January 2010

The Structure and Content of Online Child Exploitation Networks

Richard Frank
Simon Fraser University

Bryce Westlake
Simon Fraser University, bryce.westlake@sjsu.edu

Martin Bouchard
Simon Fraser University

The Structure and Content of Online Child Exploitation Networks

Richard Frank
School of Computing Science
School of Criminology
Simon Fraser University
Burnaby, BC, Canada
rfrank@sfu.ca

Bryce Westlake
School of Criminology
Simon Fraser University
Burnaby, BC, Canada
bwestlak@sfu.ca

Martin Bouchard
School of Criminology
Simon Fraser University
Burnaby, BC, Canada
mbouchard@sfu.ca

ABSTRACT
The emergence of the Internet has provided people with the ability to find and communicate with others of common interests. Unfortunately, those involved in the practices of child exploitation have also received the same benefits. Although law enforcement continues its efforts to shut down websites dedicated to child exploitation, the problem remains uncurbed. Despite this, law enforcement has yet to examine these websites as a network and determine their structure, stability and susceptibleness to attack. We extract the structure and features of four online child exploitation networks using a custom-written webpage crawler. Social network analysis is then applied with the purpose of finding key players – websites whose removal would result in the greatest fragmentation of the network and largest loss of hardcore material. Our results indicate that websites do not link based on the hardcore content of the target website; however, blogs do contain more hardcore content per page than non-blog websites.

Categories and Subject Descriptors
H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces – *Web-based interaction*

General Terms
Algorithms, Measurement, Security, Human Factors

Keywords
Child exploitation, social network analysis, target prioritization, Internet

1. INTRODUCTION
It is estimated that 1.8 billion individuals worldwide use the Internet, with 260 million users being from North America [13]. Of the 1.8 billion users, adolescents and college students make up the largest proportion [10, 24, 31]. Through access at home and at school, it is estimated that 90% of youth have regular access to the Internet [5]. Although the vast majority of individuals who use the Internet for sexual pursuits do so in a safe and legal way [4, 9], the anonymity of the Internet has resulted in a growing percentage who sexually solicit youth [23]. What makes this problem worse is the ease with which one can obtain illegal pornographic material [30, 35]. Searching the words 'boy', 'teen', or 'child', brings up countless websites and photos of youth in sexually exploitive roles [24, 34].

The growth of the Internet has resulted in a substantial increase in research aimed at understanding online networks [8, 17, 29, 33]. However, most of the research to date has focused on the structure of social networking websites such as Facebook and MySpace, and has stopped short of investigating child exploitation networks. This is despite the United Nations announcement that there are more than four million websites containing child pornography [6].

Much of the existing efforts to curb 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 roadblocks from those who argue Internet stings are a form of entrapment[1] [7]. 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. One of the ways child exploitation has been combated is with the creation of a database containing all known sexually explicit photos of children (the International Child Sexual Exploitation image database) [14]. Additionally, INTERPOL partners with the COSPOL Internet Related Child Abuse Material Project and the Virtual Global Taskforce to help coordinate multi-country investigations and spread awareness of the problem. These efforts have had some results. 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 [28]. This resulted in the conviction of seven individuals and the confiscation of 750,000 images and 1,800 videos, containing 1,263 identifiable children[3].

[1] One such example is the FBI posting fake links to explicit images of children and then raided the homes of those who clicked on the links [20].

[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" [12].

[3] The children in the images and videos ranged from 3 months to 16 years. The majority were under the age of 10 with many being 2 or 3 years old.

Algorithm CENE(*StartPage, PageLimit, WebsiteLimit, Keywords(), BadWebsites()*)

```
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 ∉ WebsitesUsed() and |WebsitesUsed| < WebsiteLimit then
6:          WebsitesUsed() ← WebsitesUsed() + D_P
7:      if D_P ∈ WebsitesUsed() and D_P ∉ BadWebsites() then          //evaluate this page
8:          PageContents ← Retrieve page P
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()
14:                 if L ∉ Queue() and L ∉ FollowedPages
15:                     Queue() ← Queue() + L
16:                     D_L ← domain of L
17:                     LinkFrequency(D_P, D_L) ← LinkFrequency(D_P, D_L) + 1
18:             KeywordsInWebsite(D_P) ← KeywordsInWebsite(D_P) + KeywordsInWebsiteCounter()
19: return WebsitesUsed(), KeywordsInWebsite(), LinkFrequency()
```

Figure 1 - Algorithm CENE

Although a lot of time and money has been placed into various units across the world, child exploitation is nowhere near under control. The best available statistics suggest that less than 1% of all virtual pedophiles are apprehended [22]. This is not necessarily an attack against law enforcement, but rather speaks to the extent of 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.

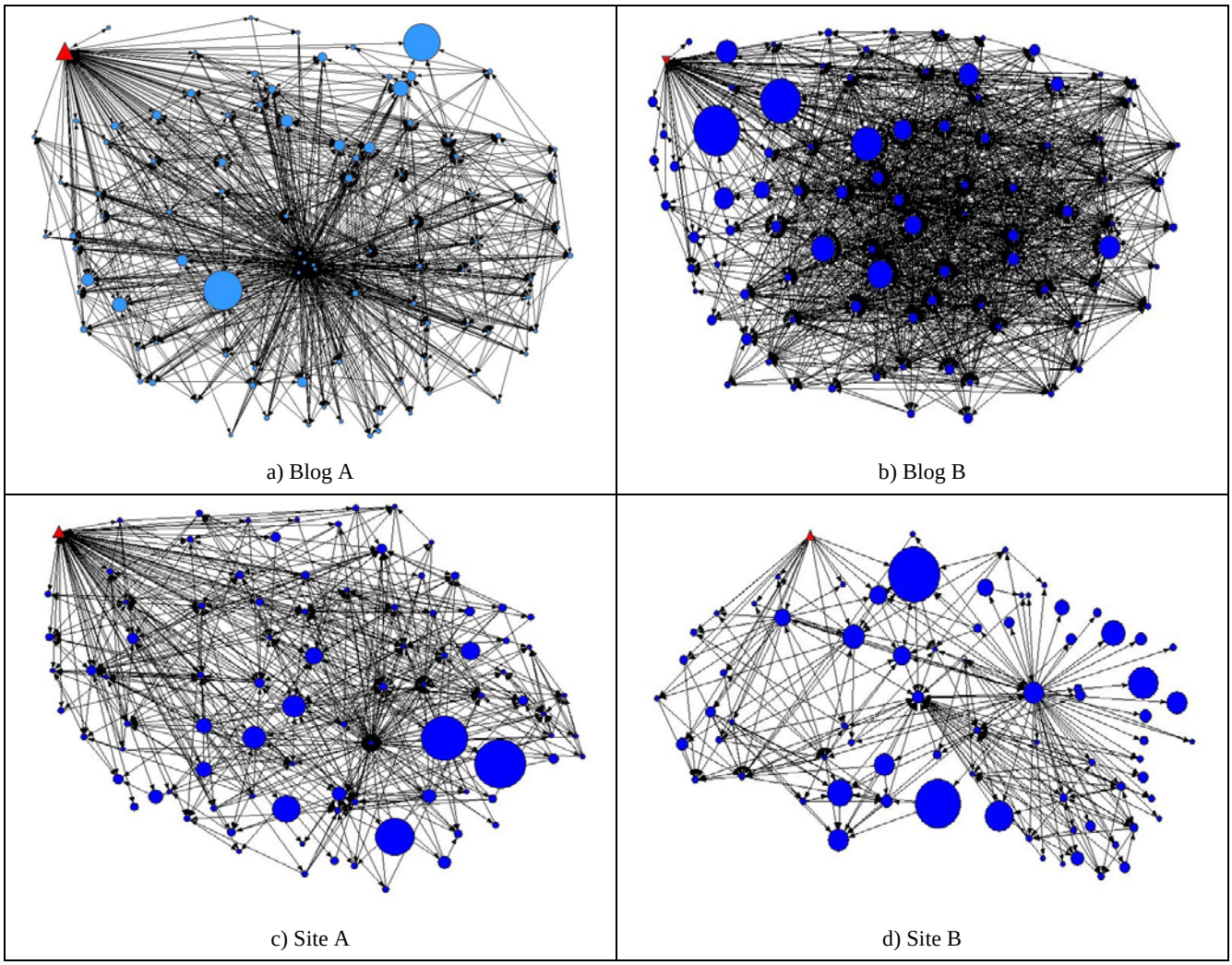
Social Network Analysis (SNA) is a tool that can be used to fulfill these objectives. SNA focuses on the patterns of connections among various entities, whether they are individuals, organizations or, in our case, websites. It has been shown to be a valuable tool for criminologists and law enforcement in determining the organizational structure of various criminal networks, from street gangs [21, 27] and drug trafficking organizations [19, 26, 25] to terrorist groups [18]. It has also been used to analyze the online communication of terrorist groups by collecting information from terrorist forums on the web [16].

Prior to the Internet, child exploitation could have been viewed as more of a solitary crime with very sparse networks [2]. 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 find and get in contact with one another. However, the advent of the Internet has changed the crime of child exploitation and sexual abuse [32]. With many websites (e.g., *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. To our knowledge, child exploitation websites have not yet been studied from the perspective of a network. Doing so has direct implications for law enforcement agencies involved in targeting websites and offenders through the Internet as it allows them to determine the key websites to target.

The current study develops a method to extract child exploitation networks, map their structure and analyze their content. 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 fragmentation of the network and largest loss of hardcore material. From a law-enforcement perspective, this would allow the prioritization of targets, to only highly connected websites that also display more harmful content.

2. METHODS

We propose a method to undertake this analysis efficiently 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. We use a custom-written web-crawler which, given a starting webpage, will recursively follow the links out of that webpage, until some termination conditions apply. 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 pages hosted on that website. The algorithm we designed to do this is described below, along with a description of the networks extracted.



a) Blog A

b) Blog B

c) Site A

d) Site B

Figure 2 – The 4 networks

2.1 NETWORK EXTRACTION

For this paper, we use a custom-written crawler (Figure 1), called Child Exploitation Network Extractor (CENE) to extract the network structure and statistics (features) of the network for analysis. A variety of starting locations can be used to extract multiple networks for comparison purposes. For each network extracted, features are collected about the content of the pages and the links between them. The statistics are then aggregated up to the website level. For example the features for www.website.com are 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 keep the network extraction time bounded, a limit was put on the number of pages retrieved (*PageLimit* – line 3). Second, the network size was fixed at a specific number of websites (*WebsiteLimit* – line 5). This was done so that the network extracted would be focused on websites dealing only with the specified topic. The end result of this process is a network where all the websites in the network are sampled approximately equally, with ($^{PageLimit}/_{WebsiteLimit}$) pages being sampled per website.

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. *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 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 one keyword from *Keywords* had to exist (line 10). If a keyword existed on the page, the page was assumed to be relevant to the network and the statistics on that webpage were calculated (line 11). 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 exist on the page, it was discarded and no further links were followed.

Website	# of Pages on Starting Website	Severity Score	% of All Websites Connected To It	Degree Centrality (Normalized)
Blog A	285	62.93	100.0	100.0
Blog B	583	1.82	81.8	81.6
Site A	237	80.19	78.8	78.6
Site B	1	2.00	27.1	68.4

Figure 3 - Description of Starting Websites

Network		Blog A	Blog B	Site A	Site B
Density (Ties)		0.13 (1214)	0.21 (2006)	0.09 (866)	0.04 (371)
Severity Score Density	**High**	0.23 (n=22)	0.37 (n=27)	0.11 (n=23)	0.08 (n=27)
	Low	0.12 (n=77)	0.16 (n=72)	0.08 (n=76)	0.02 (n=69)
Clustering Coefficient		0.39	0.48	0.28	0.22
Fragmentation		0.04	0.06	0.04	0.02
Centralization	**Out-Degree**	88.38%	61.58%	70.36%	65.03%
	In-Degree	71.89%	31.65%	60.05%	36.31%
Reciprocity		0.25	0.37	0.17	0.02

Figure 4 - Social Network Analysis Summary

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 aggregated to the website level (line 18). 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.

2.2 CHILD EXPLOITATION NETWORKS

Four websites were randomly chosen as starting points through four separate search engine searches, using the keywords 'boy' and 'love'. A search using the keywords 'girl/lolita/lolli' and 'love' was also attempted, but the results were unsatisfactory, leading to adult pornography websites. Of the four starting websites, two were based on user-generated posts, referred to as a *Blog*, and two had the traditional structure of interlinking-pages, simply called a *Site*. These were selected so that we could compare the findings within type (Blog A vs. Blog B and Site A vs. Site B) and between type (Blogs vs. Sites). Although this type of content can most likely be found via other Internet Protocols, such as IRC, NNTP (Usenets) or FTP, they require different types of analysis and hence were not included in this study. Forums are also significantly different from Blogs and Sites in that they require a slightly different approach for access, extraction, and analysis.

Each resulting network was analyzed for the presence of specific keywords which were divided between two categories: "hardcore" and "softcore" words. Although the focus of this study is on the most harmful content, it was important to collect a broader range of keywords for comparative and network extraction purposes (see above). Keywords labeled as hardcore were those with explicit sexual content: mastur*, sex, penis, vagina, anal/anus, oral, virgin, and naked/nude. The softcore words were: boy, girl, child, love, teen, lolli, young, and bath*. As 'smooth/hairless' could be found in both hardcore and softcore settings, they were included in both categories. Each network was capped at 100 websites and 50,000 pages. Therefore, the networks analyzed should not be viewed as complete networks but rather samples of larger networks. CENE retrieved up to 25 pages in parallel, requiring between 6-12 hours to extract each network.

3. RESULTS

First, we draw on SNA to examine the structure of the four extracted networks. More specifically, we derive the following measures:

- **Density**: the percentage of network connections present in relation to all possible network connections [11, 15]
- **Clustering coefficient**: the likelihood that two websites, both connected to another website, are connected to each other [11, 19]
- **Fragmentation**: the percentage of the network connections disconnected by the removal of any one website [3]
- **In-degree centrality**: for website *a* it is the number of other network websites that links to *a*
- **Out-degree**: it is based on how many other websites website *a* links to [11]
- **Reciprocity**: the proportion of websites that reference one another [11, 15]

Second, we analyze the content of the networks through the keyword analysis developed earlier. We compare the relative severity of content by examining the mean number of hardcore words present per page. Third, we turn to the main research question examined in this paper: are the Blogs/Sites with the most harmful content also the most central in the overall network? We do so by comparing network centrality measures to a severity score: number of hardcore words found per page.

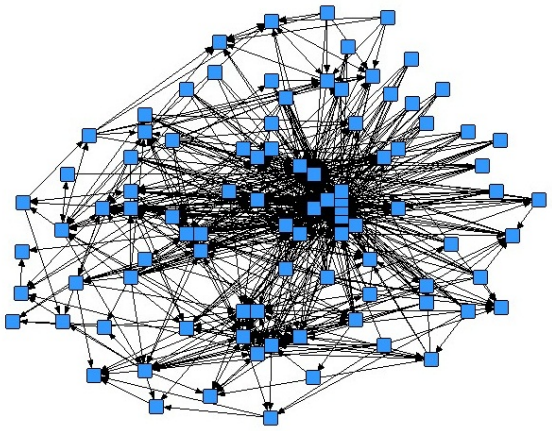


Figure 5 – Site A network after removal of the website with the highest fragmentation score

3.1 NETWORK STRUCTURE

Figure 2 shows the four networks extracted, with a triangle towards the upper left corner denoting the starting location for each network. The circles denoting the websites vary in size based on the severity score (number of harcore words/page). As shown in Figure 3, the final networks consisted of fewer than 100 websites as a few of them were aliases for each other and were consequently merged. The starting Blogs and Sites differed in size and content. For example, the starting blog for Blog A comprised 285 pages, averaging 63 hardcore words per page, while Blog B's starting blog was 583 pages in size, with an average of 2 hardcore words per page. Site A's starting point was 237 pages and 80 hardcore words per page and Site B was 1 page and averaged 2 hardcore words per page[4]. Such variety was an advantage as one of the objectives of the paper was to examine whether such differences led to different network structure and content.

A simple visual examination of the networks in Figure 2 reveals that different structures emerged for all four. Figure 4 provides more details on the similarities and differences that emerged. First, we found that the Blog networks were much denser than the Site networks. Blog A and Blog B had a density of 0.13 and 0.21, respectively, compared to 0.09 and 0.04 for the two Site networks. The clustering coefficients were also higher for the Blog networks at 0.39 (A) and 0.48 (B) compared to 0.28 (A) and 0.22 (B) for the Site networks. As the clustering coefficient for each network is more than double their network density, this indicates that the average densities of individual neighborhoods (websites) are diverse in size and dominated by several large, highly connected websites. One of the reasons why the Blog networks are denser is because of the higher levels of reciprocity. As shown in Figure 4, Blog A and B had high levels of reciprocation (25% and 37%), while Site A and B had much lower

[4] The starting page for Site B was a front page for a much larger site. For example, all 'sections' of the website www.hostsite.xxx followed the url www.section.hostsite.xxx. Therefore, the number of pages and hardcore words are low as there were no additional pages on the front page.

	Percentage of All Keywords Found in Network			
	Blog A	**Blog B**	**Site A**	**Site B**
Boy	60.82	35.89	55.78	70.59
Girl	0.61	4.90	0.43	4.54
Child	1.42	4.20	1.06	6.42
Love	6.75	30.53	19.15	7.66
Teen	4.09	2.30	4.04	0.95
Lolli*	0.00	0.15	0.00	0.04
Young	2.42	3.73	1.70	0.48
Bath*	0.02	0.01	0.12	0.04
Innocent	0.01	0.04	0.00	0.03
Smooth/Hairless	0.21	0.27	0.16	0.41
Mastur*	0.65	0.03	0.27	0.03
Sex	9.58	8.95	8.70	2.46
Penis	2.76	1.10	0.30	0.06
Vagina	0.00	0.12	0.00	0.03
Anal	4.23	3.38	1.03	4.17
Oral	0.74	0.45	0.35	0.55
Naked	5.42	3.27	6.70	0.81
Virgin	0.06	0.40	0.04	0.32

Figure 6 - Percentage of All Keywords Found in Network

levels of reciprocal ties (17% and 2%). This also ties in quite well with (similarly focused) blogs perceived as being a "community" of sorts, at least more so than other types of websites. The significantly lower level of reciprocity for Site B may be attributed to the high number of dead websites (19) or websites without any of our keywords (24). However, the lack of reciprocity may again be a precautionary tactic from the owners. 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 site increases its exposure on search engines, which in turn likely increases the possibility of being shut down.

Second, we found that content matters in determining the overall structure of the network. When dividing the network between those with higher severity scores (greater than the network average) we found that the Blogs/Sites with the most harmful content were more likely to be connected to each other. Figure 4 shows, for example, that the network density was always higher for those websites compared to others.

Overall, these results suggest that each of the networks is dominated by several 'mega' websites, or 'key players'. Initially, this does not seem to be the case as the fragmentation scores were very low for each network: Blog A (0.04), Blog B (0.06), Site A (0.04), and Site B (0.02). Recall that this score indicates that the removal of any random website would result in a loss of 2-6% of

		Blog A	Blog B	Site A	Site B
Number of Websites in Network		99	99	99	96
Number of Pages/Website	**Range**	0-651	0-470	0-1,420	0-1,575
	Average	405	265	268	394
Hardcore Words	**Average (Range)**	1501 (0-14,203)	9352 (0-133,526)	7435 (0-107,016)	1287 (0-21,226)
	Average/Page (Range)	3 (0-27)	38 (0-583)	52 (0-593)	3 (0-30)
	% of Keywords	23.64	17.98	17.55	8.83
Softcore Words	**Average (Range)**	6,847 (0-41,588)	30,214 (0-298,602)	34,934 (0-63,951)	13,283 (0-617,748)
	Average/Page (Range)	15 (0-93)	108 (0-1061)	97 (0-896)	39 (0546)
	% of Keywords	76.37	82.02	82.45	91.17
Total Words	**Range**	0-45,061	0-380,348	0-746,526	0-618,586
	Average	8,348	39,566	42,369	14,570
	Network Total	3,917,045	826,441	4,194,544	1,398,756

Figure 7 – Word content analysis for the four networks

connections within the network. However, a targeted removal of a website may produce more disruption within the networks. For example, the removal of the starting blog for the Blog A network would result in a 16% fragmentation. This was followed by a second blog, whose removal would result in a 10% fragmentation of the network – independent of the removal of the starting blog. For Site A, Figure 5 illustrates how the removal of the starting website for the network results in a 62% fragmentation of the network. For Site B, the removal of the starting website would only result in a 6% disruption; however, there were two other websites, whose removal would result in a fragmentation of 48% each. This indicates that substantial fragmentation can occur if the proper websites are targeted by law enforcement agencies. That is, removing a random website does little to disrupt the network, however, targeting specific websites that link to a lot of other websites can result in larger impacts to the online network.

3.2 NETWORK CONTENT

CENE collected statistics about every single page on the website which was crawled. This was done through the frequency of keywords on each page which was aggregated up to the website level. For a complete list of keywords, and their frequencies, see Figure 6.

Given that our initial search engine search was boy-centered, it comes as no surprise that 90.2% of the websites across networks were classified as such; while the other 10% were girl-centered. This was based on the higher ratio of 'boy' to 'girl' references on the website. Blog A and B were 100% and 83% boy-centered, while Site A and B were 99% and 75%.

It is important to note that there was some cross-over between boy and girl centered websites. However, given the small number of girl-centered websites it is unclear if the 10% of the websites classified as girl-oriented is evidence of the two network types being connected or simply by chance. Regardless, most websites within the network seemed to be predominantly boy-centered or girl-centered. This implies that child exploitation websites do not mix boy/girl material; rather they tend to focus on a specific gender, possibly impacting the choice of strategies for police investigations.

Across networks, 81.3% of keywords found belonged to the group we defined as 'softcore', while 18.7% belonged to 'hardcore'. Figure 6 shows that the most common keywords were 'boy' (58.1%), 'love' (13.8%), 'sex' (8.2%), 'nude/naked' (5.1%), and 'anus/anal' (2.3%).

Figure 7 presents the results of the content analysis of the four networks extracted. All networks contained the same average number of pages per website; however, blogs had higher counts of hardcore words, expressed as a higher severity scores per page (16.2 to 13.4) and per website (5426.3 to 4408.4). Despite this, Figure 7 shows that Site A and B had the larger ranges. Blogs were fairly consistent, while there was a wide range of values obtained from Sites. Additionally, Blogs contain more hardcore content per page. This could be attributed to the ease of setting up a blog as well as the increased anonymity afforded to the operator. This is in comparison to sites, whose operator's personal information is linked to the website and thus are at an increased risk of facing formal charges.

To determine whether severity was related to the number of links coming into, and going out of, a website, the total number of hardcore words per website and per page was correlated with in and out-degree centrality. Although none of the correlations were significant, the pattern suggested that hardcore blogs and sites have a tendency to reach out more to others (r=0.10, 0.13, 0.12, -0.05) than others reach out to them (r=-0.09, 0.04, -0.16, 0.12). These findings support the previous analyses that there are 'mega' websites with a lot of material and a lot of connections, as well as small independent websites, with only a little bit of material and relatively unconnected to the rest of the network. Put another way, the mean number of hardcore words per website and per page are mainly driven by several extreme websites on both ends (websites with a lot of content and websites with little to no content).

Website Ranking	Blog A		Blog B		Site A		Site B	
	In-degree	**Severity**	**In-degree**	**Severity**	**In-degree**	**Severity**	**In-degree**	**Severity**
1	82	1.20	50	2.05	77	80.19	38	3.50
2	80	1.20	49	14.17	23	4.01	15	1.12
3	79	1.00	48	3.61	22	62.00	13	1.17
4	79	1.13	47	4.16	19	29.00	12	10.04
5	78	0.91	46	2.34	18	17.36	10	9.64
6	30	62.93	45	2.23	17	119.00	9	0.15
7	26	1.13	45	4.31	17	36.00	9	0.33
8	25	51.21	45	3.27	16	39.56	9	0.75
9	22	20.94	43	7.89	16	12.39	8	11.00
10	22	10.07	42	1.97	15	137.00	8	3.76
Mean for Top 10	*52.30*	*15.17*	*46.00*	*4.60*	*18.40*	*53.65*	*13.10*	*4.15*
Mean for Network	*12.26*	*38.26*	*20.26*	*3.16*	*8.75*	*52.21*	*3.86*	*2.98*

Figure 8 - Top 10 In-degree websites in each network compared to the overall network

3.3 FINDING THE 'HARDCORE KEY PLAYERS'

The last set of results focuses on the question at the core of this paper: are the most central websites also the ones hosting the most serious content? Figure 8 examines the top 10 most central websites (sorted by in-degree) and compares their severity score to the overall network. The results show that the top 10 most central websites were no more or no less likely to contain offensive material. Instead, there seemed to be a more or less random variation in the four networks analyzed. Thus, offensive material on a website does not seem to influence centrality.

Figure 9 proceeds the other way around, listing the top 10 websites with the highest severity score and examines their centrality. These results are similar: the most hardcore websites were no more or no less likely to be central players in their networks. This procedure, however, allowed us to identify the hardcore key players: websites that were both central in the network and contained offensive content. For example, one of the blogs with the most offensive material in the Blog A network (62.93 score) had 30 other websites linking to it (rank 6, Figure 8). Compare this to the website ranking as number 2 for the Blog B network in Figure 8 (and ranking 4 in Figure 9): a much lower severity score of 14.12, but more websites (49) linking to it.

4. DISCUSSION

The Internet has changed the way society communicates and obtains information. Despite the positive contributions the Internet has made to society, it has also created a new avenue where individuals can engage in criminal activity. With the ease at which people can communicate with one another, from all over the world, the Internet has facilitated the proliferation of child exploitation. The simplicity of obtaining and sharing this material was clearly evident within the current study. Clearly, there is no way to effectively eliminate online child exploitation. Therefore, steps need to be taken to lessen the impact and severity, as well as maximize the efficiency of current efforts. This is where SNA can be of the greatest use to law enforcement.

The use of blog websites for child exploitation provides plenty of advantages. If an individual were to setup their own non-blog website, they would have to have some knowledge of how to design a website as well as have the financial capital to pay for the website. In addition, they would have to be cautious of detection by law enforcement. However, blogs provide a much cheaper, more efficient, and more anonymous way to distribute material. Many blog webhosts such as Blogger, LiveJournal or Sensualwritter provide members with free space to post their blogs. This eliminates the out-of-pocket expense and knowledge needed by an individual to set-up their own website. In addition, and possibly the most important advantage, blog webhosts do not verify personal information about their members. Therefore, the blog webhost allows the individual to be completely anonymous. Although each blog webhost has terms of service that state that copyrighted or illegal material is not allowed, it is usually the responsibility of patrons to report a blog containing material that violates the terms of service. If a blog is seen to be in violation of the terms of service it is usually removed by the webhost. However, there is nothing preventing the owner of the blog from creating a new account and starting the blog all over again. Therefore, the removal of the blog could be viewed as no more than a mild inconvenience for the blog creator.

As the problem of online child exploitation is continually growing and more websites are becoming hosts of material, the responsibility to combat the problem is not solely on law enforcement. Understanding the immense number of hours and resources that go into finding sexual explicit material online, one of the largest search engines in the world, Google, has begun to help. In conjunction with the National Center for Missing and Exploited Children (NCMEC), Google announced the creation of new software that would "aid in organizing and indexing NCMEC's information so that analysts can both deal with new images and videos more efficiently and also reference historical material more effectively." [1]. However, blog webhosts need to get on-board as well. Considering that one of the top blog webhost, Blogger, is owned by Google, it should be possible for

Website Ranking	Blog A		Blog B		Site A		Site B	
	In-degree	Severity	In-degree	Severity	In-degree	Severity	In-degree	Severity
1	7	583.08	6	27.50	2	593.00	8	30.00
2	5	543.14	12	24.07	1	531.00	1	26.00
3	9	193.71	36	17.33	1	431.00	1	16.00
4	8	177.53	49	14.17	1	292.45	4	14.92
5	15	130.06	18	12.97	4	244.00	8	13.48
6	10	128.05	30	11.39	9	244.00	1	12.00
7	4	125.13	3	11.11	1	182.00	8	11.00
8	14	123.95	20	10.40	11	149.88	2	10.08
9	16	113.83	4	10.00	1	146.00	12	10.04
10	7	95.81	37	8.90	1	137.00	1	10.00
Mean for Top 10	*9.50*	*221.43*	*21.5*	*14.78*	*3.20*	*295.03*	*4.60*	*15.35*
Mean for Network	*12.26*	*38.26*	*20.26*	*3.16*	*8.75*	*52.21*	*3.86*	*2.98*

Figure 9 – Top 10 severe website in each network compared to overall network.

Google Inc. to also use its image recognition software on Blogger. Obviously, this process would have to be automated as it would be very difficult, and highly inefficient, for people to have to routinely check all blogs for illegal material. However, its implementation could have a substantial impact on curbing child exploitation on blogs.

5. CONCLUSIONS

The current study drew on social network analysis to examine the content and structure of online child exploitation networks. We extracted the structure and features of child exploitation networks by performing a guided crawl of the Internet. Our crawler, CENE, was guided by a set of keywords, and exclusion websites, which kept it on topic. This provided very focused networks for analysis.

Using social network analysis we attempted to find the key players—those websites displaying a combination of connectivity and hardcore material. This analysis looked at two types of websites: blogs and sites, covering four independent starting points. Our results indicate, first, that the presence of hardcore content is not the basis for linkages between websites. Second, that blogs contain more hardcore content per page than sites.

Although this exploratory study has made substantial additions to our current understanding of online child exploitation, it has also laid the groundwork for the incorporation of SNA into future research on this topic. Subsequent research needs to expand on the network size(s) and shift to a more detailed analysis of the attributes, including the content of forums, videos and pictures, as well as data on the number of people visiting the websites. Finally, there needs to be a refinement of a) the keywords list (are hardcore words truly "hardcore"?), b) the list of websites the crawler cannot enter, and c) the criteria to reduce the occurrence of false positives.

6. ACKNOWLEDGEMENTS

Partial funding for this project was provided by the International Cybercrime Research Centre, Simon Fraser University.

7. REFERENCES

1) Baluja, S. Building software tools to find child victims. Retrieved April 2, 2010, from http://googleblog.blogspot.com/2008/04/building-software-tools-to-find-child.html

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

3) Borgatti, S.P. (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 Sciences Press.

4) 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.

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

6) 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

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

8) Garton, L., Haythornthwaite, C., & Wellman, B. (1997). Studying online social networks. Journal of Computer-Mediated Communication, 3. Retrieved from: http://jcmc.indiana.edu/vol3/issue1/garton.html

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

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

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

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

13) Internet World Stats. (2009, December 31). World Internet usage and population statistics. Retrieved April 7, 2010, from http://www.internetworldstats.com/stats.htm

14) INTERPOL. (2009b). Crimes against children. Retrieved April 2, 2010, from International Criminal Police Organization Web site: http://www.interpol.int/public/children/default.asp

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

16) Qin, J., Zhou, Y., Lai, G., Reid, E., Sageman, M., and Chen, H. (2005). The Dark Web Portal Project: Collecting and Analyzing the Presence of Terrorist Groups on the Web, Lecture Notes in Computer Science, Vol 3495, pp 623-624.

17) 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.

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

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

20) McCullagh, D. (2008 March 20). FBI posts fake hyperlinks to snare child porn suspects. Retrieved April 7, 2010 from: http://news.cnet.com/8301-13578_3-9899151-38.html

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

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

23) 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.

24) 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.

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

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

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

28) 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/

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

30) 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.

31) 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

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

33) 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.

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

35) 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: Brunner Routledge.