drug market, our networks were open and visible to any and all that wanted to view them. Specifically, within this study, we examined the structure and linkage of two types of online child exploitation networks. We did this by mimicking the path an individual might take in their search for such content: from a successful starting point to other potential target websites found through an exploration of the links provided by each website. However, this study should not be seen as a definitive conclusion, but instead, a stepping-stone in the development of innovative methods designed to map and analyze online networks. Through the continued exploration and investigation of these networks, we can better understand the evolution of child exploitation online. This will assist researchers in coming up with ways to combat the problem as well as aid law enforcement with meeting their goals of disruption and intelligence. Although the problem of online child exploitation can never be eradicated, by working together and continuing to investigate this domain, steps can be taken to reduce the prevalence, determine key players, and find children that are currently being abused.

6.1 Research Implications for Policy Development

Through the creation of a web-crawler and the use of network capital, we address three key aspects of police practices: improving on existing techniques for law enforcement, reducing the costs of conducting business, and improving officers' ability to conduct their duties. Each of these three aspects is reflected on in more detail below.

The web-crawler is a tool that can play an important role in finding child pornography online. Although the web-crawler is in its infancy stages and may require alterations as we make progress in this line of work, its use can improve on existing policing practices. More specifically, the web-crawler improves on existing techniques by reducing the need for human intervention. Using the web-crawler, law enforcement officials do not have to manually search for child pornography. Instead, the web-crawler does it automatically. This has an impact on both reducing costs of policing practices as well as allowing officers to focus their attention on more pressing issues. First, the web-crawler's efficiency means that it can analyze webpages faster than any human can, and thus allows more data to be investigated in a shorter period. This means that funding can be re-allocated to pay for more officers to investigate incidents of child pornography. Second, the web-crawler allows law enforcement agencies to focus their time on investigating cases of child pornography instead of spending their time searching for the content. This means that officers can take the information gathered using

the web-crawler and investigate the cases. While they are investigating the cases, the web-crawler is continuing to search the Internet for more child pornography. Once a case has been investigated, the officer does not have to begin a new search. Instead, while they were attending to the previous case, a new case was being found. Finally, Burns, Morley, Bradshaw, and Domene (2008) point out that officers in the ICE unit are at a greater risk of traumatic stress, due to the content they are viewing. Through the automation of finding child pornography online, the amount of stress an officer endures can be reduced, thereby decreasing health related costs associated with treatment.

The use of network capital also has important implications for police work. It improves on existing techniques whereby decisions on who to investigate was based solely on the presence of images, or simply who was found to be in possession of child pornography. Using network capital, multiple factors are taken into account –images, videos, text, and connections- and the websites high in the combination of these are identified. This maximizes law enforcement efficiency by providing evidence of who should be targeted first. In turn, this has implications for reducing costs of policing, through improved efficiency. As both resources and funds are limited, using network capital to identify key players provides high economic gain. Put another way, the pursuit of every criminal case has an economic cost associated with it (e.g., costs of officer, legal fees, and office supplies). If key offenders were targeted, the economic value of the case

would be maximized through the removal of those with the most content and connections.

6.2 Limitations & Future Research

Although substantial initial steps were made in exploring the structure of online child exploitation networks and finding the key players, there are several limitations in this study that need to be taken into consideration. These limitations fall under three broad categories: Issues with web-crawler, access to data, and social network analysis measurements. In subsequent research, we look to improve in these areas in the following ways.

6.2.1 Web-crawler

There are three key areas where improvements to the web-crawler are needed. First, the efficiency of the web-crawler needs to be improved. Second, decisions need to be made regarding how non-child pornography websites will be dealt with. Third, there needs to be a reduction in the number of false positives.

The first step for future research is to make refinements to the web-crawler. In order to develop a tool that can be used by law enforcement, and/or private organizations, a few modifications to the efficiency of the web-crawler need to be made. The current version of the web-crawler takes, on average, three to four days to analyze a network of 250,000-300,000 webpages. In addition, due to the amount of system resources required by the program, most computers can

only run two or three sessions of the web-crawler at a time --while still allowing the computer to be functional for the end-user[25]. Despite this being a considerable improvement over a manual analysis of 300,000 web pages, the web-crawler's functionality can be improved. The primary improvement is to allow for the possibility to more efficiently analyzing larger, yet 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 websites.

The second issue pertains to how non-child pornography websites will be dealt with by the web-crawler. Within the current study, if a website did not meet our requirement of seven keywords, it was ignored. This seems like the most logical decision; however, it still leaves some uncertainties. Although a website might not contain child pornography directly, they could be supporting child pornography. Therefore, a website that does not contain child pornography, but links to a lot of content may be a key player within the network.

Finally, there was the issue of false positives within the current web-crawler. The primary reason behind this was that it was designed to find and catalogue any target closely related to online child exploitation. As such, it was equally likely to retrieve websites hosting hardcore content, as well as websites

[25] If the computer is being used solely to examine networks, four or five networks can be run simultaneously.

sitting on the fence of child erotica[26] and adult pornography. Therefore, steps need to be taken to improve on the ability to ascertain true child pornography from adult pornography. One of the best ways to do this is through the use of pre-existing image databases, like those mentioned earlier: United States Division of Criminal Justice Services Database, ICSE-DB, and the NCVIP.

Although the continued refinement to 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 totalled the number of images on a website, which met a specific size criterion (greater than 150 by 150 pixels); however, the web-crawler did not distinguish between the content of the images. As previously mentioned, law enforcement practices typically focus their investigations around the images that are present on various websites. The primary reason for this is that it is easier to track images than other content, because of something called a 'hash function'. A hash function is the mathematical process of taking a large piece of data and transferring it into a single value, known as a 'hash value' (Howard, 2004). Hoffman (2010) states that these hash values act as a form of encryption that can be used to authenticate the content of an image. According to Hardy and Kreston (2004), the chances of two files having the same hash value, but different content, is 10^{38}. Therefore,

[26] 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.

utilizing known hash values, instead of just the total number of images, would help improve the validity of our severity measurement[27].

Another key area for future development is to incorporate hash values for videos. Like with images, we only incorporated the total number of videos present on a website. Although this was helpful, we were unable to easily analyze the content of the videos and hence the severity. Therefore, our severity measure increased with the *quantity* of videos, as opposed to an ideal measure of both content and quantity. Recent technological advances have resulted in hash function algorithms, known as acoustic fingerprints, being used to identify audio files (see Cano, Batlle, Kalker, & Haitsma, 2005). Currently there are no known child pornography video databases; however, steps are being taken to formulate such a database. If this occurs, the web-crawler can then connect to the database, like with the images, and catalogue the known child pornography videos as well as possible new videos. This would result in a modification of the network capital formula, to incorporate known child pornography videos and 'other' videos.

6.2.2 Data Access

In subsequent research, there are three data collection issues that need to be addressed: expanding to additional Internet domains, accessing different types

[27] 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'.

of websites, and ensuring that the type of website coincides with the type of network (i.e., sites in site networks and blogs in blog networks).

Although our priority is to design a tool that can accurately map and analyze World Wide Web networks, our long-term goal is to expand to different domains of the Internet. This is because, according to Carr (2004a), 42% of Internet sex offenders use the World Wide Web to obtain images, while 78% use Internet Chat Relay (IRC), and 39% use newsgroups. In addition, a large amount of traffic on peer-to-peer (P2P) networks is dedicated to the distribution of child pornography (Steel, 2009). Therefore, there is a need for tools that can also analyze P2P networks (Pau de la Cruz, Aller, Garcia, & Gallardo, 2010). However, for a P2P tool to be effective it needs to incorporate a function similar to image hash values but for videos, as videos are the predominantly searched media on these types of networks (Steel, 2009).

The second issue regarding data access centres on the inability to properly access two types of websites: galleries and membership (i.e., paysites and login). First, given that our web-crawler required the presence of at least seven of our keywords, photo galleries were commonly ignored by the crawler. Of course, images are a key part of online child exploitation and therefore, the likely exclusion of pure photo galleries from this study is noteworthy. Access to a hash value database could solve this issue for future studies.

The second type of website excluded in this study was websites that required login identification. As Beech et al. (2008) points out, many child pornography websites have authentications that are required for access. The purpose of this is to keep unwanted individuals off the website. This creates two research problems. First, the web-crawler is unable to create a login to access these websites. Therefore, when it comes to one of these types of websites, the web-crawler skips over it and declares that it does not contain child pornography. In some instances, this is correct; however, given that authentication is a security tactic used by child pornography websites, some of the websites that are ignored will actually be false negatives[28]. Second, because of the login identification, the website is no longer defined as being public. Therefore, accessing these websites, and recording the data found, is a breach of privacy and would require a warrant. Due to these two reasons, it is highly unlikely that any subsequent research will be able to incorporate membership-based websites.

The final issue with data involves the proper designation of websites to a network type. Despite having two 'types' of networks (blogs and sites) it is unclear what percentage of the websites within each network actually coincides with the type of network. Although each starting website was of the appropriate network type, it is doubtful that all of the websites crawled were of the same type.

[28] Websites identified as not containing child pornography, when, in fact, they do contain child pornography.

A visual inspection of the networks showed that blogs were more likely to link to blogs and sites were more likely to link to sites, as previously noted; however, there were instances where a site linked to a blog and hence that blog ended up in the site network. Therefore, when discussing the differences between site and blog networks, what we are talking about is differences when one is used as a starting website, over the other. In subsequent studies, efforts could be made to minimize this effect by including exclusionary criteria based on the type of network. For instance, when scanning site networks, we could tell the crawler to ignore any website that had *blogspot (*a common blog host) in the url. This process would be similar to the one previously discussed, that was used to exclude websites such as Google© and Microsoft©. However, given all of the independent blog websites this might not be possible. Furthermore, when constructing blog networks, there is no specific way of excluding site-based websites, unless there was an inclusion criterion. Again though, this would end up excluding all of the independent blog websites and result in many missed data. The impact of this issue within the current study is unclear as there were differences found between network types. However, as previously discussed, Site Network A had a blog rate of at least sixteen percent and seemed to follow the analysis patterns of the blog networks despite starting with a site website. Therefore, there seems to be evidence that in most cases, but not always, website types coincided with the appropriate networks.

6.2.3 Social Network Analysis

There are some inherent limitations with using social network analysis. Within the current study, these were present in three ways: the inability to obtain complete networks, the lack of weighting within the network capital formula, and the assumption of a static network. Each of these is discussed in further detail.

Sparrow (1991) notes that criminal network data tends to be incomplete, lacking all the connections or relationship actually present amongst nodes. When relating this to online networks, and our study, this was evident. Given the enormous size of the Internet, it is virtually impossible to obtain an entire network. Although we limited the networks to 250,000 pages and 200 websites, this is probably not the full extent of each network. Despite many of our networks not consisting of 200 valid nodes, it is important to remember there were issues with invalid nodes and nodes without child exploitive content that were included in the 200 nodes. Therefore, it is possible that, if these issues are better taken into account, we can obtain larger networks of strictly child exploitive websites. This seems to be particularly important for site networks where several networks consisted of only thirty to forty websites. However, there is a trade-off associated with larger networks. Although we might be able to examine a greater percentage of the overall network, the additional resources needed may not outweigh the additional data that can be collected. More specifically, analyzing a network of 500 websites may not provide us with any

additional knowledge that we cannot obtain with a network of 200 websites. Therefore, a careful balancing act needs to occur to maximize the information we collect while minimizing the resources needed to carry out the analysis. We believe that the first step of this balancing act is ensuring that the 200 websites that are included in the network are actually child exploitive websites. Once this has occurred, expanding to larger network sizes is a possible area for future research.

As previously discussed, the use of image, and possibly video, hash values may result in a modification to the way that *node_severity* is calculated in the future. Currently, we evenly weight images, (hardcore) keywords, and videos; however, the uncertainties attached to the keywords and videos –whether they are child pornography or not—could lead to a modified weighting system or the exclusion of keywords from the final formula[1]. By using image hash values, we could effectively eliminate false positives from our analysis, resulting in a more accurate measurement of content severity.

Finally, Carley (2006) notes that SNA typically assumes a static relationships between the actors within a network; however, in most networks this is not the case (see also Easton & Karaivanov, 2009). For example, when examining covert terrorist networks, Carley points out that the network will typically adapt to attacks by laying low, breaking away from those being attacked,

engaging in different actions, and joining with other groups. Of course, the Internet is a highly dynamic domain and thus is prone to efforts by offenders to avoid detection and modify their behaviour because of law enforcement actions. Therefore, future research needs to examine how these online networks evolve. As the Internet is a very dynamic environment, that is constantly changing, a suitable research design could involve repeated measures taken on the networks at different points in time. Through this, we would be able to determine how frequently websites are shutdown, what types of websites are usually targeted, and what facilitates their removal. By following several networks, over an extended period, and collecting data at fixed time intervals (e.g., every thirty days), we could better understand the nuances of the dynamic structure. Comparing to non-child exploitation networks, we could clarify what is unique about child exploitation networks and how they are similar to non-child exploitation networks. For example, Leskovec, Kleinberg, and Faloutsos (2005) found that as online social networks evolve, they become denser and the overall distances between websites become smaller. Due to the illegal nature of child exploitation websites, it is unclear whether the networks would follow this pattern or whether densification would facilitate their detection and thus removal.

The issue of combating child exploitation seems to be a topic that many people support. As technology continues to advance, individuals who exploit children will find new places and ways to abuse children. Although the current

research is in its infancy, we have proposed two important tools (web-crawler and network capital) that can help shed like on the growing problem. Through the continued growth of the project, we can have a large impact on this issue and continue to aid individuals who investigate child exploitation cases on a daily basis.

REFERENCE LIST

Akdeniz, Y. (1999). *Sex on the net: The dilemma of policing cyberspace*. Reading, U.K.: Garnet Publishing

Ali-Hasan, N., & Adamic, L. (2007). Expressing social relationships on the blog through links and comments. ICWSM 2007, Boulder CO.

Baker, W.E., & Faulkner, R.R. (1993). The social organization of conspiracy: Illegal networks in the heavy electrical equipment industry. *American Sociological Review, 58,* 837-860.

Balkundi, P., & Kilduff, M. (2006). The ties that lead: A social network approach to leadership. *The Leadership Quarterly, 17,* 419-439.

Baluja, S. (2008). Building software tools to find child victims. Retrieved from http://googleblog.blogspot.com/2008/04/building-software-tools-to-find-child.html

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.

Bill C-15A: An Act to amend the Criminal Code and to amend other Acts. 1st Reading, June 4, 2002, 37th Parliament, 1st Session, 2001-2002. (Online). Ottawa. Available at: http://laws.justice.gc.ca/PDF/Annual/2/2002_13.pdf

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 (pp.241-252).* Washington D.C.: National Academy of Science Press.

Borgatti, S. (2006). Identifying sets of key players in a social network. *Computational and Mathematic Organization Theory, 12*, 21-34.

Borgatti, S., & Lopez-Kidwell, G (in press). Network theorizing. In P. Carrington, and J. Scott (Eds.), *The sage handbook of social network analysis.* Thousand Oaks, CA: Sage Publishing

Bruinsma, G., & Bernasco, W. (2004). Criminal groups and transnational illegal markets: A more detailed examination on the basis of social network theory. *Crime, Law, and Social Change: An Interdisciplinary Journal, 41,* 79-94

Burke, A., Sowerbutts, S., Blundell, S., & Sherry, M. (2001). Child pornography and the internet: Policing and treatment issues. *Psychiatry, Psychology and Law, 9,* 79–84.

Burt, R. (1992). *Structural holes: The social structure of competition.* Cambridge, MA: Harvard University Press.

Burns, C.M., Morley, J., Bradshaw, R., & Domene, J. (2008). The emotional impact on and coping strategies employed by police teams investigating in Internet child exploitation. *Traumatology, 14,* 20-31.

Butts, C. T. (2009). Revisiting the foundations of network analysis. *Science, 325*. doi: 10.1126/science.1171022.

Carley, K.M. (2006). Destabilization of covert networks. *Computational Math and Organization Theory, 12,* 51-66.

Carley, K.M., & Hill, V. (2001). Structural Change and Learning Within Organizations. In A. Lomi, and E.R. Larsen (Eds.), *Dynamics of organizations: Computational modeling and organizational theories (pp.63-92).* Cambridge, MA: MIT Press.

Carr, A. (2004). Internet traders of child pornography and other censorship offenders in New Zealand. Wellington, NZ: Department of Internal Affairs. Retrieved from: http://www.dia.govt.nz/diawebsite.nsf/wpg_URL/Resource-material-Our-Research-and-Reports-Internet-Traders-of-Child-Pornography-and-other-Censorship-Offenders-in-New-Zealand

Carr, J. (2004). Child abuse, child pornography and the internet. London: NCH.

Cooper, A. (1998). Sexuality and the Internet: Surfing into the new millennium. *Cyber Psychology & Behaviour, 1,* 181-187.

Cooper, A., Scherer, C.R., Boies, S.C., & Gordon, B.L. (1999). Sexuality on the Internet: From sexual exploration to pathology expression. *Professional Psychology, Research and Practice, 30,* 154-161.

Criminal Code, R.S. c. C-46 (1985). Retrieved from http://laws.justice.gc.ca/PDF/Readability/C-46.pdf

Cross, A. (2010, August, 27). 2 arrests in Canada in crackdown on international child porn network. *Postmedia News*. Retrieved from: http://www.vancouversun.com/news/arrests+Canada+crackdown+international+child+porn+network/3450024/story.html

Dretzin, R. (Writer), & Dretzin, R., & Maggio, J. (Directors). (2008). Growing up online [Television series episode]. In D. Fanning (Executive Producer), *Frontline*.

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., & Karaivanov, A.K. (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*. Retrieved from http://abcnews.go.com/Technology/wireStory?id=8591118

Ferguson, N.M., & Garnett, G.P. (2000). More realistic models of sexually transmitted disease transmission dynamics: Sexual partnership networks, pair models, and moment closure. *Sexual Transmission of Disease, 27,* 600-609.

Freeman, L.C. (1979). Centrality in social networks: Conceptual clarifications. *Social Networks 1,* 215-239.

Freeman-Longo, R. E., & Blanchard, G. T. (1998). *Sexual abuse in America: Epidemic of the 21st century.* Brandon, VT: Safer Society Press.

Frith, M. (2003 October 8). 20,000 child porn images a week put on Internet, says NSPCC. Retrieved from: http://www.independent.co.uk/news/business/news/20000-child-porn-images-a-week-put-on-internet-says-nspcc-582609.html

Fulda, J.S. (2005). Internet stings directed at pedophiles: A study in philosophy and law. *Widener Law Journal, 15*, 47-84

Furukawa, T., Ishizuka, M., Matsuo, Y., Ohmukai, I., & Uchiyama, K. (2007). *Analyzing reading behavior by blog mining*. 22nd Annual Conference on Artifical Intelligence (AAAI-07), 1353-1358.

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

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

Greenfield, P.M. (2004). Inadvertent exposure to pornography on the Internet: Implications of peer-to-peer file-sharing networks for child development and families. *Applied Developmental Psychology, 25*, 741-750.

Griffiths, M.D. (2000). Excessive Internet use: Implications for sexual behavior. *Cyber Psychology and Behavior, 3*, 537-552.

Gross, E.F. (2004). Adolescent Internet use: What we expect, what teens report. *Journal of Applied Development Psychology, 25*, 633-649.

Gruhl, D., Guha, R., Liben-Nowell, D., & Tomkins, A. (2004). Information diffusion through blogspace. In S.I. Feldman, M. Uretsky, M. Najork, and C.E. Wills (Eds.), *Thirteenth International World Wide Web Conference (pp.491-501).* New York, NY: ACM Press.

Hafner-Burton, E., Kahler, M., & Montgomery, A.H. (2009). Network analysis for international relations. *International Organization, 63*, 559-592.

Hardy, R.L., & Kreston, S.S. (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. Retrieved from: http://www.sapsac.co.za/geeks.pdf

Hanneman, R.A., & Riddle, M. (2005). *Introduction to social network methods.* Retrieved from University of California, Riverdale (http://faculty.ucr.edu/~hanneman/)

Hoffman, S. (2010). An illustration of hashing and its effect on illegal file content in the digital age. *Intellectual Property and Technology Law Jou*rnal, 22, Retrieved from: https://lawlib.wlu.edu/works/426-1.pdf

Hooked. (2001 January 6). *Free spirits: Boylove on the Internet.* Retrieved from: http://www.freespirts.org

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.

Internet World Stats (2010 June 30). World Internet usage and population statistics. Retrieved from: http://www.internetworldstats.com/stats.htm

INTERPOL. (2010 February 2). *Crimes against children.* Retrieved from: http://www.interpol.int/public/children/default.asp

Izquierdo, L.R., & Hanneman, R.A. (2006). *Introduction to the formal analysis of social networks using mathematica.* Retrieved from: http://www.luiz.izquierdo.name

Jensen, R. (2004). Pornography and sexual violence. National Electronic Network on Violence Against Women. Retrieved from: http://www.oneangrygirl.net/jensenlong.pdf

Johnson, B. (2008 June 6). Time taken to shut child abuse sites criticised. *The Guardian*. Retrieved from: http://www.guardian.co.uk/technology/2008/jun/06/internet.childprotection

Keeling, M.J. (2005). Networks and epidemic models. *Journal of the Royal Society, 2,* 295-307.

Kempe, D., Kleinberg, J., & Tardos, E. (2003). Maximizing the spread of influence through a social network. Presented at the Association for Computing Machinery SIGKDD. Washington, D.C.

Krebs, V.E. (2002). Mapping networks of terrorist cells. *Connections, 24,* 43-52.

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., & Taylor-Butts, A. (2009). Child luring through the Internet, 2009.Juristat, 29 (1) (Cat No. 85-002-X). Ottawa, ON: Statistics Canada

Malm, A.E., Kinney, J.B., & Pollard, N.R. (2008). Social network and distance correlates of criminal associates involved in illicit drug production. *Security Journal, 21,* 77-94.

Malm, A., & Bichler, G. (in press). Networks of collaborating criminals: Assessing the structural vulnerability of drug markets. *Journal of Research in Crime and Delinquency.*

McCullagh, D. (2008 March 20). FBI posts fake hyperlinks to snare child porn suspects. Retrieved from: http://news.cnet.com/8301-13578_3-9899151-38.html

McGloin, J. (2005). Policy and intervention considerations of a network analysis of street gangs. *Criminology and Public Policy, 4,* 607-636.

McLaughlin, J. (2004). Cyber child sex offender typology. Retrieved from: http://www.ci.keen.nh.us/police/typology.html

Media Awareness Network (2010). Criminal Code of Canada: Child pornography and luring of children on the Internet-Summary. Retrieved from: http://www.media-awareness.ca/english/resources/legislation/canadian_law/federal/criminal_code/criminal_code_child.cfm

Medina, R., & Hepner, G. (2008). Geospatial Analysis of Dynamic Terrorist Networks. In I.Karawan, W.McCormack, & S.E.Reynolds *Values and Violence: Intangible Aspects of Terrorism (pp.151-167)* Berlin, Germany: Springer.

Microsoft. (2005 April 7). Tool thwarts online child predators. Retrieved from: http://www.microsoft.com/presspass/features/2005/apr05/04-07CETS.mspx

Microsoft. (n.d.). Microsoft technology helps in fight against child pornography. Retrieved from: http://www.microsoft.com/industry/publicsector/government/cetsnews.mspx

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

Ministry of Labour and Citizens' Services. (2006). *Detecting pornographic images on the network.* Victoria: Information Security Branch: Office of the Chief Information Officer

Mitchell, K.J., Finkelhor, D., & Wolak, J.W. (2003). The exposure of youth to unwanted sexual material on the Internet. *Youth & Society, 34*, 330-358.

Mitchell, K.J., Finkelhor, D., & Wolak, J.W. (2007). Youth Internet users at risk for the most serious online sexual solicitations. *American Journal of Preventative Medicine, 32,* 532-537.

Mitchell, K.J., Finkelhor, D., & Wolak, J.W. (2008). Are blogs putting youth at risk for online sexual solicitation or harassment? *Child Abuse & Neglect, 32*, 277-294.

Morselli, C. (2001). Structuring Mr. Nice: Entrepreneurial opportunities and brokerage positioning in the cannabis trade. *Crime, Law, and Social Change, 35*, 203-244.

Morselli, C. (2009). *Inside criminal networks.* New York: Springer

Natarajan, M. (2006). Understanding the structure of a large heroin distribution network: Quantitative analysis of qualitative data. *Journal of Quantitative Criminology, 22,* 171-192.

Newman, M.E.J. (2003). The structure and function of complex networks. *SIAM Review, 45,* 167-256.

O'Brien, G. (2011 January 14). Child porn suspect arrested. *The Castlegar Source.* Retrieved from: http://castlegarsource.com/news/crime/child-porn-suspect-arrested-9220

Papachristos, A. (2009). Murder by structure: Dominance relations and the social structure of gang homicide. *American Journal of Sociology, 115,* 74-128.

Reuters. (2001 February 13). Child porn gang face jail. *CNN.com*. Retrieved from: http://archives.cnn.com/2001/WORLD/europe/UK/02/13/england.pornography/

Ropelato, J. (n.d.). *Internet pornography statistics.* Retrieved from: http://www.internet-filter-review.toptenreviews.com/internet-pornography-statistics.html)

R. v. Alicandro, ON.C.A. 133 (2009). Retrieved from: http://scc.lexum.umontreal.ca/en/bulletin/2010/10-07-23.bul/10-07-23.bul.html

R. v. Legare, S.C.C. 56 (2009). Retrieved from: http://csc.lexum.umontreal.ca/en/2009/2009scc56/2009scc56.html

R. v. Sharpe, 1 S.C.R. 45 (2001). Retrieved from: http://scc.lexum.umontreal.ca/en/2001/2001scc2/2001scc2.html

Saltman, J. (2011 January, 12). Students apologize to victim in Facebook-rape case. *Postmedia News*. Retrieved from: http://www.montrealgazette.com/news/Students+apologize+victim+Facebook+rape+case/4100096/story.htm.

Schwartz, D.M., & Rouselle, T. (2009). Using social network analysis to target criminal networks. *Trends in Organized Crime, 12,* 188-207.

Shiels, M. (2008 April 14). Google tackles child pornography. *BBC News.* Retrieved from: http://news.bbc.co.uk/2/hi/7347476.stm

Sinclair, R.L., & Sugar, D. (2005). *Internet based child sexual exploitation environmental scan.* Ottawa, ON: Royal Canadian Mounted Police. Retrieved from:

Smith, M.A., & Kollock, P. (1999). *Communities in cyberspace*. London: Routledge.

Sparrow, M. (1991). The application of network analysis to criminal intelligence: An assessment of prospects. *Social Networks, 23*, 251-274.

Spink, A., Ozmutlu, H.C., & Lorence, D.P. (2004). Web searching for sexual information: An exploratory study. *Information Processing and Management: An International Journal, 40*, 113-123.

The Office of the Attorney General (2010, June 17). *Attorney General Cuomo announces groundbreaking initiative to enable social networking sites to eliminate thousands of images of child pornography*. New York: Author. Retrieved from: www.ag.ny.gov.

Taylor, M., Holland, G., & Quayle, E. (2001). Typology of paedophile picture collections. *The Police Journal, 74*, 97-107.

Taylor, M., & Quayle, E (2003). Child pornography: An Internet crime. East Sussex: Brunner-Routledge.

Technology Quick Response Team. (2005, January). Youth Internet usage statistics. From: http://ces.ca.uky.edu/extension_regions/Technology_Resources/Yth_Internet_StatS_UsU.pdf

Technorati (2010). Sports blogs. Retrieved from: http://technorati.com/blogs/directory/sports

Techterms Dictionary (2010). Bittorrent. Retrieved from: http://www.techterms.com/definition/bittorrent

The Hague. (2011 March 16). More than 200 children identified and rescued in worldwide police operation. *Europol.* Retrieved from: http://www.europol.europa.eu/index.asp?page=news&news=pr110316.htm

Top Sites Blog. (2010 June 19). Top 11 most popular sports websites. Retrieved from: http://topsitesblog.com/best-sports-websites/

TopTenREVIEWS (2004 February 6). TopTenREVIEWS Releases Porn Industry Statistics. Retrieved from: http://www.toptenreviews.com/2-6-04.html

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

Van Hulst, R. (2009). Introduction to social network analysis as an investigative tool. *Trends in Organized Crime, 12*, 101-121.

Wasserman, S., Faust, K., Iacobucci, D., & Granovetter, M. (1994). *Social Network Analysis: Methods and Applications.* Cambridge, MA.*:* Cambridge University Press.

Wellman, B., Salaff, J., Dimitrova, D., Garton, L., Gulia, M., & Haythornthwaite, C. (1996). Computer networks as social networks: Collaborative work, telework, and virtual community. *Annual Review of Sociology, 22*, 213-238.

Wolak, J., Finkelhor, D., & Mitchell, K. J. (2005). Child pornography possessors arrested in Internet-related crimes: Findings from the National Juvenile Online Victimization Study (NCMEC 06-05-023). Alexandria, VA: National Center for Missing & Exploited Children.

Wolak J., Finkelhor D., Mitchell K., Ybarra M. (2008). Online predators and their victims: Myths, realities and implications for prevention and treatment. *American Psychologist, 63*, 111-128.

Xu, J., & Hsinchun, C. (2008). The typology of dark networks. *Communications of the ACM, 51,* 58-65.

Young, K.S. (2005). Profiling online sex offenders, cyber-predators, and pedophiles. *Journal of Behavioral Profiling, 5*, 1-15.

Young, K.S., Griffin-Shelley, E., Cooper, A., O'Mara, J., & Buchanan, J. (2000). Online infidelity: A new dimension in couple relationships with implications for evaluation and treatment. In A. Cooper (Ed.), *Cybersex: The dark side of the force (pp. 59-74).* Philadelphia PA.: Brunner-Routledge.

APPENDIX A: FIGURES

Figure 1. 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** $L \notin$ *Queue*() **and** $L \notin$ *FollowedPages*
15:	*Queue*() ← *Queue*() + L
16:	D_L ← domain of *L*
17:	*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*
18:	*KeywordsInWebsite*(D_P) ← *KeywordsInWebsite*(D_P) + *KeywordsInWebsiteCounter()*
19:	**return** *WebsitesUsed(), KeywordsInWebsite(),LinkFrequency(), VideosInWebsite(), ImagesInWebsite()*

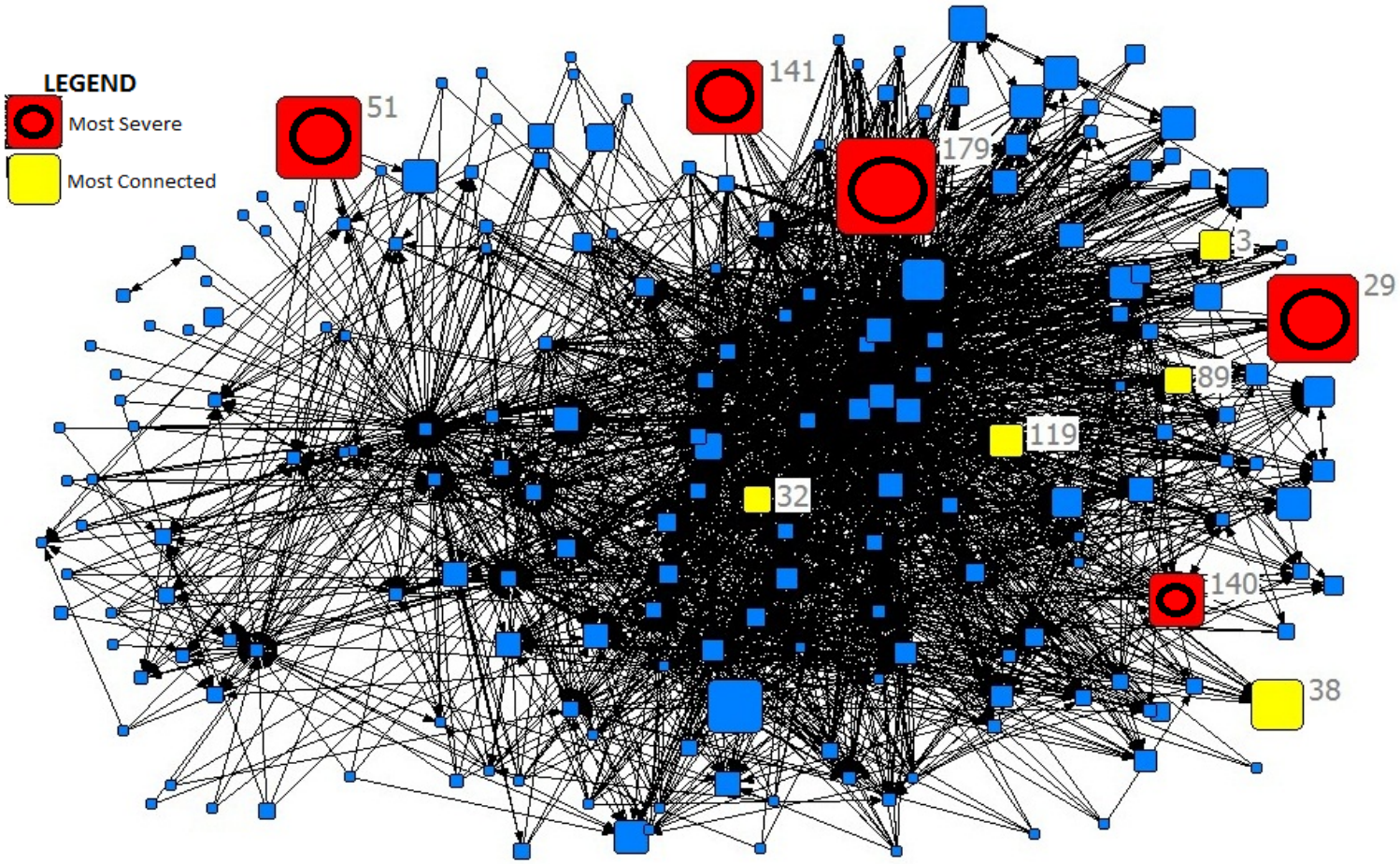

Figure 2. Severity scores with top five severity scores (red) and connectivity scores (yellow) highlighted for Blog Network A.

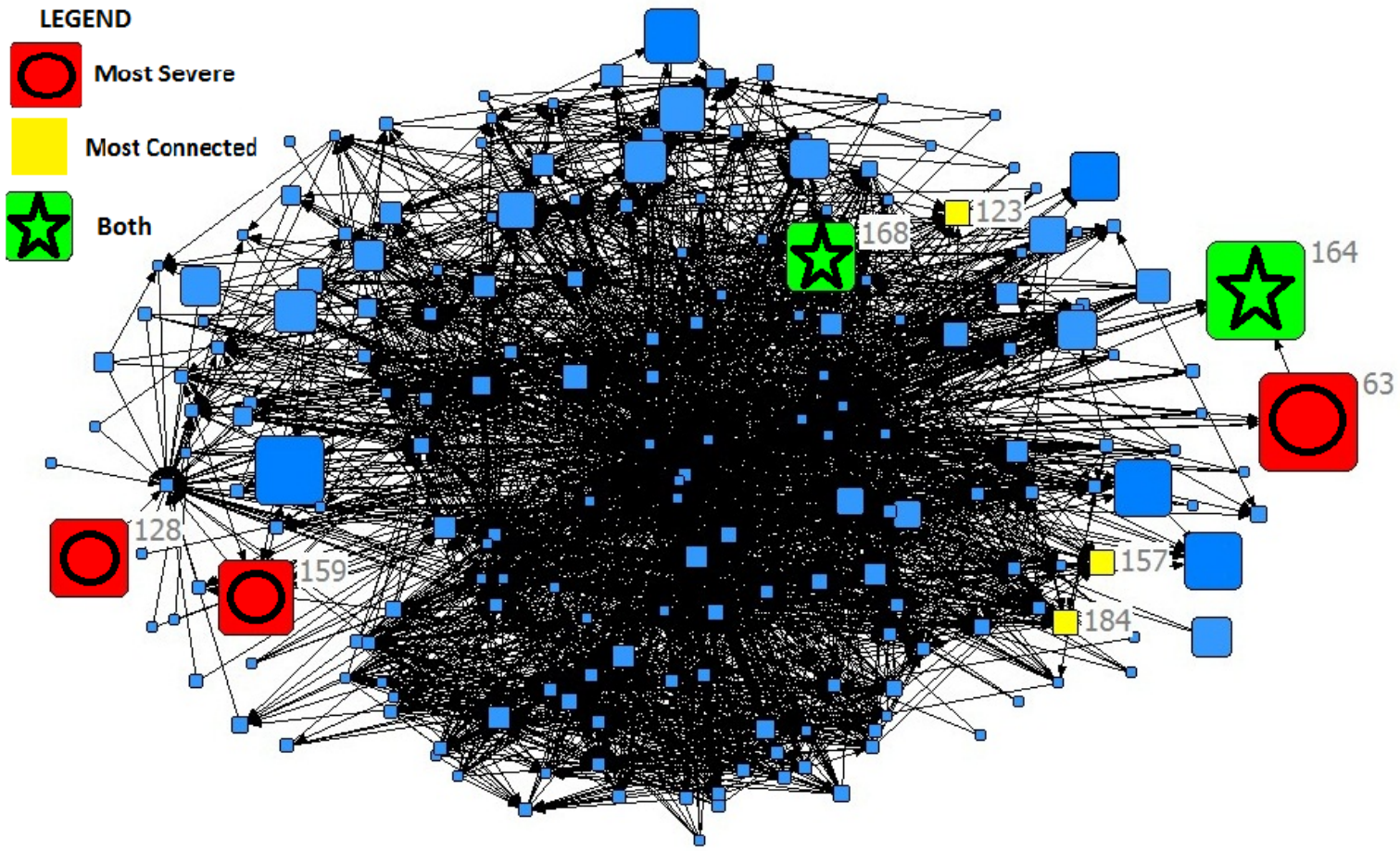

Figure 3. Severity scores with top five severity (red), connectivity (yellow), those in both (green) highlighted for Site Network A.